Internet and Human Rights III

Contributions to the discussion in Latin America

Agustina Del Campo
Compiler

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

CELE

UP
Universidad de Palermo

Internet and Human Rights III
Contributions to the discussion in Latin America

Internet and Human Rights III

Contributions to the discussion in Latin America

Agustina Del Campo
COMPILER

Law School
Centro de Estudios en Libertad de Expresión y Acceso a la Información

CELE

UP Universidad de Palermo

Compiler:
Agustina Del Campo

General design:
Departamento de Diseño
de la Universidad de Palermo

Correction:
Carla Ortiz Rocha

Translation:
María Soledad Váquez

Edited by Universidad de Palermo
March 2020,
Buenos Aires, Argentina

Universidad de Palermo
Rector
Ing. Ricardo H. Popovsky

Facultad de Derecho
Dean
Dr. Leandro Vergara

Center for Studies on Freedom of Expression
and Access to Information (CELE)
Director
Agustina Del Campo

Mario Bravo 1050
(C1175ABW) Ciudad de Buenos Aires
Argentina
Tel.: (54 11) 5199-4500 | cele@palermo.edu |
www.palermo.edu/cele

Index

Foreword

iLEI

In recent years, discussions about Internet regulation have become more complex and they have gained particular visibility and relevance in the public agenda. The first volume of this series highlighted the early state intervention in Internet regulation and the need to address such intervention from a human rights perspective. Volume II addressed some of the most relevant issues of the legislative agenda and the problems associated with Internet governance on a broader level. Volume III delves into the study of a few topics that enjoy some permanence on the public agenda, as the poorly named "right to be forgotten", and includes new issues such as misinformation, regulation of Over-The-Top services (OTTs) or the challenges presented by humorous speech online.

The first article, *Implementing "Digital Oblivion": Forgetting Details*, written by Carlos Cortés and Luisa Isaza, studies the judicial implementation of the "right to be forgotten", focusing on its various meanings and approaches at a comparative level in Latin America. The authors declare: "There is a risk that the implementation of a technical order will cause unforeseen effects, especially when this does not take into account the praxis of the digital environment and the context in which it must be executed. That is the problem that we identify in our region: the risk of not understanding the technical implications of an order can lead to disproportionate measures or be an incentive for self-censorship".

In the second article, *From comics to memes: old and new problems regarding humor and freedom of expression*, authors Agustina Del Campo and Paula Roko evaluate the public and private regulation that currently impacts the production and dissemination of humor, political humor in particular. The authors conclude that "Setting up short terms for action in the style of NetzDG, the resolution of disputes in the hands of companies and not of State justice, and more recently the incentives for the adoption of filters (used before and after loading) without prior complaint or notification, are strong incentives for automating processes of detection, removal, and moderation of content. Accordingly, the contexts, symbolisms, and language that characterize satirical expressions, parody and humor in general — but particularly political humor — are held back, become invisible and smothered".

Fake news on the Internet: the strategy to battle misinformation is the third article in this publication. Authors Cortés and Isaza identify, systematize and contrast the main policies adopted by Google, Facebook and Twitter to combat disinformation on their platforms and conclude with a call for greater transparency.

The fourth article, *Over-the-Top Services: fundamental principles for discussing their regulation in Argentina*, by Maia Levy Daniel, analyzes the tensions between the positions of the various actors in the debate concerning the regulation of OTTs, with special emphasis in the need to frame these discussions from a human rights perspective.

Finally, this volume also includes three contributions drafted from within iLEI linked to rules of self-regulation and corporate liability regarding the protection and promotion of freedom of expression. The first one is addressed to the United Nations Rapporteur for Freedom of Expression and Opinion for his report on moderation of content on the Internet (2018); the second one is in response to a query opened by Twitter regarding the possibility of modifying its community rules about dehumanizing discourse; the third one is in response to Facebook's proposal to create an Oversight Board of independent experts to interpret and make concrete recommendations regarding its community rules and their implementation.

"Internet and human rights" has become a series of publications by CELE's Initiative for Freedom of Expression Online (iLEI). These papers are intended to become useful tools to support legislators, civil society, journalists, government institutions and the private sector in the difficult work they face in designing, implementing and monitoring the policies that regulate the Internet, particularly from a human rights perspective. This third volume was done thanks to the generous support of the Ford Foundation.

Chapter One

Implementing "Digital Oblivion": Forgetting Details

Carlos Cortés* and Luisa Isaza**

I. Introduction

"In a connected world, someone's life can be ruined in a matter of minutes and a person stuck frozen in time," says American academic Meg Leta Jones.[1] The Internet offers us examples on a daily basis: a humiliating moment that is captured on a cell phone and multiplies in social networks like a virus; a news scandal that nobody remembers but that persecutes its protagonist every time someone Googles their name; an old tweet that is unearthed to embarrass the author... It seems that the past on the Internet does not exist. We live in an eternal present.

We have been talking about the "right to be forgotten" on the Internet for several years. In a broad sense, this is the power to silence a person's past events; this "right" has been making its way into our region.[2] Its origin is found primarily in the protection of an individual's personal data. The international benchmark that allowed the use of "digital oblivion" took place in

* Carlos Cortés is a researcher of the Iniciativa por la Libertad de Expresión [Initiative for Freedom of Expression] (iLEI) of CELE. He is a Lawyer from the Universidad de Los Andes (Colombia), with a Master's degree in Media Governance from the London School of Economics. He is a consultant in freedom of expression and Internet regulation.

** Luisa Isaza is a Lawyer from the Universidad Javeriana (Colombia). She is a Legal advisor for the Coordinación de Defensa y Atención a Periodistas de la Fundación para la Libertad de Prensa [Defense and Attention to Journalists Office, Foundation for Freedom of the Press] (FLIP), Colombia.

*** This article was originally published by CELE in December 2018.

[1] Leta Jones, Meg, Ctrl + Z: The Right to Be Forgotten, New York, NYU Press, Kindle edition, 2016, loc. 105.

[2] Cf. Pino, Giorgio, "The Right to Personal Identity in Italian Private Law: Constitutional Interpretation and Judge-Made Rights", in: Mark Van Hoecke and François Ost (eds.). The Harmonization of Private Law in Europe, Oxford, Hart Publishing, 2000, pp. 225-237.

Europe: in 2014, the Court of Justice of the European Union decided that the Google search engine should avoid showing old news referred to a person's past debt when searching for their name (known as the "Costeja ruling").

In that ruling, the European justice urged Google to develop a private system whereby European citizens can request the search engine to de-index inaccurate, inadequate, irrelevant or excessive information about them. In general terms, de-indexing implies that the search engine disassociates a particular link from the search results.[3] Between mid-2014 and the end of 2017, Google received requests for de-indexing 2.4 million Internet addresses (URLs), 43% of which were executed. 89% of the cases were undertaken by private citizens.[4]

The Costeja ruling made the search engine liable for guaranteeing "digital oblivion" as opposed to the party which had produced the information or was hosting it. That is to say: according to this decision, the search intermediary must prevent the "harmful" information from being associated with the name of the affected party. The information still exists, but finding it is much more complicated by not being part of the search results under that person's name.[5]

This precedent has been applied in some administrative and judicial rulings in Latin America. While the private Google de-indexing system is in place in Europe, in our region there are specific cases where an authority gives an equivalent order. In any case, neither the administrative nor judicial authorities in our region have reached the point of ordering the creation of a private system similar to the European one.

In contrast, other decisions in the region focused the responsibility of "digital oblivion" on the author or publisher of the information. As a result, they ordered the responsible party to update or delete specific information. In this scenario, the search engine is not part of the controversy, and if the author of the information eliminates it, this information simply disappears from the search results.

In the midst of these two types of orders, those focused on the search

[3] Although the term "unindexing" is also used, this document chooses to use the word "de-indexation", either partially or as a whole. Part of the purpose of the paper is to make it clear what this action means in each case.

[4] Smith, Michee, "Updating our 'Right to Be Forgotten' Transparency Report", February 26, 2018, retrieved from: https://bit.ly/2SGwtLT.

[5] It is currently being debated whether the de-indexation measures adopted by Google based on this ruling should be applied only in their European domains (such as Google.fr, Google.es, etc.) or if, on the contrary, there is a right to request a global de-indexation. The Court of Justice of the European Union is close to deciding on a preliminary issue raised by the Council of State of France in this regard after Google sued the National Commission for Information Technology and Liberties of France (CNIL) for having imposed a penalty for limiting the de-indexing to European domains.

engine and those directed to the publisher, there are combined, intermediate, and openly contradictory decisions. When trying to implement "digital oblivion", the judges of Latin America are going through unfamiliar territory. Understanding the technical implications and exploring their implementation and consequences is the objective of this paper.

It is important to delimit the scope of this analysis: the digital "right to be forgotten" includes several discussions about the need to return control over the data and the information that circulates on the Internet to the individual. In this paper, we want to focus on the "digital oblivion" that impacts the exercise of freedom of expression and, in particular, the activity of the media and those who disseminate information and opinions of general interest. That is to say, "digital oblivion" is not studied in relation to the discussion about the personal information that companies gather and keep in databases. We focus on the orders that resolve conflicts between freedom of expression and individual rights such as the right to a good name, privacy and the free development of personality, due to information that is published openly on the Internet.

This paper focuses on decisions about "digital oblivion" in its broadest sense: those that seek to eliminate, obscure or hinder access to certain information. On the other hand, we do not address other components of those rulings that seek to balance public debate. This is the case of rulings about information update, those that grant a right of reply or those that require the publication of a warning (flagging). It is worth adding, however, that these last rulings do not exclude those pertaining to "digital oblivion" (think, for example, of a provision that requires the elimination of information and at the same time requires granting affected party the space to exercise their right to reply).

Furthermore, this document does not address in a critical manner the concept of the "right to be forgotten" or the legitimacy of the agent that implements it. This has been dealt with in other publications of the Center for Studies on Freedom of Expression and Access to Information (CELE) and in previous works by the authors.[6] We believe that the idea of "digital oblivion" faces many theoretical questions, and that the incorporation of this "right" in the

[6] *See*, Cortés, Carlos, "*Derecho al olvido: entre la protección de datos, la memoria y la vida personal en la era digital*", in: Bertoni, Eduardo (comp.). *Internet y derechos humanos. Aportes para la discusión en América Latina*. Buenos Aires, CELE-Universidad de Palermo, 2014; Ferrari, Verónica and Schnidrig, Daniela, "Responsabilidad de intermediarios y derecho al olvido. *Aportes para la discusión legislativa en Argentina*", in: Bertoni, Eduardo (comp.). *Internet y derechos humanos II. Aportes para la discusión en América Latina*, Buenos Aires, CELE-Universidad de Palermo, 2016; Botero, Catalina, Camilleri, Michael J. and Cortés, Carlos, "*Democracia en la era digital. Libertad de expresión y el derecho al olvido europeo*", in: *El Diálogo. Liderazgo para las Américas*, Informe del Programa de Estado de Derecho, November, 2017.

context of the Inter-American Human Rights System should continue to be discussed. On this occasion, we want to focus on the implementation issue.

"Digital oblivion" implies a series of actions on information, data and content: elimination, updating, obscuring and invisibility. These actions can operate on different levels, involve different actors and generate variations in the type of "oblivion". Do judges, administrative authorities and interested civil society understand the implementation of these orders? Are the proportionality and possibility of these orders being evaluated? Are the impact they have and the incentives they generate being analyzed? Do the environment and practices of private companies allow the understanding of technical problems? Is there a transparent and honest discussion about it?

To understand where oblivion actions are implemented, the first part of this paper describes the concepts of transactive memory and open and closed systems. The purpose is to explain, on the one hand, how the closer the method of "digital oblivion" is to the storage core, the more domains will be affected. And, on the other, how that "oblivion" usually involves the action of several players, which cannot be impacted by a single decision or technical measure.

Subsequently, "digital oblivion" is addressed in practice towards the two agents involved: the search engine and the author of the information. In both cases, emphasis is placed on the nature of the orders issued and their possible implementation. Finally, the last part offers some conclusions focused on the problems of proportionality and incentives.

This paper attempts to simplify a complex issue in order to bring the discussion to judicial operators and members of civil society. In that regard, it omits some technical details, but also part of the assumption that the audience has a minimal previous frame of reference on the debate on "digital oblivion".

II. Transactive memory and open and closed systems

Throughout the centuries, human beings have created and perfected techniques and tools that support our perishable memory. The clearest example is the external objects that "memorize" information for us: a phone book, the alarm clock or a post-it with a reminder. In the digital environment, these objects have been replaced by applications and services: phone numbers stored in cell phones, the cell phone's alarm or its list of tasks.

Memory also operates, above all, in cooperation with other people. In our group of friends or colleagues, a decentralized process of storage and

consultation is constantly taking place: "I remember that Pablo's birthday is July 15 because that day my niece was born, but I do not remember the date for Felipe's day," someone says. "It's November 19, like the song," someone else replies. In a conversation as simple as this, a transactive memory process is performed, meaning the "combination of individual minds and the communication between them".[7]

Transactive memory refers to how information is coded, stored and retrieved collectively. According to the psychologists who coined the term, transactive memory is defined by two components: (i) an organized storage of knowledge that is housed entirely in the individual memory systems of the members of the group, and (ii) a set of relevant transaction processes for knowledge that occur between members.

As it is a space in which people can consult or provide information, the Internet is a system of transactive memory.[8] This is relevant to understanding "digital oblivion", what its effects are and who is the agent responsible for taking the necessary measures to "forget". We are not confronted with the external memory of a single agent but multiple agents, each of which may have different interests in the elimination or retention of the information in question. "The problem of fulfilling an individual's need to 'be forgotten' from an external network memory storage is that it is not the external storage of the individual, but a transactive one."[9]

Transactive memories can reside in open or closed access systems. According to Vijfvinkel, in an open system the information can be copied by anyone, without there being the possibility of keeping tags on the copies.[10] This is the case, for example, of an open discussion forum in which anyone

[7] Cf. Wegner, Daniel M., Giuliano, Toni and Hertel, Paula T., "Cognitive Interdependence in Close Relationships", in: Ickes, William J. (ed.). Compatible and Incompatible Relationships, New York, Springer-Verlag, 1985, p. 256. Retrieved from: https://bit.ly/2sf9qMQ, last access: December 28, 2018.

[8] The notion of the Internet as a transactive memory was popularized by a famous study published in *Science* magazine in 2011 that revealed that people tend to forget the information they hope to be able to consult again using the web. Sparrow, Betsy, Liu, Jenny and Wegner, Daniel M., "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips", in: *Science*, Vol. 333, No. 6043, 2011, retrieved from: https://bit.ly/2xTCB8i, last access: December 28, 2018.

[9] Korenhof, Paulan, Ausloos, Jef, Szekely, Ivan et al., "Timing the Right to Be Forgotten: A Study into 'Time' as a Factor in Deciding About Retention or Erasure of Data", in: Gurwirth, Serge, Leenes, Ronald and de Hert, Paul (eds.). *Reforming European Data Protection Law*, Vol. 20, Dordrecht, Springer, 2015.

[10] Vijfvinkel, M.M., "Technology and the Right to be Forgotten", thesis submitted to obtain the Master's degree in Computer Science, Radboud University, Nijmegen, Netherlands, 2016.

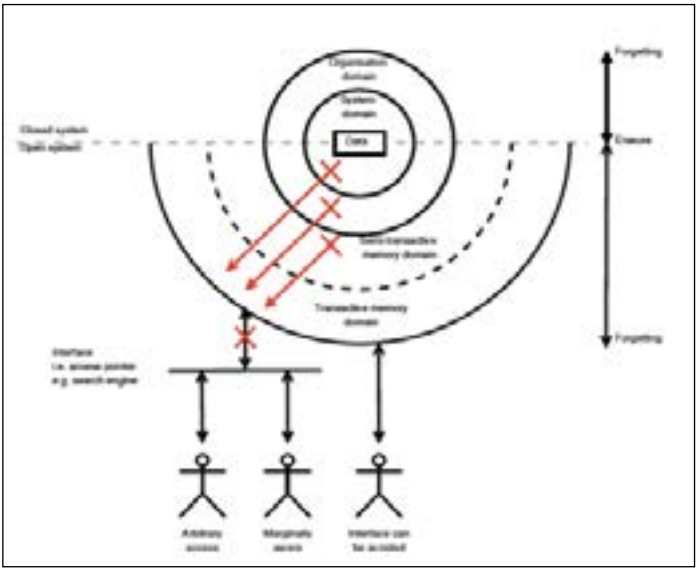
can read the information published by its participants. Meanwhile, in a closed system the dissemination of information is limited to controlled spaces, and all its operators are limited by space restrictions and by conventional or legal obligations of non-disclosure.[11] Think about the system for consulting and editing medical records of a hospital's patients.

Vijfvinkel explains that both open and closed access systems operate on different levels: the first is the layer of the physical medium where the data is stored — the "domain of the system" — and the layer where the data can be consulted — the "domain of the organization". In the medical records system, the domain of the system would be, for example, the server where those are stored, and the domain of the organization would be for the members of the medical and administrative personnel authorized for consults. This is the end of access levels in a closed system like this, where there is no space for public consults.

In an open access system, however, there are additional layers where the possibilities of consultation and exchange of information increase. And the more open the access to the system, the more transactive it will be. Facebook, for example, requires user authentication to access its contents, this would be a semi-transactive memory. Wikipedia, which in contrast does not require this authentication neither to consult the contents nor to produce them, would be a transactive memory.

Another example of an open access system of transactive memory, also relevant to this paper, is the one between the web page of a media outlet and the search engine used to locate a newspaper article from a term introduced by a user. Although knowledge is housed in one place — the media outlet's server —, the search engine generates the process so that knowledge is available to enable users to "remember".

[11] Druschel, Peter, Backers, Michael and Tirtea, Rodica, "The Right to Be Forgotten. Between Expectations and Practice", European Network and Information Security (ENISA), 2012.



“The closer the method of ‘oblivion’ is to the data storage location, the more domains will be affected,” explains Vijfvinkel.

The domain system allows us to understand the type of measures implemented by “digital oblivion”: “The closer the method of ‘oblivion’ is to the data storage location, the more domains will be affected,” explains Vijfvinkel.[12] Depending on the layer where the measures are implemented, the effects will vary and be unequal among the users of the information. For example, an order to delete information in the domain of the system — in the digital file of a media outlet, for instance — will affect all other domains and all users. On the other hand, a measure applied to the domain of the transactive memory — a de-indexing order for the search engine — will not affect the information: it will still be available to those users who move in the semi-transactive domain (they are subscribers of the media outlet or they know the access link, for example), or in the domain of the organization (media journalists with access to the server). However, both will have to know how to look for it.

Another important aspect of the domain system is to understand who controls the information and who is really in a position to eliminate it from the system or to control how it is exchanged in a transactive system. Following the previous example, the search engine that indexes Internet pages for the search results cannot modify or delete the information in the domain of the system. On the contrary, if the media outlet that published the article removes

[12] Vijfvinkel, *supra* note 10.

it from its servers — that is, from the domain of the system — this will affect the search engine result. However, if that piece of news is available in an open access system, the media outlet cannot control all subsequent actions that are built in the transactive memory: an article published in a media outlet can be shared in screenshots or copied and published in a third-party blog, and very possibly these contents will in turn be indexed by the search engine.[13]

III. The "digital oblivion" in practice

Implementing the digital "right to be forgotten" requires entering the engine room of a media outlet or an intermediary. In other words, it means that these agents have to be made aware of the terms in which their service must relate to the content. As we stated in the introduction, many of these decisions are made without taking into account the capabilities and technical consequences. Next, we will provide examples of judicial or administrative decisions to classify and analyze those orders taking into account the responsibility assigned to the two relevant actors: the search engine and the author or publisher of the information. In the latter case, as was already stated, we want to focus on the media, but we will refer to some decisions that center on other agents and that are relevant to the present analysis.

1. Focus placed on the search engine

1.1. De-indexing (total and partial)

In the Costeja ruling, the Court of Justice of the European Union considered that search engines (and Google in particular) carry out the processing of personal data when indexing the websites of third parties. Consequently, they are responsible for complying with European regulations in this matter.[14] And although the court recognized that authors of external content — like the media — can prevent their publication from being indexed by search engines

[13] The media outlet would have the possibility to control subsequent actions if it works as a closed access system, for example, in those cases where it has implemented digital rights management measures or anti-copy programs (also known as DRM for digital rights management), through which it can control the access and use of its publications.

[14] Tribunal de Justicia (Gran Sala), "Google Spain, S.L., Google Inc. vs. Agencia Española de Protección de Datos (AEPD), Mario Costeja González", judgment, May 13, 2014, retrieved from: https://bit.ly/2sug3Nu, last access: December 28, 2018.

(through measures that will be discussed later), it concluded that this option does not remove liability from search engines, especially when those authors do not respond to the questioned content.[15]

The ruling then states that Google has to adopt the measure of "digital oblivion". The technical instruction given by the court was not very thorough: the search engine is forced to evaluate and, when appropriate, eliminate from the search results — obtained from a person's name — the 'links' to web pages that contain the questioned information.[16] This means that, according to the ruling, Google must not de-index the questioned "link" from the search results completely. It is a partial de-indexation: Google restricts the results of a search associated with a particular name, but maintains the links if other related terms are searched.[17]

Let's illustrate the point with an example: John Doe is a citizen who several years ago was involved in the crime of tax evasion, this piece of news was reported by a local media outlet and it appears indexed as Google's first search result when typing his name. Following the criteria established by the Costeja ruling, the information about Doe is excessive and inadequate, so Google must partially de-index the news: when someone types "John Doe" that result should not appear in the search engine. But if someone performs a search and uses different terms ("tax evasion" and the year or city where the event occurred), the result is still available.

Based on the precedent set by the Costeja ruling, other courts in our region adopted analogous or similar criteria. In 2014, for example, Mexico's

[15] The court justified the search engine's liability stating that: (i) the author of a web page can benefit from the exception that protects the processing of data for journalistic purposes, (ii) including this content in the results of a search of someone's name enables any Internet user to access this information, which represents a greater breach in privacy than the one carried out by the author of the web site, and (iii) since the information can be copied easily by sites that are not subject to European Union law, effective protection of data owners could not be achieved if they had to resort to such websites.

[16] According to the court, the decision must be made in relation to "links" to web pages with truthful information, legitimately published by third parties, when said information is "inadequate, not pertinent, or is no longer pertinent, or is excessive in relation to the purposes of the treatment in question carried out by the search engine". And the judgment states: "The manager of a search engine is obliged to remove from the list of results obtained after a search made of the name of a person links to web pages published by third parties and containing information relating to this person, also in the event that this name or this information is not previously or simultaneously deleted from these web pages, and, if applicable, even if the publication on said pages is in itself lawful."

[17] The European court requires that partial de-indexing be made in all European domains of Google Search (Google.fr or Google.es) and even searches from non-European domains that are made within the territory of the person who requested the measure of "digital oblivion".

data protection authority issued a partial de-indexation order for media information.[18] The administrative authority ordered Google Mexico to refrain from "treating the personal data of the Holder, consisting of their name and surname, in such a way that, when said name is typed into the search engine of the Responsible Party, the links or URL (indexation) do not appear as was requested by the Holder".[19] In 2016, the decision was annulled by a court protection known as *amparo*.[20]

The orders issued by the European court and the Mexican authority are directed to the search engine and not to the author of the website that published the information, which puts us in the domain of transactive memory. The measure of "digital oblivion" does not affect internal domains: the information will remain available for those who have the direct "link" to the article or for those who perform a search on the same web page of the media outlet, but will not be available via a Google search of the name of the person affected.

1.2. Degradation of the search ranking

A well-known Internet "meme" says that the best place to hide a dead body is page two of Google's search results. Why is this so? Nobody will look for it there. According to a 2014 study, 95% of a search engine's traffic is concentrated on the first page of results.[21]

Following this logic, the Privacy Commissioner of Canada recently proposed that search engines be forced to remove certain "links" in the ranking of search results as a way of protecting the reputation of people on the Internet.[22] However, to date this solution has not been ordered by any authority.

---

[18] The authority is the *Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales* [National Institute of Transparency, Access to Information and Protection of Personal Data], (INAI).

[19] INAI, file: PPD.0094/14, Responsible Party: Google México, S. de R.L. de C.V., 2014, retrieved from: https://bit.ly/1jr0U6E, last access: December 28, 2018.

[20] *"¡Ganamos! Tribunal anula resolución del INAI sobre el falso 'derecho al olvido"*, Red en Defensa de los Derechos Digitales, August 24, 2016, retrieved from: https://bit.ly/2VCS0qW, last access: December 28, 2018.

[21] Jacobson, Madeline, "How Far Down the Search Engine Results Page Will Most People Go?" in: *Leverage Marketing*, August 14, 2017. Retrieved from: https://bit.ly/2scfc1F, last access: December 28, 2018.

[22] Office of the Privacy Commissioner of Canada (OPC), "Draft OPC Position on Online Reputation", January 26, 2018, retrieved from: https://bit.ly/2CU0rXg, last access: December 28, 2018.

"The best place to hide a dead body is page two of Google's search results."

This solution would have an impact on the domain of the transactive memory, although with a different effect than the partial or total de-indexing, since the "link" is kept up, even if the name of a person were searched. The question would be if this form of "hiding" the "link" would work for any search or if, like in the case of the European ruling, its ranking was lowered only when searching for the name of the person concerned. For the moment, this proposal has not been further developed.

2. Focus placed on the author or publisher

2.1. Removal

In January 2016, the Third Chamber of the Supreme Court of Justice of Chile ordered a media outlet to "eliminate the digital record of the news that adversely affected the appellant, within a period of three days."[23]. The information had been published ten years earlier, and referred to the participation of a citizen of unusual surname in a crime. Once the sentence was served, the citizen requested the elimination of the information arguing that the news was no longer relevant.

According to the Court, if criminal law indicates the duration of a penalty and allows its elimination from public records once it is complied with, it is

[23] Corte Suprema de Justicia de Chile [Supreme Court of Justice of Chile], judgment of January 21, 2016, retrieved from: https://bit.ly/2MbDW6m, last access: December 28, 2018.

consistent that the media do the same and thus allow the individual's social reintegration.[24] The Court argues that maintaining that digital registry does not report any benefit for freedom of expression and in any case the information "can be consulted by analogous methods through professional investigative exercise by whoever is interested in it". In any case, the Court clarifies that this is not about removing information from all digital records, but that access to it is limited to official sources and to the terms set forth in the law.

The agent in charge of implementing the order is the media outlet that originated the information. And that's what the Chilean media outlet did: it eliminated the news about the citizen's criminal conviction.[25] The elimination of information is the most radical measure of "digital oblivion". If it is applied where the data and information are stored — in the domain of the system — it impacts all other layers and affects all users. On the other hand, if the elimination applies only to public access, the information will remain available to those who move within the domain of the organization (that is, the information is maintained in a closed system).

Without overlooking the radical nature of the measure, it is important to bear in mind that as it is information that was part of an open system, the elimination in the core does not cover other downstream locations where it would have been replicated, such as a blog or a social network. These arenas exceed the control that the publisher or author initially had over the information. A judicial or state order cannot then expect the originator of the information to respond technically to these publications, with different actors in control. Furthermore, if the media outlet took the information from another publication — that is, if the domain of the original system is a different one — the act of elimination will only affect its republication orbit.

A complementary Colombian judicial precedent allows us to better understand the limitations inherent to the elimination order. The case originates with an unpaid debt. In 2014, "Esther" decided to publicly expose "Lucía" on Facebook for an unpaid loan.[26] In a common practice of online humiliation — known as online shaming — "Esther" published a photo of "Lucia" in which she denounced the debt and her refusal to respond to messages and

[24] Justice María Eugenia Sandoval voted against the decision, she stressed that the news dealt with a topic of public interest.

[25] From a technical perspective, the absolute suppression of information from a system is not as simple as it seems. Different methods have been proposed to ensure the irrevocability of the removal, ranging from the destruction of the physical medium and overwriting it to the encryption of the information with the subsequent elimination of the keys.

[26] The Court modified the names of the protagonists to preserve their privacy.

calls. After the creditor refused to withdraw the publication, "Lucía" filed an appeal for court protection known as "*tutela*". Although "Lucia" did not deny the existence of the debt, the Constitutional Court approved the request for protection. According to this court, the publication of this information on Facebook put "Lucía" in a state of defenselessness and violated her rights to privacy, good name, image and honor. Consequently, "Esther" had to eliminate the image, the message and publish a public apology.

The removal carried out by "Esther" affected the domain of the Facebook system: once deleted, these contents cannot be recovered or extracted by a third party. However, screen captures and information related to the debt that has been exchanged in other spaces — a very common practice in social networks — will not be affected by the elimination at the source and, therefore, will not be impacted by the court order.

2.2. Pseudonymization

In 2014, a Colombian citizen filed a court protection known as *amparo* against the Supreme Court of Justice for publishing on its website information about her related to a criminal conviction for the crimes of extortion, falsehood in a public document and procedural fraud. The claimant had already served her sentence, but this precedent was easily accessible by searching her name on Google.

The Constitutional Court accepted the citizen's claim. The Court argued that, although enforceable sentences are governed by the principle of publicity, they must be subject to the general framework of data management. The Court stated that maintaining certain data available for consultation on the Internet violates the principles of purpose limitation, of restricted access and circulation of data. To that extent, a parallel system of judicial background consultation was being created outside the existing controls on the matter. The Constitutional Court then ordered the Supreme Court of Justice to replace the name of that person with a succession of letters or numbers that would prevent her identification in the public versions of the sentence found on the Internet.[27]

This decision of the Colombian court is an order of pseudonymization, which consists of hiding the real name of a person, by changing it for another or by making only its initials public. An order of this type must be executed

[27] Corte Constitucional de Colombia [Constitutional Court of Colombia], judgment T-020, January 27, 2014. Judge writing for the court, Luis Guillermo Guerrero, retrieved from: https://bit.ly/2AAZ3as, last access: December 28, 2018.

by the author of the website, and would operate in the domain of the system if it is applied to all versions of the document. On the contrary, it would be limited to the domain of the organization if an integral version of the document is kept for its members, but a pseudonymous version is made public.

Pseudonymization already existed in the media as a form of protection of sources and subjects mentioned in newspaper articles. This self-regulatory measure, deployed voluntarily by the media from its operational capacity, could prevent an eventual request of "digital oblivion". By protecting an identity from the beginning, the media outlet decides to avoid exposure of an individual. Depending on how it is implemented, this measure could operate in the domain of the system or the organization.

Although this order was not addressed to a media outlet, it is relevant because of the type of precedent it establishes for journalistic activity. On the one hand, searching for information of public interest about a citizen becomes more difficult. In fact, after this ruling and similar requests, the Supreme Court generalized the rule: when it is proven that a sentence was served or has prescribed, the names of the people in open access databases must be suppressed.

According to the decision, "the condemnatory sentences issued by the Court or the files which references them (...) will be offered in full to the community on its public access server", with the defendant's full names. However, "when it is proven that the sentence has been served or has prescribed; the names of the convicted persons shall be removed from open access databases, except where the law requires that such information be kept public in all cases".[28] The Court maintains that this version will not affect the internal files of the entity, which may be physically consulted according to the rules of access to information. Following the classification proposed in this paper, it is a decision that affects the transactive-memory system, but not the domain of the system or the organization.

2.3. De-indexing (total and partial)

In August 2000, the Colombian newspaper El Tiempo published the article "*Empresa de trata de blancas*", where it reported that the citizen "Gloria" — as she was identified in the file — had been linked, among others, to a criminal investigation for the crime of human trafficking. In 2013,

[28] Corte Suprema de Justicia [Supreme Court of Justice], judgment August 19, 2015. Justice writing for the court: Patricia Salazar Cuéllar. Abstract of record retrieved from: https://bit.ly/2Rydi9Q, last access: December 28, 2018.

"Gloria" filed a request for a court protection known as *amparo* against El Tiempo and Google, since the process against her had ended by prescription and the news was still published — easily accessible with Google. In 2015, the Constitutional Court of Colombia heard the case.[29]

The high court ruled out two solutions before resolving the case. On the one hand, it released the search engine from any liability, thereby setting a precedent against the Costeja ruling. According to the Court, forcing Google to de-index search results would turn it into a content censor. This could jeopardize freedom of expression and information and affect the architecture of the Internet. On the other hand, and also from the perspective of freedom of expression, the Court considered that an order for the elimination of content did not conform to national and international standards.

It was clear then that the Court was going to center the responsibility on the content's author. However, the Court claimed that an update of the article (with the introduction of the clarification that the criminal action had prescribed in favor of "Gloria", for example) was insufficient to avoid stigmatization. Therefore, it ordered the media outlet that, in addition to the update, it had to take measures so that the search engine did not index that content when a search for the name "Gloria" was made.

It is a partial de-indexation order equal to that of the Costeja ruling, but placing the liability on a different agent. In this case, the court requires the media outlet to instruct the search engine not to index a specific website on its site when searching for a name. The effect should be the same as with the European precedent: the news should be hidden in the Google results if the claimant's name is searched ("Gloria Pérez", for example), but it must be shown for other terms related to the information ("human trafficking in Colombia "). However, that the liable actor is the media outlet and not the search engine has technical implications that were not foreseen by the Court.

The order issued by the Court is technically inaccurate. The ruling provides that the media outlet, "through the technical tool 'robots.txt', 'metatags' or another similar tool, should neutralize the possibility of free access to the article "*Empresa de trata de blancas'"* just by typing the plaintiff's name in

[29] Corte a de Colombia [Constitutional Court of Colombia], judgment T-277, May 12, 2015. Justice writing for the court: María Victoria Calle, retrieved from: https://bit.ly/1iQCR1b, last access: December 28, 2018.

the Internet search engines".[30] The problem is that the "robots.txt" protocol, the use of tags or another similar tool, do not allow partial de-indexing that excludes certain words from the content of a "link". The use of these protocols do allow for indexing instructions to the search engine, but in a different sense.

2.3.1. "Robots.txt" and "metatags"

The robot exclusion protocol — known as the "robots.txt" protocol — is the way in which a website informs search engine's crawlers which pages of their website someone does not want to be indexed. That they are not indexed implies, in principle, that they will not appear in the search results under any search term.[31] That is to say that the consequence is a total de-indexation.

The "robots.txt" protocol emerged as a technical response for those website administrators who wanted their content to have a certain degree of privacy without necessarily being secret. In the same way that two people can have a conversation in a public place that can eventually be heard, a web page that does not want to be indexed wants to remain relatively hidden, but not entirely private.[32] American scholar Jonathan Zittrain explains

[30] At this point, the Court accepted the search engine's argument: "In this regard, the Court takes note of the response given by Google Colombia Ltda., which states that through the use of tools such as techniques like 'robots.txt' 'and' metatags' it is possible for the owners and administrators of a website to prevent specific contents from being displayed as results when making a query through an Internet search engine. In relation to this issue, the claimant states that by using the tool robots.txt what is achieved is that certain content is not tracked by the search engine. However, despite the use of this tool, the search engine continues to recognize that the information exists and, therefore, may show the title of the article or URL in the search results, even when it could not be accessed because the content was not indexed. Similarly, in relation to the use of metatags, it states that '(...) what is achieved is that a certain URL, despite being indexed, is not shown as a search result'".

[31] We say "in principle" because the search engine could still include in its results pages that "robots.txt" indicated it not to include. This happens when many other third-party sites — which were indexed — include that link in their content, which the algorithm considers to be a relevant result. In this case, the link to the site in question will appear in the list of results, but without the description that usually accompanies it. With this measure it will not appear as a direct consequence of the content found on that site, but indirectly because it has been externally referenced on account of this content. Technically, it is not about "indexing" but about "listing". From Valk, Joost, "Preventing your Site from Being Indexed, the Right Way", June 2017, retrieved from: https://bit.ly/2AyaCiG, last access: December 28, 2018.

[32] According to Zittrain, the lesson of "robots.txt" is that the creation of a basic and simple standard can mean a significant advance in solving or anticipating a problem with important ethical and legal dimensions. Zittrain, Jonathan, "Privacy 2.0", in: University of Chicago Legal Forum, Vol. 2008, article 3, 2008, p. 104

that “robots.txt” is not an official standard: there is no guarantee that current or future crawlers will respect it.[33] However, in the self-regulatory and consensual logic of the Internet, the most popular search engines accept the non-indexing requests included in “robots.txt”.[34]

“Robots.txt” works through a text file that is inserted in the root of the website and which includes the web addresses (URL) that have to be de-indexed. The exclusion of the URL is complete: it is not possible, and this is important to reiterate, to define parts of a page that should be indexed and others that cannot be indexed. If we assume that the search engine is a large telephone directory, “robots.txt” only allows you to request that an apartment in a particular building not be listed, but it does not allow you to choose in which cases the apartment should appear and in which it should not.

“Robots.txt” page from nytimes.com. The “disallow” command instructs to not index a particular URL.

[33] *Ibid*.

[34] On the Google search help page, see “*Información sobre los archivos robots.tx*t [Information about robots.txt files]” in the section “*Cómo bloquear URLs con robots.txt* [How to block URLs with robots.txt]”, retrieved from: https://support.google.com/webmasters/answer/6062608, last access: December 28, 2018. In Yahoo, consult “*Elimina tu sitio web del* índice *de búsqueda* [Remove your website from the search index]”, retrieved from: https://es.ayuda.yahoo.com/kb/SLN2214.html?guccounter=1, last access: December 28, 2018. In Bing, “How to Create a Robots.txt File”, retrieved from: https://binged.it/2FcaCcc, last access: December 28, 2018.

On the other hand, "metatags" are labels (in a programming code known as HTML) that are used to record important information about a website: the description of the page, the keywords and the author's name, among others. This information is included in the page header code and not in the body of the page, so it is information invisible to the user. Using "metatags" and "robots" with values as "noindex" or "nofollow" can give directions by using labels or categories, for example, to prevent one or more URLs to be indexed or tracked by search engines.[35]

Both "robots.txt" and "metatags" do not allow compliance with the order of the Colombian court. Nor do other "similar" tools seem to exist — as the court suggests — to achieve partial de-indexation. In specialized circles, there is a mention of tags that allow the search engine to be instructed not to index some parts of a web page.[36] These tags, however, have so far not been recognized by Google nor are they present in the technical forum on the solutions. And, in any case, it is not about tools that allow someone to dictate to the search engine under what words it can list or index the content.[37]

Bearing this in mind, what was the solution in the Colombian case? Faced with an order that could not be met technically, El Tiempo opted simply to use a pseudonym for the claimant's name: "The Prosecutor's Office has just captured 16 people accused of committing the crime of human trafficking and conspiracy to commit a crime, including Ms. Gloria", reads a section of the new version of the text.[38]

[35] Google Webmaster Central Blog, "Using the robots meta tag", 2007, retrieved from: https://bit.ly/2AvMRHV, last access: December 28, 2018. Wikipedia, "Noindex", retrieved from: https://bit.ly/2CX37Uh, last access: December 28, 2018.

[36] Wikipedia, "Noindexing Part of a Page", retrieved from: https://bit.ly/2LVC9z9, last access: December 28, 2018.

[37] We cannot say categorically that there is no possible technical solution. The certain thing is that at this point and with the tools and knowledge within our reach we did not find an answer, which says a lot about how difficult a media outlet would find complying with an order of this nature.

[38] "*Empresa de trata de blancas*" [The human trafficking industry], *El Tiempo*, 2015, retrieved from: https://bit.ly/2AB6H4S, last access: December 28, 2018.

IV. Unwanted effects of technical ignorance

One of the theoretical questions faced by "digital oblivion" is the proportionality of its protection. In relation to the right to information, a measure is proportional when it "does not imply a cost for freedom of expression greater than the benefit achieved"[39] Therefore, the objection is that a de-indexation order aimed at protecting the privacy or reputation of a person may end up disproportionately affecting access to general interest information.

There is a risk that the implementation of a technical order will cause unforeseen effects, especially when this does not take into account the praxis of the digital environment and the context in which it must be executed. That is the problem that we identify in our region: the risk of not understanding the technical implications of an order can lead to disproportionate decisions or an incentive for self-censorship. Faced with contradictory or technically impossible decisions, a media outlet may opt, at best, to modify texts of the journalistic archive beyond the scope of the judicial mandate. And, at worst, it may simply prefer to delete the information.

The Colombian precedent of the "Gloria" case brings into the spotlight the problem of not taking into account the technical orbit. Similarly — but in a more worrying aspect — the Peruvian data protection authority has been making decisions about "digital oblivion" without an adequate technical basis. In a recent order against the newspaper El Comercio, the authority of Peru ordered the updating of the headline of an article, but it also seems to assume that it is the media outlet that carries out the indexing.[40]

In addition to the underlying problems of these decisions — which tend to impact corruption allegations that have not been judicially settled — the rulings are encouraging the media to choose to eliminate content to settle

[39] Botero, Camilleri and Cortés, *supra* note 6. This is the standard of the inter-American system according to which limitations on freedom of expression in a democratic society must not only be necessary, but also strictly proportional. According to the Inter-American Court of Human Rights, the sacrifice inherent in the limitation of freedom of expression must not be excessive compared to the advantages derived from the limitation of this right (Kimel v. Argentina case, judgment of May 2, 2008, § 83).

[40] It states: "Said update was placed below the article's headline, which causes the Internet search engine Google to render this headline as the first result of a search for the claimant's name, because "El Comercio" had indexed the article's headline". *Dirección de Protección de Datos Personales, Ministerio de Justicia y Derechos Humanos de Perú* [Directorate of Protection of Personal Data, Ministry of Justice and Human Rights of Peru], directorial resolution No. 453 of March 12, 2018.

a judicial controversy.[41] In the face of confusion and ambiguity, journalism finds a defense strategy in content subtraction.

The actions to guarantee "digital oblivion" implemented directly in the newsrooms should not be ruled out. There should be new alternatives that are less burdensome for freedom of expression and the journalistic exercise. One of them would be, for example, that the media outlet publishes both a partial — or pseudonymous — version of the article and a complete one. While the first would be indexed by the search engine, the second would include some non-indexing instruction.[42] That the full version is not indexed by the search engine would partially affect the transactive memory: the information cannot be found if the name of the person is searched, but by other related terms because it would refer to the partial or pseudonymous version. However, it would allow preserving it in the domains of the system and the organization of the media outlet (something which apparently did not happen in the final resolution of the "Gloria" case). In other words, the solution would not lead to the total removal of information or its unavailability for open access.

This type of solutions should be analyzed bearing in mind the reality of the newsrooms, since executing them would demand the reorganization of internal processes and a certain degree of progress in programming. They are not, of course, tasks impossible to carry out. A generalized order can generate high costs or entry barriers for the generation and dissemination of information.

---

[41] A Google search with the headline of the aforementioned case — ""*Renzito: ex asesor ministerial sería uno de sus testaferros*" [Renzito: former ministerial advisor is allegedly one of its front men]" — allows us to verify this. Additionally, it does not appear in the newspaper's digital file of the original publication date. See, https://bit.ly/2TwOpsn, last access: December 28, 2018. The information, nevertheless, continues to exist, since other portals copied it. This is an example of content that was in the domain of the transactive memory and that, therefore, subsists despite the removal. This type of solution is not the exception in the media. Last June, one of the authors of this paper spoke in Lima with lawyers and journalists from various media outlets regarding the matter. Several of them accepted that the removal of content has become a form of defense against administrative orders of "digital oblivion".

[42] These two versions could be used to produce a system of partial de-indexing performed by the media outlet that is less restrictive for freedom of expression. As was stated in the paper, the version indexed in Google would be one with partial or pseudonymous information. However, when someone clicked on that link, the page could redirect the user to the complete information that was not indexed by the search engine. These types of redirecting instructions are very common in Internet browsing. Thus, if someone searches for "Gloria" using her own name, the news will not appear in the Google results. But if someone searches for "human trafficking", the modified article would be indexed and, when clicked, would redirect the user to the complete information. This idea came from a conversation with the Colombian programmer David Avellaneda.

V. Conclusion

In addition to the already complex question of what is the "right to be forgotten" and what is its scope in the digital environment, we have the technical questions of how to guarantee this right and what would be its effects. This paper tried to approach the issue of its implementation.

Approaching this issue from the transactive memory and the information domains underlines two relevant analysis elements: (i) the closer the method of "digital oblivion" is to the storage core, the more domains will be affected and, therefore, the greater the restriction on freedom of expression and access to information, and (ii) in many cases "digital oblivion" encompasses a "memory" that is constructed collectively among many parties.

The technical approach offers elements to understand the proportionality of the concrete dimension of the impact of a judicial or administrative order, and allows us to understand that, in practice, there is no absolute solution for "digital oblivion". When establishing the role played by the State in promoting technological changes that respond to social expectations, judges and administrative authorities must consider this part of the discussion for decision making.

However, the participation of the State and civil society in this debate requires greater transparency on the part of Internet intermediaries and, in particular, search engines. Talking about the technical implications of a decision is often confused with the inevitability of a present configuration or the impossibility of promoting changes. However, the goal is to identify the starting point to find alternatives. In that process, the responsibility of Internet intermediaries in the design of the algorithms cannot be ignored: "While the algorithms represent calculations and processing that no human could do alone, ultimately humans are the arbiters of the inputs, the design of the system and the results".[43]

Adding the issue of the implementation of "digital oblivion" to the main discussion will also allow seeing it in the context of the impact of the measures taken to achieve technological oblivion and the need for them. That is to say, this will also enable a conversation around caution. "The question about memory and oblivion is not a question that must be solved by a specific date, but one that society must keep open for future generations to

[43] Caplan, Robyn et al. "Algorithmic Accountability: A Primer", in: *Data and Society*, April, 2018, p. 10.

decide according to their own circumstances."[44] In the midst of the fragile present and our countries' conflictive past, keeping this question open is even more relevant.

[44] Thouvenin, Florent et al. "Remembering and Forgetting in the Digital Age. A Position Paper", Open Repository and Archive, Zurich, University of Zurich, March, 2016, p. 5. This document is the basis of a book with the same name published in 2018.

Chapter Two

From comics to memes: old and new problems regarding humor and freedom of expression

Agustina Del Campo* y Paula Roko**

I. Introduction

In 2018, the "Save the meme" campaign gained global attention of the Internet community of human rights activists.[1] Through the slogan "Without memes, there is no democracy" many international organizations came together to collect signatures against the reform of the European directive on copyright for considering it a serious threat to freedom of expression.[2] The United Nations Special Rapporteur for Freedom of Expression also expressed concern about the negative consequences that would follow

* Agustina Del Campo is CELE's Director. She is a lawyer and has an LLM in International Law and Human Rights from American University Washington College of Law. She is professor on un-dergraduate and postgraduate courses in International Law and Internet & Human Rights.

** Paula Roko is a Researcher at CELE. She is a lawyer and a journalist. She specialized in Investi-gative Journalism and is currently pursuing a Master in Constitutional Law and Human Rights at Palermo University Law School.

*** Additional contributions by Carolina Botero, director of the Fundación Karisma in Colombia and Vladimir Cortés, Officer of the Digital Rights Program in Article 19 Mexico.

[1] "Save the meme" campaign, retrieved from: https://savethememe.net/es, last access: November 4, 2019.

[2] Europapress, "Una nueva campaña carga contra la reforma de la ley de copyright de la UE bajo el lema 'Sin memes, no hay democracia'"[A new campaign charges against the reform of the EU copyright law under the motto 'There is no democracy without memes], March 7, 2017, retrieved from: https://bit.ly/36TBYyr, last access: November 4, 2019; Cadena Ser, "'Sin memes no hay democracia', la nueva campaña contra la reforma de la ley de copyright de la UE" ["Without memes there is no democracy', the new campaign against the reform of EU copyright law], March 8, 2017, retrieved from: https://bit.ly/2X4hJJF, last access: November 4, 2019.

this rule and made a public plea to protect humorous expressions.[3] In June 2019, the New York Times chose to stop publishing political cartoons in its international edition after an accusation of anti-Semitism.[4] The problem was originated by a vignette in which the Israeli Prime Minister appeared as a guide dog leading a blind Donald Trump wearing a yarmulke, which generated a barrage of criticism.

It seems that this is not the best time for humor. Is this a limit for this type of expression? What are the causes of this restriction and how does it fit within a historical-legal perspective? This article aims to investigate the main conflicts between humor and other rights, many of which seem to arise or become more entrenched by the surge of "memes" on the Internet in recent years. First, this paper will explore the peculiarities of humorous discourse, from its most traditional conception to the most current meanings, along with the so-called "meme culture." Then we will analyze the impact of some recent measures, in the public and private sector, both of which with different purposes regulate the dissemination and circulation of content on the Internet and therefore have a great impact on the circulation of humorous discourse.

II. Humor and its peculiarities

Humorous discourse, especially political humor, has historically been conceived as an instrument of public accusation and social criticism in different and varied means of communication: journalism, literature, TV, cinema,

[3] In a document addressed to the European Commission at the time of the discussions before the approval of the Copyright Directive, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, expressed serious concerns about article 13. And he stressed that what further aggravates his concerns is that content filtering technologies are not equipped to make comprehensive interpretations of limits and exemptions to copyright, such as content of an educational, critical, or satire and parody nature. See Palais de Nations, "Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression," June 13, 2018, p. 7, retrieved from: https://bit.ly/2KccFhh, last access: November 4, 2019.

[4] Chappatté, Patrick, "The end of political cartoons at The New York Times," June 10, 2019, retrieved from: https://bit.ly/2q5W47Q, last access: November 4, 2019; Prieto, Ana, "El New York Times dejará de publicar caricaturas políticas, tras una acusación de antisemitismo" [The New York Times will stop publishing political cartoons, after an accusation of anti-Semitism], *Clarín*, June 11, 2019, retrieved from: https://bit.ly/2CBaLT2, last access: November 4, 2019.

music, theater, etc.[5] Defining what is humor is a complex task since it can vary according to the times, the theoretical trends and the authors who study it.[6] This has to do, in part, with the necessary distinction of concepts such as irony, comedy, satire, and parody. However, they all share certain semantic features and a communication purpose: to cause laughter.[7] On the other hand, this type of manifestations can take on different — and even antagonistic — meanings according to who interprets them and the context in which they appear: what causes someone to laugh may bother others. When and to what extent should certain expressions be tolerated is a question that seems difficult to answer.

In recent decades we have witnessed a huge growth of humorous discourse, largely driven by the Internet.[8] What until recently circulated mostly through

[5] Valero Heredia, Ana, "Libertad de expresión y sátira política: un estudio jurisprudencial" [Freedom of expression and political satire: a jurisprudential study], *Revista Internacional de Historia de la Comunicación*, No. 2, Vol. I, 2014, pp. 86-96, retrieved from: https://bit.ly/2wBK8lq, last access: November 4, 2019.

[6] Theofylakti Zavitsanou explains that different definitions of humor have been divided into three basic theories: the superiority theory, the relief theory, and the incongruous juxtaposition theory. In the first of them, laughter is seen as the expression of a feeling of superiority over other humans or circumstances regarded as inferior. Henri Bergson is considered one of the main exponents of this trend. The relief theory is based mainly on the works of Sigmund Freud and derives from the idea that humor serves to release the tension created by social inhibitions and restrictions in humans. Miguel Billig, Rod Martin, and Ofra Nero were also followers of this approach. Finally, the incongruous juxtaposition theory seeks to define the essence of humor and establish the conditions that allow its existence. According to this approach, humor relates disparate ideas in a way that "violates expectations" of the person receiving the message. Immanuel Kant and Arthur Schopenhauer are the main exponents of this theory. See, Zavitsanou, Theofylakti, "Humor y discurso político: el humor como recurso de opinión y crítica en la prensa contemporánea griega y española" [Humor and political discourse: humor as a resource for opinion and criticism in the contemporary Greek and Spanish press], Repositorio Universidad Pompeu Fabra, 2016, retrieved from: https://www.tdx.cat/handle/10803/385361, last access: November 9, 2019.

[7] Zavitsanou, *ibid*. For this article, we will adopt a generic concept of humor, although with special emphasis on satire and the problems it raises. Willibald Ruch, psychologist and professor at the University of Zurich, argues that this is the current trend that prevails internationally. See Ruch, Willibald, "Humo(u)r Research," article presented in the 14th Conferencia de la Sociedad Internacional de Estudios de Humor, Italy, 2002.

[8] According to Thomas Kilian, humor is one of the essential elements of digital communication since it satisfies almost all of people's Internet needs. The author summarizes these needs as follows: (i) informative: to seek information and advice, to satisfy a sense of curiosity; (ii) social integration: to gain a sense of belonging and social ties, to connect with friends, family, and social circles; (iii) personal identity; (iv) entertainment: to relax, to emotionally unwind. The fact that Internet users are constantly in touch with humorous content — especially through "memes" —, along with the phenomenon of mass viralization, has made them true producers of humorous content. See Kilian, Thomas, Hennigs, Nadine and Langner, Sascha, "Do millennials read books or blogs? Introducing

magazines and newspapers now is also disseminated in the form of "memes" generated and viralized by the users in social media and instant messaging services. The term "meme" was born in 1976 in a publication by British biologist Richard Dawkins, who combined the English words "mimesis" and "gene" and understood memes as gene-like culture units.[9] Currently, the word has many meanings.[10] Some argue that the main characteristic of the meme is its viral nature. Others add that, when it comes to content subject to copyright, the content must have a substantial modification concerning the original to be considered a meme. This modification is interpreted by some as literal (for example, photo manipulation), others admit that it can refer to the meaning or interpretation that the work acquires in a new context or format (for example, the case of "Pepe the frog," which we will refer to later).[11]

For several years now, the area of study on the singularities of "memes" and

a media usage typology of the internet generation," *Journal of Consumer Marketing*, Emerald Group Publishing Limited, No. 2, Vol. 29, 2002, pp. 114-124, retrieved from: https://bit.ly/2XYh6nT, last access: November 4, 2019. There is also the argument that the proliferation of humorous discourse on the Internet has to do with the fact that a large part of people's social interaction happens in social media. See Krotoski, Alex, "What effect has the internet had on comedy?" *The Guardian*, April 3, 2011, retrieved from: https://bit.ly/2JDws8i, last access: November 4, 2019.

[9] This term, which arises from the combination of the words *mimesis* and *gene*, was coined in 1976 by Richard Dawkins, British biologist, in his book *The selfish gene*. Dawkins argued that "memes" are units of information found in the brain which can spread on their own and control human behavior. The meme was understood, then, as a unit of culture similar to a gene. Those who ascribe to this theory consider memetics to be an approximation to the evolutionary models of cultural information transfer. See Nooney, Laine and Portwood-Stacer, Laura, "One does not simply: an introduction to the special issue on internet memes," *Journal of Visual Cultures*, No. 3, Vol. 13, December 16, 2014, pp. 248-252, retrieved from: https://bit.ly/2XKJOES, last access: November 4, 2019; MDX Online, "La cultura del meme"[Meme culture], December 3, 2018, retrieved from: https://bit.ly/2ZSZqr3, last access: November 4, 2019; Dean, Jonathan, "Sorted for memes and gifs: visual media and everyday digital politics," *Political Studies Review*, October 25, 2018, retrieved from: https://bit.ly/2Gd2qY1, last access: November 4, 2019.

[10] Lantagne, Stacey, "Famous on the Internet: the spectrum of internet memes and the legal challenge of evolving methods of communication," *University of Richmond Law Review*, November 27, 2017, pp. 387-424, retrieved from: https://bit.ly/36UXxyv, last access: November 9, 2019. See also Jackson Johnson, Shontavia, "Memetic theory, trademarks & the viral meme mark," *The John Marshall Review of Intellectual Property Law*, No. 13, Vol. 96, 2013, pp. 104,106, retrieved from: https://bit.ly/2Q7IVWH, last access: November 4, 2019.

[11] Lantagne, *op. cit*. The author points to two specific examples of American culture; see also Carpenter, Julia, "Meme librarian is a real job, and it's the best one on the internet," *The Washington Post*, December 2, 2015, retrieved from: https://wapo.st/2O0thK6, last access: November 4, 2019; Shifman, Limor, *Memes in the digital culture*, Cambridge, MA, The MIT Press, 2013; Madison, "What makes a meme go viral?" Medium, January 22, 2018, retrieved from: https://bit.ly/2CxzGHd, last access: November 4, 2019.

the dynamics in which they function has extended to linguistics and sociology, which is currently known as the "meme culture."[12] However, in the field of law, there are still few studies in this area. The dramatic increase in Internet memes and their unprecedented discursive nature have introduced some gray areas in the law. On the one hand, the lack of clarity in its definition makes the concept all-embracing. But, additionally, due to the great variety of nuances, law concepts that could traditionally be applied to some of them are difficult to adapt to the peculiarities of others. As Professor Stacey Lantagne points out, the theory of fair use of copyright, for example, seems easier to apply to a "static" meme — one in which the image did not undergo changes in its physiognomy nor in its meaning when reproduced — than in a "mutating" meme whose aspect or meaning has evolved when adopted as a "meme."[13] On this last point, some authors identify differences between "static uses" and "mutating uses," which directly affects the legal arguments that can be made in their defense.[14]

III. Humor in times of social media: old and new controversies

In general, humorous expressions have comparatively always enjoyed high levels of legal protection since they play a crucial role in keeping under control abuses of power and in the formation of public opinion. However, certain limits have also been set, which are permanently in dispute, particularly rights regarding personality, reputation and honor, copyright, or the crime known as *apologia* [TN the defense or praise of criminal acts] in some countries. These restrictions, which could be identified as traditional, have recently been joined by others such as disinformation and discriminatory and politically incorrect discourses, mainly motivated by the assumption that their widespread presence on the Internet could have harmful social effects.

[12] Zittrain, Jonathan L., "Reflections on internet culture," *Journals of Visual Culture*, No. 3, Vol. 13, 2014, pp. 388-394, retrieved from: https://bit.ly/2NSqeGM, last access: November 4, 2019; Daneci, Marsel, *Understanding media semiotics*, London, Bloomsbury Academic, 2nd ed., 2018; Wiggins, Bradley E., *The discursive power of memes in digital culture: ideology, semiotics, and intertextuality*, Abingdon, Routledge, 2019.

[13] Lantagne, *op. cit*.

[14] See Lantagne, *ibid*. The author points out that "static uses" are those in which a mere reproduction of an image is made, without altering it or providing it with new meanings. While the image may have been slightly modified, in essence, there is no added value to the original, it is simply reproduced with its original meaning and context. In contrast, "mutating uses" are those that drastically transform the original image to give it its own meaning. A mutating meme is any new creation imbued with a new meaning that exceeds its original one, also driven by the collaborative creativity of various Internet communities that add their characteristics.

In this context, new rules emerged intended to solve the problems of the digital world, and they started to suppress humorous expressions. These rules vary not only in their character but also in their objectives. Next, we will address some of the most relevant measures that both states and private companies (platforms) are taking and the impact they can have on the right to freedom of expression.

1. Protections to image and honor

The rights to honor and reputation have historically been in tension with freedom of expression and freedom of the press, humor in particular, and especially political humor. The study of judicial precedents in this matter is rich and extensive. In the United States, for example, satire and parody enjoy comprehensive judicial protection. One of the ground-breaking cases regarding satire was "Hustler Magazine, Inc. v. Falwell,"[15] where the Court decided that Hustler magazine was not liable for parodying public figures. The case arose when the magazine published a parody that mimicked a Campari campaign under the title "Jerry Falwell talks about his first time." The parody, part of a series, consisted of a false interview in which Falwell, a religious and renowned political commentator, reported that his "first time" had been an incestuous encounter with his mother in a drunken state. At the bottom of the page and in the index of the magazine it was stated that it was a parody.

[15] Supreme Court of the United States, "Hustler Magazine, Inc. v. Falwell," 485 U.S. 46, 1988, retrieved from: https://bit.ly/2CxVMcP, last access: November 4, 2019.

In its ruling, the Supreme Court argued that the case was not about true or false statements, but about alleged damage as a result of the publication of a caricature whose mission was clearly to satirize or distort reality. They argued that the United States Constitution prohibits public figures and public officials from receiving compensation for intentional psychological distress by a caricature if it cannot be demonstrated that the publication contains a false claim made with actual malice and that even this element individually is not enough to attribute liability, since "graphic depictions and satirical cartoons have played a prominent role in public and political debate (…). From a historical perspective, it is clear that our political discourse would have been considerably poorer without them."[16] In this case, as in general in the United States cases on the subject, the analysis is focused on the distinction between public figures and private persons and gives the former a lower threshold of protection than private persons. Undoubtedly, this approach is in line with the American meaning of freedom of speech. This distinction is also made by many other courts in cases linked to effects on image and honor caused by humorous expressions, including many of the courts in Latin America.

In Europe, the case of Charlie Hebdo probably received the most attention in recent years. In 2006, the French weekly publication launched a special issue that showed on its cover a cartoon of the Prophet Muhammad covering his face with his hands saying: "It's hard to be loved by jerks," under the title "Muhammad, overwhelmed by the fundamentalists." Additionally, the magazine reproduced in its pages the drawings of the Danish newspaper *Jyllands-Posten*,[17] including one of the most controversial ones in which Muhammad appeared with a bomb-shaped turban. In a few hours, more than 400,000 copies were sold, which tripled the usual sales. "This success proves the interest that people have on their freedom. It is a citizen response," said the then director of the magazine.[18]

[16] *Ibid*.

[17] The cartoons had originally been published in 2005 by the Danish newspaper Jyllands-Posten and provoked passionate protests around Asia, Africa and the Middle East that left at least fifty people dead. Many European media outlets republished these cartoons as a way of defense and to demand respect for freedom of speech. See De Andrés, Francisco, "La ruta mortal de las caricaturas" [The deadly route of cartoons], *ABC Internacional*, January 8, 2015, retrieved from: https://bit.ly/2KJx7V5, last access: November 4, 2019.

[18] "Chirac tacha de 'provocación' el número especial con caricaturas de Mahoma de una revista francesa" [Chirac labeled the special issue with Muhammad cartoons by a French magazine as 'provocation'], *El Mundo*, February 9, 2006, retrieved from: https://bit.ly/32Db2PS, last access: November 4, 2019.

The Union of Islamic Organizations of France and the Grand Mosque of Paris denounced three cartoons: the one that appeared on the cover and two others that had been reproduced by the magazine but were originals of the newspaper *Jyllands-Posten*.[19] The two Muslim entities called for the recognition of the crime of insults for religious reasons, with sentences of up to six months in jail and fines.

In 2007, the *Tribunal Corrrectionel* of Paris acquitted the director of Charlie Hebdo, among other reasons, because it was an eminently satirical publication.[20] They pointed out that nobody is under the obligation of buying or reading this type of magazines, which differentiates it from other media, such as public road signs. They further emphasized that the ultimate goal of the protection of freedom of expression also includes those speeches that

[19] In one of them, Muhammad is seen with a turban from which the wick of a bomb comes out, the other depicts the prophet asking terrorists to "not immolate themselves because there are no more virgins left in paradise." This has to do with the more traditional doctrine of Islam that teaches their faithful that if they immolate themselves and give their lives for Allah, they will receive 72 virgin maidens in paradise as a special reward. Women, on the other hand, will receive only one man "with whom they will be satisfied."

[20] The synthesis of the main arguments was extracted from Noorlander, Peter, "When satire incites hatred: Charlie Hebdo and the freedom of expression debate," Center for Media, Data and Society, February 2, 2015, retrieved from: https://bit.ly/2lPo3fv, last access: November 4, 2019; and Fourlon, Armelle, "The Charlie Hebdo case: freedom of expression and respect for religious faith," International Law Office (ILO), June 26, 2008, retrieved from: https://bit.ly/2z5Yoj2, last access: November 4, 2019.

may offend or shock some people. As for the cartoons, the court adopted the classical position about the public interest at stake and argued that they did not constitute defamation or a personal and direct attack against a group of people for religious reasons. It is interesting to note the emphasis of the sentence regarding the context in which the cartoons should be analyzed, which in this case was a special issue dedicated to religious fundamentalism. In general, the ruling did not come as a surprise since the prosecution itself had requested the acquittal because they considered that the cartoons did not attack Islam but fundamentalists.[21] Some analysts commented on the extensive tradition of the French courts regarding the separation of Church and State, as well as their repeated court decisions in favor of freedom of speech in the face of claims based on respect for religion.[22]

This case is one of the latest global landmarks in terms of satire and it has seeped deeply into legal precedents and doctrine in the world, largely due to the attack on Charlie Hebdo's editorial office years later (an issue that has been explored in length and which exceeds the framework of this analysis). The infamous "Muhammad cartoons" were later replicated by other media worldwide,[23] in some cases with legal consequences.[24]

The European Court of Human Rights, for its part, also ruled on this issue in the case "Vereinigung Bildender Künstler v. Austria" in 2007.[25] This case originated with the exhibition entitled "The century of artistic freedom" in a prominent independent gallery in Vienna. One of the paintings exhibited consisted of a collage of public figures immersed in sexual activities, including Mother Teresa of Calcutta, an Austrian cardinal and

[21] Marti Font, José María, "Un tribunal francés absuelve a la revista que publicó las caricaturas de Mahoma" [A French court acquits the magazine that published Muhammad's cartoons], *El País*, March 23, 2007, retrieved from: https://bit.ly/2MFuvYP, last access: November 4, 2019.

[22] Leveque, Thierry, "French Court clears weekly in Mohammad cartoon row," *Reuters*, March 22, 2007, retrieved from: https://reut.rs/2IS6ZFA, last access: November 4, 2019.

[23] In the United States, for example, some of these cartoons appeared in the *Philadelphia Inquirer* newspaper and the *Philadelphia Jewish Voice*, as well as being shown on TV shows and university newsletters such as at Harvard University. See Freedman, Leonard, *The offensive art: political satire and its censorship around the world from Beerbohm to Borat*, Santa Barbara, CA, Praeger, 2008, p. 53 and 54.

[24] See Mourenza, Andrés, "Cárcel para dos periodistas turcos por reproducir una caricatura de Mahoma de 'Charlie Hebdo'" [Jail for two Turkish journalists for reproducing a cartoon of Muhammad from 'Charlie Hebdo], *El País*, April 28, 2016, retrieved from: https://bit.ly/2KFgGsN, last access: November 4, 2019.

[25] European Court of Human Rights, retrieved from: https://bit.ly/33Bdd80, last access: November 4, 2019.

renowned representatives of the Austrian Liberal Party. The painting caused controversies and incidents and was even damaged by a visitor. The former secretary-general of the Liberal Party requested a ban on the exhibition of the painting based on the Austrian Copyright Law, which extends this right not only to the photographer but to the person portrayed, and compensation. Regardless of the legal framework, the national courts decided the case in light of the claimant's personal and non-transferable rights and held that the work constituted a defamatory action. The European court declared that "it is the common understanding of the national courts in all instances that the paintings in question were not intended to reflect or suggest reality" and, therefore, "these were caricatures that used satirical elements"[26]. The court described satire as a form of artistic expression and social criticism that, due to its inherent characteristics of exaggeration and distortion of reality, is destined to "provoke and agitate." Therefore, any interference with the right of an artist to such an expression must be examined with strict and special care.

In Latin America, the Inter-American Court of Human Rights has no cases related to humor. However, the case "The last temptation of Christ v. Chile"[27] assesses a group's right to honor against artistic expression. The case in question was a result of the remedy for protection filed by a group of people on behalf of and representing Jesus Christ who requested the Consejo de Calificación Cinematográfica in Chile [Cinematographic Classification Council][28] to ban the distribution of the film *The Last Temptation of Christ* for considering it contrary to honor and the reputation of Jesus Christ and contrary to religious freedom. The Council accepted the request and prohibited the film's exhibition, and the Chilean Supreme Court confirmed such a ban. In its judgment, the Inter-American Court stated that freedom of expression is the cornerstone of a democratic society and that it protects those expressions that are shocking. The Court concluded that in the case there was prior censorship incompatible with the American Convention, and ordered the reform of the Chilean Constitution. In addition to this case, the Inter-American Court has several cases linked to honor and reputation as possible limits to freedom of expression, where it adopted the legal principle that distinguishes between public and private figures

[26] *Ibid*.

[27] I/A Court H.R, case "La última tentación de Cristo (Olmedo Bustos y Otros) c. Chile" ["The Last Temptation of Christ" (Olmedo-Bustos et al.) v. Chile], Judgment of February 5, 2001, Series C, No. 73.

[28] Authority constitutionally responsible for the evaluation and approval or disapproval of exhibition of cinematography in Chile.

and information of public interest, in line with the main Supreme Courts in Latin America.[29]

There are interesting cases in Argentina and Colombia regarding humor and freedom of expression, even recent ones. The case of Cecilia Pando (2017) in Argentina resulted from a photomontage published by the magazine Barcelona in its back cover showing a naked body tied with ropes bearing the face of Pando, an activist for the rights of people convicted of crimes against humanity in Argentina. Pando filed a civil suit for damages and considered that her honor and image were injured.

The first ruling was favorable to the plaintiff and in 2017 the National Federal Court of Appeals for Civil Matters confirmed the sentence.[30] The court decided that the legal principle of actual malice was not appropriate since it was not a publication of false or inaccurate news, but rather a parody of Pando through an altered image and with titles that exceeded the limits of

[29] I/A Court H.R, "Herrera Ulloa v. Costa Rica," Judgment of July 2, 2004, Preliminary exceptions, Merits, Reparations, and Costs; I/A Court H.R, "Ricardo Canese v. Paraguay," Judgment of August 31, 2004, Merits, Reparations, and Costs; I/A Court H.R, "Kimel v. Argentina," Judgment of May 2, 2008; among others.

[30] Cámara Nacional de Apelaciones en lo Civil [National Federal Court of Appeals for Civil Matters], Courtroom D, "Pando de Mercado, María Cecilia v. Gente Grossa SRL on damages," March 23, 2017, retrieved from: https://bit.ly/2lN0H0Z, last access: November 4, 2019. The sentence ordered the magazine to pay the sum of 70 thousand Argentine pesos as compensation.

the press. The court declared that it was an imprudent exercise of freedom of expression, to the point of affecting personal and non-transferable rights.

The sentence was appealed and is pending resolution. In 2018, the public prosecutor before the Supreme Court of Justice of Argentina issued his opinion on the case and declared that the Supreme Court should reverse the decision of the Court of Appeals.[31] He pointed out that this was a satire on a topic of public interest about a public figure, so there should be greater tolerance for criticism. He emphasized that social or political satire is specially protected by freedom of expression because it enriches the public debate and that in Argentina there is a long tradition of it.

In Colombia, there was a very famous case when a cartoonist, César González, better known as "Matador," published in the newspaper El Tiempo a cartoon of President Iván Duque. The drawing depicts the then-presidential candidate turned into a pig, with a speech bubble that reads: "Oh, no! I am the only *uribista* who has not become a 'pig'."

A Colombian lawyer sympathetic to the Centro Democrático party filed a writ for the protection of constitutional rights against Editorial El Tiempo and the cartoonist, claiming it constituted an offense and a threat to the image of the political party's candidates. He claimed that the cartoon was intended

[31] Opinion of the public prosecutor before the Supreme Court, "Pando v. Gente Grossa SRL on damages," February 20, 2018, retrieved from: https://bit.ly/2z5Frgn, last access: November 4, 2019.

to influence "arbitrarily, and under the pretext of making a humorous or jocular manifestation, the freedom of conscience of Colombians and their right to exercise their power to choose without interference, manipulation or pressure from the media."[32] The judge deciding on the case denied the writ.[33] She noted that it was "a type of censorship" and that "a cartoon of a candidate published in a newspaper cannot qualify as a violation of fundamental rights, since qualifying it, or limiting it, or diminishing its critical potential, would infringe on the right of readers to the pluralism of ideas, opinions, and would be contrary to the social rule of law and therefore to democracy."[34] Furthermore, she mentioned that the creativity and ingenuity of the cartoonist cannot be limited by people who might not like the cartoon or the point of view, and that the cartoon only intended to express an opinion on a current issue in a context of impending presidential elections, from which favorable or unfavorable criticism can be drawn without injury to honor.

In Ecuador, there are multiple cases of cartoons addressed to former president Rafael Correa, who has publicly considered this type of criticism as "part of a systematic irony-filled smear campaign."[35] One of the stories that stands out is that of Bonil, a cartoonist for the newspaper El Universo. In 2013, Bonil published a vignette about the raid on the house of Fernando Villavicencio, who had been accused by Correa of hacking his emails, those of Vice President

[32] The judge was asked to order a public apology in the media, to protect "the fundamental rights to a good name, freedom of conscience and the freedom to choose of the members of the Centro Democrático party (known as 'uribistas') and of Colombian voters in general." See Judges Opinion "Tutela contra el caricaturista Matador: libertad de expresión en época electoral" [Writ for the protection of constitutional rights against the cartoonist Matador: freedom of expression in electoral time], *El Espectador*, March 8, 2018, retrieved from: https://bit.ly/2Hk4ADk, last access: November 4, 2019.

33 Some civil society organizations such as the Fundación para la Libertad de Prensa (FLIP) and the Centro de Estudios de Derecho, Justicia y Sociedad (Dejusticia) submitted writings to assist in the process of the writ for the protection of constitutional rights and argued in defense of the cartoonist. See texts written by FLIP, March 10, 2018, retrieved from: https://bit.ly/2GANv9f, last access: November 4, 2019; and Dejusticia, March 12, 2018, retrieved from: https://bit.ly/2SltNnM, last access: November 4, 2019.

[34] 4 Civil Circuit Court of Enforcement of Judgments, Bogota, March 20, 2018. Writ for the protection of constitutional rights filed by José Luis Reyes Villamizar against Julio César Quiceno "Matador" and Casa Editorial El Tiempo-CEET S.A.

[35] "Suspenden por varias horas cuenta de Twitter de página crítica de Correa" [The Twitter account of a website which criticizes Correa is suspended for hours], *Panam Post*, January 29, 2015, retrieved from: https://bit.ly/2rrJSyD, last access: November 11, 2019; Fundamedios, "Fundamedios condena censura de Twitter tras suspensión de cuenta de Crudo en Ecuador" [Fundamedios condemns Twitter 's censorship after suspension of Crudo's account in Ecuador], August 11, 2017, retrieved from: https://bit.ly/2q4bNnV, last access: November 11, 2019.

Jorge Glas and those of the former legal secretary of the Presidency, Alexis Mera. The graphics show armed and uniformed characters and the message "Prosecutor and Police raid the house of Fernando Villavicencio and find allegations of corruption." Following the publication, the Superintendencia de la Información y Comunicación (Supercom, the governing body) sanctioned Bonilla and Diario El Universo. The cartoonist was given 72 hours to recant himself, while the newspaper received a fine equivalent to 2% of the average turnover of the previous three months.[36] Almost four years later, the 2nd District Court on Administrative Matters of Ecuador rendered the resolution issued by Supercom void. The Court argued that "what has been sanctioned is a matter of opinion" and added that "the cartoon suggests a critical position on a previous source (...)." The ruling argues that as a value judgment, the cartoon is not liable to be true and is not and cannot be legally proven or reprehensible.[37]

Undoubtedly, the tension between honor and reputation with freedom of expression is an issue that, despite being addressed in comparative jurisprudence, remains to be solved by the court. Interestingly, the cases allow us to take a glimpse into the diversity of meanings of honor and reputation. In some cases they are described as personal and non-transferable rights, in others, they are regarded as violations of collective rights, such as the honor of a group (be it religious or another kind).

Although many of the examples given are from print publications or analog media, the cases that address digital expressions follow the same line. We should mention, for example, the case of Ali Ziggi Mosslmani of Australia, based in the Sydney District Court.[38] Ziggi, as he is known, is an 18-year-old teenager who acquired worldwide fame on the Internet in 2016 after a photo of him went viral in which he sports an extravagant hairstyle. The photo was posted and circulated with different captions referring to the haircut and various puns. He decided to sue three Australian media outlets for the inclusion of the picture in their print and online editions. The District Court judge dismissed the case against the online publication and only

[36] See article 25 of the Ley Orgánica de Comunicación de Ecuador [Organic Law of Communication of Ecuador]. According to this law, the media outlet is forced to take an institutional stance regarding the innocence or guilt of Xavier Bonilla.

[37] "Jueces declaran nula actuación de Supercom sobre caricature de Bonil" [The Court renders void the Supercom action on Bonil's caricature], *El Universo*, August 23, 2017, retrieved from: https://bit.ly/2Q8PxUP, last access: November 11, 2019.

[38] Whitbourn, Michaela, "Mullet defamation case gets a haircut as Sydney teen Ali Ziggi Mosslmani suffers setback," *The Sidney Morning Herald*, December 18, 2016, retrieved from: https://bit.ly/2qEo2Yz, last access: November 4, 2019. See decision, "District Court New South Wales," 2016, retrieved from: https://bit.ly/36TlTbZ, last access: November 4, 2019.

allowed the claim against the print edition of one of the media outlets. She argued that "To attempt to draw an imputation of stupidity from one or more of these [articles] when they are all humorous adaptations of the plaintiff's hairstyles with no deeper meaning is an exercise in futility."[39]

In Argentina, the senator and current candidate for vice president Miguel Ángel Pichetto reported in 2018 to the Federal Justice that he had suffered "personal, moral and political damage" by a Twitter account that used his image and his name as a parody. The account, which was described as "unofficial" in the biography, published alleged phrases of the legislator that users quickly reposted as jokes. Although the legal action was not on the grounds of slander and insult but, on the contrary, it was based on the crime of "identity theft," it is another form of attack and restriction of humorous expressions on the Internet to protect the image and honor of a public official. Furthermore, Senator Pichetto argued that in addition to the individual harm was "the deceit (...) suffered in good faith by each one of his followers." In July 2019, the judge decided to dismiss the charges against the creator of the fake account since identity theft in social media does not fit into any legal description of a crime in the Argentine Criminal Code.[40] Nevertheless, the issue continues to be present on the public agenda because it led to various bills to criminalize digital "identity theft" or "impersonation," essentially arguing that the use of someone else's identity in social media could be the prelude to crimes such as fraud, "slander and insult" or child abuse.[41]

As we can see in these cases, humor enjoys wide protection against defamation or insult. Unlike other issues, there is a consensus that protection and balance in these cases must necessarily be at a judicial level. Thus, in most Latin American countries and the United States, the law rules out the strict liability of the platforms for defamatory content. And the terms and conditions of service of large platforms do not provide for self-regulation measures for defamation,[42] although a lot of private persons and public officials insist that they should be included.

---

[39] In its original wording: "To attempt to draw an imputation of stupidity from one or more of these [articles] when they are all humorous adaptations of the plaintiff's hairstyles with no deeper meaning is an exercise in futility." Withbourn, Michaela, *op.cit*.

[40] Argentine Judiciary Branch, 4th Federal Court for Criminal and Correctional Matters, March 27, 2018, retrieved from: https://bit.ly/2Czo0DP, last access: November 4, 2019.

[41] See CELE, "Argentina Proyecto de Ley Usurpación Digital (2449/18)" [Argentina Digital Theft Bill of Law (2449/18)], 2018, retrieved from: https://bit.ly/36TGlnl, last access: November 4, 2019.

[42] For example the terms of service of platforms such as Facebook, YouTube or Twitter.

2. Protection of copyright

Another of the traditional reasons why humor has been in tension with freedom of expression is the use of texts, images, music and, in general, audiovisual content subject to copyright without proper authorization. These laws grant the owners of literary, musical, scientific or artistic creations a series of powers or rights to promote artistic creation and protect the authors' economic rights over their works and their reproductions. Notwithstanding this, comparative legislation establishes limitations on the individual rights of the owners of the works to make them compatible with the right of access to knowledge and information, uses for educational purposes, and even parody and satire, among other primarily collective interests.[43]

In many laws of the world, satire and parody are considered exemptions to the exclusivity of use proposed by copyright. With an intrinsically educational and cultural value, in general, it is considered that the works subject to satire and parody can be reproduced without licenses or authorizations. This exemption has different names, such as the legal principle known as fair use.

The distinction between satire or parody and abusive uses of copyrighted material is not always clear. Quite the contrary, the limits are permanently being debated. Unlike defamation cases, in these cases, Internet companies have had greater prominence than the courts. There are documented cases of uses and abuses of copyright protection policy to limit satire, parody and humorous criticism in general.[44] Eduardo Bertoni and Sophia Sadinsky mention the case of the Tourism Office in Alberta, Canada, which issued an injunction against YouTube to take down a satirical video that used part of a commercial to criticize environmental degradation in the city. Also in Canada, the Canada Post requested YouTube to remove a video broadcast by members of a union in which they made fun of their executive director

[43] In Colombia, for example, Law 23 on Copyright (1982) contemplates a series of exemptions and limitations that allow the use of protected material without authorization in cases such as quotes (art. 31), illustration in works intended for teaching (art. 32) or taking notes during conferences or lessons in educational establishments (art. 39). Also, in Costa Rica, Law No. 6,683 (1982), then amended by Law No. 8,686 (2008) and in Mexico, the Federal Copyright Law provide for fair use exemptions, just to name a few examples.

[44] Bertoni, Eduardo and Sadinsky, Sophia, "El uso de la DMCA para limitar la libertad de expresión" [The use of DMCA to limit freedom of expression], *Internet y Derechos Humanos II*, Buenos Aires, CELE, Universidad de Palermo, 2016. See also Keller, Daphne, "Empirical evidence of 'over-removal' by Internet companies under intermediary liability laws," Center for Internet and Society, Stanford University, October 12, 2015, retrieved from: https://stanford.io/2fBMNhk, last access: November 4, 2019.

and her corporate policies, by displaying a retouched photograph of her.[45]

Globally, some countries without the fair-use exemption recently incorporated it. For example, in 2017 in New Zealand, a bill of law was brought before Parliament to extend existing exemptions to copyright infringements and include satire and parody in cases of art criticism and reviews.[46] This debate also took place in Australia, and fair use exemptions were finally included in 2006.[47]

In Latin America, Mexico amended its federal law in 2018 and implemented precautionary measures in cases of use of copyrighted material without the owner's consent. Due to its scope and vague wording, the law leaves open the possibility for memes to be subject to copyright. The law passed despite having been severely criticized by organizations dedicated to the promotion of human rights in the digital space.[48]

Since 2012, Colombia has also tried to modify and update copyright laws, but so far the attempts have not been successful. The last bill submitted to Congress, and still held in the Senate, is the well-known "Ley Lleras 6."[49] This bill includes the exemption for parody, an unresolved issue according to activists, since the previous bills maintained as a rule the prohibition of reproduction of protected works and did not establish any exemptions for criticism, parody or caricature.[50]

Recently, the discussion about satire, parody and copyright became more heated following the proposal and approval of the European directive on copyright,[51] and in Latin America due to the impact that the directive could have on companies and users in this region.[52] This rule modifies the 2001

[45] Bertoni and Sadinsky, *op. cit.*, p. 63.

[46] New Zealand Parliament, "Copyright (Parody and Satire) Amendment Bill," retrieved from: https://bit.ly/2JAQWQh, last access: November 4, 2019. See especially section 42

[47] Austin, Graeme, "A copyright exemption for parody and satire," *Newsroom*, May 11, 2018, retrieved from: https://bit.ly/2GcrAWG, last access: November 4, 2019.

[48] Sinembargo.MX, "Nueva ley censura hasta a los memes" [New law censors even memes], *Noroeste*, April 30, 2018, retrieved from: https://bit.ly/2X4H46l, last access: November 4, 2019; Notimex, "Por esta razón podrían demandarte si usás memes" [This is why you could be sued for using memes], *Milenio*, March 6, 2019, retrieved from: https://bit.ly/2Y2Ft41, last access: November 4, 2019.

[49] Law No. 206 from 2018.

[50] Botero, Carolina, "Una de las Ley Lleras por fin será Ley y reformará el derecho de autor en Colombia" [One of the Lleras Laws will finally become Law and will reform copyright in Colombia], Fundación Karisma, May 24, 2018, retrieved from: https://bit.ly/2qFl6ts, last access: November 4, 2019.

[51] Eturopean Union, directive No. 2019/790 of the European Parliament and the Council on copyright and related rights in the Digital Single Market and amending Directives No. 96/9/CE and No. 2001/29/CE, April 17, 2019, retrieved in Spanish from: https://bit.ly/30D1qEb, last access: November 4, 2019.

[52] CELE, "La Directiva Europea de Derecho de Autor y su impacto en los usuarios de América Latina y el Caribe: una perspectiva desde las organizaciones de la sociedad

directive[53] and updates the intellectual property legal framework. Although it is not yet in force and there is a two-year period for the different member states of the European Union to amend their legislation, its wording has caused extensive and severe criticisms.[54] Article 17, one of the most controversial ones, requires platforms that offer services for publishing and exchanging content on the Internet (such as YouTube, Twitter or Facebook) to monitor and take all the proactive measures necessary to prevent copyrighted content to be hosted on their servers and shared by users. A large part of the community drastically opposed the policy, claiming primarily that such wording promotes active monitoring and content filtering, including pre-upload filtering (in other words, filtering the content before it is published).[55] Regarding filters, Wikipedia, one of the most active parts in this debate, stated that:

> The world should worry about new proposals to introduce a system that automatically filters information before it appears online. Through mandatory loading filters, platforms would be forced to create expensive and often biased systems to automatically review and filter possible copyright infringements on their websites.[56]

civil" [The European Copyright Directive and its impact on users in Latin America and the Caribbean: a perspective from civil society organizations], April 2019, retrieved from: https://bit.ly/36YtNRA, last access: November 4, 2019.

[53] European Union, directive No. 2001/29/CE of the European Parliament and the Council on the harmonisation of certain aspects of copyright and related rights in the information society, May 22, 2001, retrieved in Spanish from: https://bit.ly/2SPkUm8, last access: November 4, 2019.

[54] Office of the United Nations High Commissioner for Human Rights, "EU must align copyright reform with international human rights standards, says expert," March 11, 2019, retrieved from: https://bit.ly/2LtKlcH, last access: November 4, 2019. See also Lara, Juan Carlos, "Directiva de Derechos de Autor de la UE: avanza la internet filtrada en Europa" [EU Copyright Directive: filtered Internet gains space in Europe], Derechos Digitales, February 14, 2019, retrieved from: https://bit.ly/2LVP1aD, last access: November 4, 2019.

[55] See European Digital Rights (Edri), "Copyright: open letter calling for the deletion of articles 11 and 13," January 29, 2019, retrieved from: https://bit.ly/2FWqlXR, last access: November 4, 2019; Doctorow, Cory, "More than 130 European businesses tell the European Parliament: reject the #CopyrightDirective," Electronic Frontier Foundation (EFF), March 20, 2019, retrieved from: https://bit.ly/2ulSZ2k, last access: November 4, 2019.

[56] Sefidari, María, "Tu internet está bajo amenaza. Estas son las razones por las que deberías preocuparte por la Reforma Europea de Derechos de Autor" [Your Internet is under threat. These are the reasons why you should worry about the European Copyright Reform], Wikimedia, September 4, 2018, retrieved from: https://bit.ly/2xQex8K, last access: November 4, 2019.

In general, filtering poses at least two problems. On the one hand, there are the problems of transparency and accountability that are difficult to correct. On the other hand, due to the complexity of this type of system, probably very few companies have the technical capacity to develop this technology, so these measures would increase the concentration of power of the sector in a few large players. If we look at our region, the prior filtering encouraged by this regulation would be very similar to a system of direct censorship and incompatible with the Inter-American standards for freedom of expression.[57]

According to this rule and the regulations in this matter, satire and parody — including memes — would fall beyond the scope of this rule due to the fair-use exemption. However, the ability to detect this content before publication and distinguish it from others does not seem simple. Nor does it seem clear that all copyright holders want to exercise this right in this way in every case. Consider, for example, the case of "Pepe the frog" in the United States. The cartoon was created in 2005 by an American cartoonist, Matt Furie, and was initially published on his webcomic site. It quickly became popular, the face was adapted to fit different scenarios and emotions and it became a meme that worked in any context. In the following years, it was already viral on MySpace, Tumblr, and other social media as a funny way to express multiple moods. Despite being a work subject to copyright, "Pepe the frog" circulated freely without any action from its author demanding payment of usage fees or the cessation of its use. Undoubtedly, there was already some tension regarding copyright as the character was used and reused by people from different parts of the world. In 2016, however, the image of the Frog took an unexpected turn when the alt-right movement in the United States appropriated the character and made it their symbol to promote the campaign of the then-presidential candidate Donald Trump, by dressing the frog in Ku Klux Klan robes and military helmets. This controversial version of Pepe the Frog was reproduced in magazines and posters. It was then that Furie, its creator, sued the founder of a far-right site that used the image for political purposes for infringing copyright laws.[58]

[57] CELE, "La Directiva Europea de Derecho de Autor y su impacto en los usuarios de América Latina y el Caribe: una perspectiva desde las organizaciones de la sociedad civil" [The European Copyright Directive and its impact on users in Latin America and the Caribbean: a perspective from civil society organizations], *op. cit.*

[58] Leon, Harmon, "Alex Jones' Pepe the Frog Battle & the Co-Opting of Other Innocent Symbols," *Observer*, May 23, 2019, retrieved from: https://bit.ly/2O0T6th, last access: November 4, 2019; Thomser, Jacqueline, "Judge rules Pepe the Frog copyright lawsuit

There are many other cases like the previous one that expose the difficulty to detect: 1) the kind of use; and 2) the different connotations of certain uses, especially when they largely depend on the context.[59] It should also be noted that, in the case of memes, the legal principle of fair use varies according to whether it is referring to "static" or "mutating" memes, previously explained.[60] The fear caused by automatic filters is that they are not able to identify either the kind of user, nor can they recognize irony or purely satirical expressions, or distinguish them from other types of copyrighted material, resulting in the removal of memes from the Web.[61]

The European directive is an unprecedented policy, reversing the logic that, until now, governed the rules of intermediary liability — reactive logic — and moving instead towards a proactive logic. The systems in force so far follow a somewhat similar model to the US system of the Digital Millennium Copyright Act (DMCA) that provides for a "safe harbor" system. In this type of system, the intermediary is in principle not liable for third-party content until it is notified that the material already exists and that it infringes someone else's copyright. Several studies account for the abuses of these systems since they do not require a court order for the removal of content, and due to their architecture, they encourage the platform to block content and links when presented with mere notifications.[62] However, the governing principle is that of the platform's immunity against the content of third parties until there is a notification from the allegedly aggrieved party. On the other hand, users can appeal these orders to remove content by invoking fair use. Unlike these systems, the European directive is moving in a new direction, holding platforms directly liable for hosting copyrighted content

against InfoWars will go to trial," *The Hill*, May 17, 2019, retrieved from: https://bit.ly/2KbV4pM, last access: November 4, 2019.

[59] Dewey, Caitlin, "How copyright is killing your favorite memes," *The Washington Post*, September 8, 2015, retrieved from: https://wapo.st/33Nqgnb, last access: November 4, 2019.

[60] Lantagne, *op. cit.*, p. 395. The author mentions that while the fair-use principle could fit perfectly for static memes, where what is reproduced is an identical or almost identical imitation of an image, it might not be easily applicable if the use is mutating and there is a drastic transformation of the original image.

[61] For more information about the criticism of the directive: Reynolds, Matt, "What is Article 13? The EU's divisive new copyright plan explained," *Wired*, May 24, 2019, retrieved from: https://bit.ly/2xlQdM4, last access: November 9, 2019; Schaeffer, Joe, "The EU's Article 13 signals the death knell of political satire," October 8, 2018, in *Liberty Nation*, retrieved from: https://bit.ly/2LmSmjm, last access: November 9, 2019.

[62] Electronic Frontier Foundation (EFF), "NBC issues takedown on viral Obama ad," September 30, 2008, retrieved from: https://bit.ly/1SVR8bC, last access: November 9, 2019.

even in the absence of any notification and forcing them to "take all possible actions" to block access to shared content that infringes the pertinent laws. Additionally, the implementation of automatic and loading filters presents the risk that their large scope, lack of transparency and precision result in disproportionately high levels of removal of legitimate content. Undoubtedly, this will create a new scenario for freedom of expression on the Internet and humor in all its forms.

3. Disinformation

The discussion about fake news became widespread in 2016 after the surprise victory of Donald Trump.[63] Weeks before the elections, headlines such as "The Pope supports the candidacy of Donald Trump" or "Hillary Clinton sold weapons to ISIS"[64] circulated through fake accounts on Twitter, propagated on Facebook, and even appeared as search results on Google.[65] The issue took a turn after the circulation of a story of a pizzeria in central Washington DC that supposedly was a cover for a child sex ring and human trafficking run by Hillary Clinton and other members of the Democratic Party. As a result, in an episode later known as "Pizzagate," a 28-year-old man from North Carolina traveled there to rescue the children who were being exploited and shot three times in the pizzeria.[66] Various media outlets informed that disinformation had been decisive in the outcome

[63] The term fake news has different meanings and some have classified them under different criteria. For this article, we refer to fake news when describing deliberately false content, with the appearance of authenticity and intending to deceive the user. We also consider fake news those pieces of content disseminated through instant messaging services in the form of images, videos or memes, without an identified or identifiable author.

[64] See, for example, Ritchie, Hannah, "Read all about it: the biggest fake news stories of 2016," CNBC, September 30, 2016, retrieved from: https://cnb.cx/2wE3PDl, last access: November 9, 2019; Roberts, Hannah, "This is what fake news actually looks like – we ranked 11 election stories that went viral on Facebook," *Business Insider*, November 17, 2016, retrieved from: https://bit.ly/2xNhwPl, last access: November 9, 2019; Silverman, Craig, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook," Buzzfeed, November 16, 2016, retrieved from: https://bit.ly/2Ocl1dp, last access: November 9, 2019.

[65] Rogers, Kathy and Bromwich, Jonah, "The Hoaxes, Fake News and Misinformation We Saw on Election Day," *The New York Times,* November 8, 2016, retrieved from: https://nyti.ms/2o4NhOA, last access: November 9, 2019.

[66] Siddiqui, Faiz and Svrluga, Susan, "N.C. man told police he went to D.C. pizzeria with gun to investigate conspiracy theory," *The Washington Post*, December 5, 2016. Retrieved from: https://wapo.st/2k2clCq, last access: November 9, 2019.

of the elections,[67] and soon it all resulted in censorship measures in some countries[68], and it even began to be used by political figures to upset those media outlets whose journalistic coverage did not please them.[69]

Especially during elections, disinformation is an issue that impacts public discourse and, therefore, expressions of humor, particularly political humor, whose relevance in electoral times might be even greater. It is clear that when talking about fake news, humor is exempted, however, since fake news gained global relevance, the distinction between satire and false information has generated uncertainty and has become a gray area for journalists, researchers, and fact-checkers.[70] Some specialists consider that satire and parody constitute a type of misinformation because they are art forms that can become disinformation if audiences misunderstand the message.[71] Even though they do not intend to cause real damage, they can potentially misinform.[72] A recent UNESCO report issued a warning that satire can often be misunderstood by social media users, which in turn disseminate it as if it were true information.[73] Indeed, some people came to wonder if it is still possible to make humor and satire in the era of fake news.[74]

[67] Cortés, Carlos and Isaza, Luisa, "Noticias falsas en internet: la estrategia para combatir la desinformación" [Fake news on the Internet: the strategy to battle misinformation], CELE, December 2017, quoting: Read, Max, "Donald Trump won because of Facebook," *New York Magazine*. *Intelligencer*, November 9, 2016, retrieved from: https://nym.ag/2XQCK9N, last access: November 9, 2019; Dewey, Caitlin, "Facebook fake-news writer: 'I think Donald Trump is in the White House because of me,'" *The Washington Post*, November 17, 2016, retrieved from: http://wapo.st/2f3NllC, last access: November 9, 2019; Parkinson, Hannah, "Click and elect: how fake news helped Donald Trump win a real election," *The Guardian,* November 14, 2016, retrieved from: http://bit.ly/2fSyaDH, last access: November 9, 2019.

[68] Fraser, Matthew, "Debate: the legal fight against 'fake news' must not veer into censorship," *The Conversation*, June 11, 2018, retrieved from: https://bit.ly/2WFeQCn, last access: November 9, 2019.

[69] Annenberg School for Communication, *Understanding and addressing the disinformation ecosystem*, Philadelphia, PA, December 15-16, 2017, retrieved from: https://bit.ly/2GbeyJ2, last access: November 9, 2019.

[70] Smith, Justin E.H., "The end of satire. The toxic disinformation of social media has rendered traditional forms of humor quaint and futile," *The New York Times*, April 8, 2019, retrieved from: https://nyti.ms/2G7QM0V, last access: November 9, 2019.

[71] Wardle, Claire, "Fake news. It's complicated," *Medium*, February 16, 2017, retrieved from: https://bit.ly/2HRc2HJ, last access: November 9, 2019.

[72] Wardle, *ibid;* see also Ireton, Cherilyn and Posetti, Julie (eds.), *Journalism, 'fake news' & disinformation*, UNESCO, 2018, p. 46, retrieved from: https://bit.ly/2QmgNwM, last access: November 9, 2019.

[73] Ireton and Posetti, *op. cit.*, p. 17.

[74] Semley, John, "There's a deep, dark joke at the heart of Sacha Baron Cohen's 'Who is America?'" *Maclean's*, July 25, 2018, retrieved from: https://bit.ly/2GlDQ7w, last access: November 9, 2019.

Last year, for example, Snopes, a well-known data verification or fact-checking organization, verified a story of an American satire portal that was accusing the CNN of biased publications.[75] The article read: "CNN invested in an industrial-sized washing machine to help their journalists and news anchors spin the news before publication."

The article received more than 22 thousand interactions on Facebook. Hours later, users began receiving warnings before sharing the article and finally, the administrators of the page received a notification that their publication could be reduced in scope.[76] According to its founder, Snopes' policy is to verify any content that can be interpreted as true.[77]

[75] Mikkelson, David, "Did CNN purchase an industrial-sized washing machine to spin news?" Snopes, March 1, 2018, retrieved from: https://bit.ly/2NDMD8T, last access: November 9, 2019.

[76] Poynter, "Should satire be flagged on Facebook? A Snopes debunk sparks controversy," March 2, 2018, retrieved from: https://bit.ly/2JCO7Ow, last access: November 9, 2019; Alcazaren de Leon, Marguerite and Hapal, Kevin, "Satire vs. fake news. Can you tell the difference?" *Rappler*, May 15, 2018, retrieved from: https://bit.ly/2LtvKxS, last access: November 9, 2019.

[77] Poynter, *op. cit*.

This and similar incidents have sparked a debate about whether there is a need to check or "flag" satirical content or parodies. The mere existence of this debate impacts humor; in short, having to explain the joke even before it generates a reaction is an attack against its nature and goals.

The Snopes case is another example of the measures taken by platforms in response to the growing demand for transparency and fact-checking, some platforms are also implementing different actions to mitigate the circulation of fake news.[78] Facebook launched a pilot test to check news that users reported as fake, which is currently in force in fourteen countries.[79] In turn, Google recently adjusted its algorithms to make the results of the search engine as accurate as possible. To this end, it not only evaluates if the results coincide with the user's needs, but it also takes into account the quality of the sites that it shows ("quality" refers to the sites with the most experience, authority, trust, and reputation).[80] To achieve this, it tries to exclude results with content that is fake or that promotes misinformation, especially after various media outlets, a few years ago, reported on how the search engine yielded extremist and racist content among the first results to the question "Did the Holocaust really happen?"[81] Some elements common to these public and private initiatives include difficulties in clearly defining what constitutes misinformation; problems determining the nature of the content that would fall into this category; and an approach that would seem, at least in some cases, not to look at the content but rather the possible interpretations of it.

On the other hand, in addition to self-regulation measures, there were bills of law and public policies promoted in different countries to deal with this phenomenon. In Germany, for example, Congress adopted in 2018 the NetzDG, a rule that prohibits, among other things, misinformation and sanctions companies with large fines if they do not remove such content within a period not exceeding 24 hours after user notification.[82] Other examples

[78] Cortés and Isaza, *op. cit*.

[79] Lyons, Tessa, "Hard questions: what's Facebook's strategy for stopping false news?" Facebook Newsroom, May 23, 2018, retrieved from: https://bit.ly/2KP8JB2, last access: November 9, 2019.

[80] Hern, Alex, "Google tweaked algorithm after rise in US shootings," *The Guardian*, July 2, 2019, retrieved from: https://bit.ly/2Y958YR, last access: November 9, 2019.

[81] Cadwalladr, Carole, "Google is not 'just' a platform. It frames, shapes and distorts how we see the world," *The Guardian*, December 11, 2016, retrieved from: https://bit.ly/2hzO7Ot, last access: November 9, 2019.

[82] "Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG)" was passed in 2017 and entered into force on January 1, 2018, retrieved from: https://bit.ly/2qNxOHB, last access: November 9, 2019.

of laws with similar purposes appeared in France, where the National Assembly passed in 2018 a "law on the manipulation of information"[83] and in Singapore, where legislation against false news has just come into force.[84] All these initiatives were severely criticized by the international community and particularly by defenders of freedom of expression.[85] The main criticisms lie in the difficulty in defining fake news; the possibility that a public authority is in charge of shaping "truthful discourse"; the outsourcing of Justice that some of the measures imply; and the deadlines set for companies to adopt measures in this regard.

Regionally, in September 2018 the European Union announced its "Code of practice against disinformation," which made representatives of the main companies (Google, Facebook, Twitter, etc.) commit to adopting certain self-regulation standards on the subject.[86] The Code highlights the importance of transparency in the public debate and the need to combat disinformation as a threat to the democratic system. In addition, it establishes a wide range of commitments that include from guaranteeing transparency on sponsored content — especially the advertisements from political parties — to measures to identify and remove fake accounts. It also envisions follow-up mechanisms, so there are studies carried out to assess the situation, and signatories must report and compare their efforts monthly.[87] Signatory companies commit to investing in technology that would allow them to highlight news from "more reliable and accurate" sources and "reduce" the visibility of those with false or "misleading" content. Although parody and humor, in general, are explicitly exempt from the definition of misinformation adopted by the Code, ultimately, the elements used by the Code to characterize misinforma-

[83] Ayuso, Silvia, "La Asamblea Nacional francesa aprueba la ley contra las 'fake news'" [The National Assembly of France passes the law against 'fake news], *El País*, July 4, 2018, retrieved from: https://bit.ly/34Oo7Yj, last access: November 9, 2019; J.M.S., "La ley francesa contra las 'noticias falsas' se vuelve contra el propio gobierno" [The French law against 'fake news' turns against its own government], ABC, April 5, 2019, retrieved from: https://bit.ly/32xjssc, last access: November 9, 2019.

[84] Agence France-Presse, "'Chilling': Singapore's 'fake news' law comes into effect," *The Guardian*, October 2, 2019, retrieved from: https://bit.ly/32BpwQo, last access: November 9, 2019.

[85] See, for example, McAuley, James, "France weighs a law to rein in 'fake news,' raising fears for freedom of speech," *The Washington Post*, January 10, 2018, retrieved from: https://wapo.st/2rwWWCV, last access: November 9, 2019.

[86] European Commission, "Código de buenas prácticas para combatir la desinformación" [Code of Practice against disinformation], retrieved in Spanish from: https://bit.ly/2xEjvpw, last access: November 9, 2019.

[87] See monthly reports, *ibid.*

tion do not seem to exclude it: 1) fake and verifiable information; 2) created, presented or disseminated for profit or to intentionally deceive the public; and 3) that may be a threat to the public interest.[88] Many of the cases cited throughout this article, precisely, deal with publications and content that meet the three elements of the recommended test.

In Latin America, despite some bills that are in circulation, there are no "anti-fake news" laws or codes of conduct on the matter. However, States and companies allied during the electoral campaigns of 2018 and 2019 in order to generate systems of alerts, verification, and dissemination of checked content during electoral periods. Some examples are Reverso in Argentina (2019), Check in Brazil (2018) and Verified in Mexico (2018). The problems faced by these initiatives are similar to those of the other fact-checking efforts mentioned above.

On the other hand, and regardless of where or how they originate, the self-regulation policies of companies directly impact the users of the platforms globally. Initiatives such as the code of conduct or the European directive on copyright, while promoting that the companies incorporate technologies and practices, they indirectly impacted the users of these companies far beyond the European Union. This reality undoubtedly reveals a huge inequality between those who adopt and promote policies and those who suffer the consequences.

4. Discriminatory, offensive and politically incorrect expressions

"The trouble with satire, though, is that we all love it when it is directed at our enemies – and at those who are objectively ludicrous"[89] pointed out a journalist of The Guardian. What happens when satire addresses vulnerable groups and minorities or attacks the fundamental principles of a religious belief? In times of #NiUnaMenos, #MeToo, of old patriarchal models and

[88] "Disinformation: As provided under the Commission's Communication, for the purpose of this Code, the Commission as well as the High Level Expert Group in its report define "Disinformation" as "verifiably false or misleading information" which, cumulatively,(a) "Is created, presented and disseminated for economic gain or to intentionally deceive the public"; and (b) "May cause public harm," intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security." See European Union, "Code of Practice on Disinformation," retrieved from: https://bit.ly/33Dv3r4.

[89] Groskop, Viv, "A new satire must emerge – one that breaks out of the liberal bubble," *The Guardian*, February 13, 2017, retrieved from: https://bit.ly/2O43nZ3, last access: November 9, 2019.

stereotypes that have begun to be deconstructed progressively, is it possible to do the same humor as a few decades ago?

"We have become more cautious than at the beginning. People are very sensitive," states Mike Reiss, screenwriter and producer of "The Simpsons" one of the most influential series of recent times and famous for its unique brand of irony.[90] Humor to make fun of certain people or groups of people no longer seems to be received in the same way as before. Readers, viewers, and users are increasingly critical of sexist, racist or homophobic jokes. In this situation, some people invite us to carefully reconsider the legal boundaries of freedom of expression, further claiming that it must be "strictly within narrower limits when it is detrimentally in conflict with social integrity."[91]

There are many examples of expressions on social media that have caused controversy because of their discriminatory content. A few years ago, an American woman traveling to South Africa published a tweet that spread throughout the world and became the target of harsh criticism: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!" Soon, her account was deleted and she was fired from the company where she worked. In Mexico, memes that used images of indigenous people accompanied by messages that mocked their language became viral.[92] In Argentina, one of the most controversial sarcastic figures of recent years is "Doctora Alcira Pignata," an anonymous Twitter account that presents herself to her followers as being "against Arabs, Hebrews, homosexuals, blacks, *Peronists*, and scourge in general."[93]

Faced with this phenomenon that has recently stirred great social outcry, both State and private initiatives have emerged. A prime example of this in Europe is the NetzDG in Germany, which addresses the issue of hate speech and forces the platforms to delete content within 24 hours under penalty of high fines. As we mentioned before, the rule was criticized for the

[90] Ruiz Jiménez, Eneko, "Los guionistas de 'Los Simpsons' nos hemos vuelto más cautos. La gente está muy sensible" [The writers of The Simpsons' have become more cautious. People are too sensitive], *El País*, July 12, 2019, retrieved from: https://bit.ly/2CxXHhx, last access: November 9, 2019.

[91] Yoo, Kwanghyuk, "When does cultural satire cross the line in the global human rights regime?: the Charlie Hebdo controversy and its implication for creating a new paradigm to assess the bounds of freedom of expression," *Brooklyn Journal of International Law*, No. 2, Vol. 42, 2017, p. 764, retrieved from: https://bit.ly/30GMmpf, last access: November 9, 2019.

[92] Gaona, Pável M., "¿Son discriminatorios los memes que usan imágenes de indígenas?" [Are memes that use images of indigenous people discriminatory?], Consejo Nacional para Prevenir la Discriminación (Conapred), México, retrieved from: https://bit.ly/2O56WuZ, last access: November 9, 2019.

[93] The tweet was posted on May 29, 2014, by @drapignata, and was later eliminated. This account is also suspended.

vagueness of its terms and the incentives for companies. A few days after it entered into force, the Twitter account of a German satirical magazine was blocked after parodying Beatriz von Storch, a member of the Alternative Far-Right party for Germany, who a few days earlier had criticized the police for "appeasing the hordes of barbarian men, Muslims, and rapists."[94] Twitter suspended the account for approximately 48 hours for violation of the supposed hate speech included in the NetzDG. The German Journalists Association declared that such a measure implied censorship and warned that the regulations could become extremely dangerous.[95]

In Latin America, the issue of hate and discriminatory speech is of great interest in local congresses. In Mexico, a bill of law "to prevent and eliminate discrimination" was presented in 2018.[96] With a similar legislative language, since 2016 and up to now there have been multiple initiatives in Argentina to modify the current Ley de Actos Discriminatorios [Law on Discriminatory Acts] and explicitly include the "promotion of non-discrimination on the Internet."[97] In Ecuador, former President Rafael Correa presented a bill to regulate "acts of hate and discrimination on social media on the Internet," including the platforms' obligation to remove and block content within 24 hours, including any replicas that have been made.[98] All these bills of law are based on a common idea: that the progress of the Internet has been accompanied by a growing confrontational spirit, and that it allows people to easily carry out discriminatory and hateful acts due to the possibility of anonymity, with a huge impact thanks to its possible immediate viralization. Despite pursuing legitimate objectives, these initiatives' intended measures pose serious issues regarding their legality and proportionality. They carry

[94] Thomasson, Emma, "German hate speech law tested as Twitter blocks satire account," *Reuters*, January 3, 2018, retrieved from: https://reut.rs/2rm6Al2, last access: November 9, 2019.

[95] *Ibid*; see also "German satire magazine Titanic back on Twitter following 'hate speech' ban," *Deutsche Welle*, January 6, 2018, retrieved from: https://bit.ly/2X5FdOL, last access: November 11, 2019.

[96] See Observatorio Legislativo CELE, "México proyecto de ley federal para prevenir y eliminar la discriminación" [Mexico federal bill to prevent and eliminate discrimination], 2018, retrieved from: https://bit.ly/2CEECKh, last access: November 9, 2019.

[97] See Observatorio Legislativo CELE, "Argentina proyecto de ley sobre modificación de Ley de Actos Discriminatorios" [Argentina bill of law to amend the Law on Discriminatory Acts], 2018, retrieved from: https://bit.ly/33Nr1N3, last access: November 9, 2019.

[98] See Observatorio Legislativo CELE, "Ecuador proyecto de ley que regula los actos de odio y discriminación en redes sociales e internet" [Ecuador bill of Law for the regulation of acts of hate and discrimination on social media and the Internet], 2017, retrieved from: https://bit.ly/2Q7F6kc, last access: November 9, 2019.

very high risks for the circulation of discourse, including cases of prior censorship, political persecution, and excessive latitude of public bodies in their application.

On the other hand, a distinctive attribute of regional legislative activity is its tendency to be reactive. In this sense, in the Regional Legislative Observatory for Freedom of Expression of CELE, we have witnessed little previous debate about the rules presented to regulate content in the digital space. Given the complexity of the Internet's architecture, its design and operation, it would be desirable that any attempt at regulation be accompanied by a broad and vigorous debate, ensuring the participation of all relevant actors, especially the technical sector, civil society, and academic institutions.

In the private sector, companies have adopted clear policies to reject segregation, exclusion, and discrimination. Facebook, for example, has a hate speech policy that catalogs content on three levels.[99] The first level includes violent or dehumanizing content based on one of the characteristics of the protected or semi-protected group (gender, nationality, religion, country of origin, etc.)[100]; the second level includes content with pejorative generalizations directed to protected groups or persons with references to deficiencies of physical (hygiene, for example), mental or moral nature; referencing handicaps; or offending the group in some other way; and in the third level language that excludes (for example, saying that women are not allowed.) The policy is complex and excessively detailed with examples of the type of content that would not be tolerated on the platform. Its implementation has generated major controversies, particularly because it does not make any distinction based on the level of vulnerability between the groups. Under this policy, a comment excluding women would be equivalent to a comment excluding men. Along the same lines, expressions of white supremacists regarding Afro-descendants would be as repressed as expressions of Afro-descendant activists about white supremacists or even stories of racism.[101] Although the policy has an explicit exemption for humor, in many of the publicly reported cases, it was not able to grasp neither the broader nor the more subtle cases of irony.

[99] Facebook, "Lenguaje que incita al odio" [Hate Speech], retrieved in Spanish from: https://bit.ly/2KeosMa, last access: November 9, 2019.

[100] Facebook distinguishes between protected and semi-protected characteristics. See *ibid*.

[101] See, for example, Guynn, Jessica, "Facebook while black: users call it getting 'Zucked', say talking about racism is censored as hate speech," April 24, 2019, retrieved from: https://bit.ly/2QiZ3F1, last access: November 9, 2019.

On the other hand, between 2018 and 2019, Twitter carried out a study based on the proposal to modify its terms and conditions of service regarding "dehumanizing content." After a year of inquiries on the subject, and having received more than a thousand responses from academics, activists, non-governmental organizations, experts and officials, the company decided to limit the scope of the proposed rule and apply it only to those tweets that use "dehumanizing" language on religious grounds.[102]

This array of measures shows that there are certain expressions that both States and companies choose not to accommodate. There is an evident interest in protecting values like equality and non-discrimination, and humor seems unable to pass this obstacle. In this context, it is important to strengthen respect for the basic principles of legality, necessity, and proportionality for any type of restriction on freedom of expression, especially given the crucial role of expressions of political humor in the public interest discourse.

IV. Conclusions

Humor, as an expression intended to shock, question and provoke, has caused major divisions throughout history. The immediate viralization, the amplification of the messages and the possibility of anonymity that the Internet provides have contributed to aggravating some of these conflicts. However, much of the "new" threats to humor come with policies and practices that affect our object of study more as collateral damage than as a deliberate measure.

The speed, permanence, and accessibility of content on the Internet, coupled with a decentralized and universal structure, have raised issues regarding the means to protect individual rights such as honor, dignity, and copyright against fake news, misinformation, and discrimination. These debates are giving rise to public and private policies that undermine a deeper deliberation of every case. Setting up short terms in the style of NetzDG, the resolution of disputes in the hands of companies and not of State justice, and more recently the incentives for the adoption of filters (used before and after loading) without prior complaint or notification, give strong incentives for automating processes of detection, removal, and moderation of content. Accordingly, the contexts, symbolisms, and language that characterize satirical expressions, parody and humor in general — but particularly politi-

[102] Twitter Blog, "Updating our rules against hateful conduct," July 9, 2019, retrieved from: https://bit.ly/2YJdBPt, last access: November 9, 2019.

cal humor — are held back, become invisible and smothered. Humor, like many other types of expressions, including criticism and opinion pieces, necessarily needs to be evaluated and take a place on its own context to fulfill its purposes.

The privatization of the analysis and the decisions regarding the removal, visibility, and scope of content on the platforms is a common factor in at least three of the cases analyzed in this article: misinformation, discriminatory discourse, and copyright. This outsourcing, endorsed and promoted by certain States, is currently subject to merely partial requirements of accountability and transparency. Oftentimes, companies cannot even notify authors about the limitations that apply to their content. On the other hand, the success of these actions is often measured only concerning the decrease in formal judicial claims, and they have the power to review, evaluate, reject or implement changes.

This document tried to review the measures, both public and private, that affect humorous discourse. Nonetheless, many of the questions and conclusions outlined in this article are transferable to several other problems in the digital arena.

Chapter three

Fake news on the Internet: the strategy to battle misinformationn

Carlos Cortés* y Luisa Isaza**

I. Introduction

On December 4, 2016, a month after the US presidential election, Edgar Maddison Welch, a 28-year-old man from North Carolina, fired three shots at the Comet Ping Pong pizzeria in Washington DC. He was determined to enter, investigate and rescue the children who were being exploited by an alleged child sex trafficking network run by Hillary Clinton and other members of the Democratic Party.[1]

Maddison Welch was convinced of the existence of such network from news he had read on the Internet. After his arrest and despite the fact that the police denied the story - known as *Pizzagate* - the man apologized but never admitted that the information that had motivated his attack was fake.[2] The outrageous story was spread through social media and discussion forums on the Internet, along with hundreds of false stories related to the two

* Lawyer, University of Los Andes (Colombia), Master's Degree, Media and Communication Governance, London School of Economics. He is a consultant in freedom of expression and Internet regulation, and external advisor to Twitter's public policy team.

** Lawyer, Universidad Javeriana (Colombia). Legal advisor, Defense and Attention to Journalists Office, Fundación para la Libertad de Prensa (FLIP), Colombia.

*** This article was originally published by CELE in December 2017.

[1] Siddiqui, Faiz and Svrluga, Susan, "N.C. man told police he went to D.C. pizzeria with gun to investigate conspiracy theory", *Washington Post*, December 5, 2016. Retrieved from: http://wapo.st/2gW0yFo.

[2] Goldman, Adam, "The Comet Ping Pong Gunman Answers Our Reporter's Questions", *The New York Times,* December 7, 2016. Retrieved from: http://nyti.ms/2Er1xrM.

candidates or other members of their parties.[3]

Weeks before the election, millions of people saw in their Facebook News Feed a piece of news about an unprecedented statement from Pope Francis proclaiming his support for the candidacy of Donald Trump. This piece of fake news received 960,000 interactions in the social network (comments, reactions and shares), more than any other real news about the election.[4] As a matter of fact, according to a BuzzFeed study released days after the election, the top twenty fake news available in Facebook during the three months prior to the election had more engagement than the twenty main actual stories from the most well-known media outlets (New York Times, Washington Post, Los Angeles Times, Wall Street Journal, FOX News, among others) published on the same social media network.

The Guardian and BuzzFeed revealed that many of these stories were being produced by a group of young Macedonians, who through suggestive headlines that produced clicks - a deceptive technique known as clickbait - made thousands of dollars in advertising thanks to the traffic on their Internet sites.[5] Over a hundred websites were created for this purpose in the city of Veles, Macedonia, all designed to look like authentic news portals. Other fake news factories operated directly within the United States.[6] According to their own creators, much of the traffic in these sites came from clicks originated in Facebook, where they also had hundreds of thousands of followers.

The false information was spread until the day of the election in the form of hoaxes, fake Twitter accounts, misinformation tweets and even Google search results.[7] After the vote, Mediaite reported that the first result in the Google search

[3] Other pieces of fake news related to the elections and widely circulated stated, for example, that it had been confirmed that Hillary Clinton had sold weapons to ISIS, that Donald Trump was offering one-way tickets to Africa and Mexico for those who wanted to leave the country and that an ISIS leader was asking Muslims to vote for Clinton. Silverman, Craig, "Here Are 50 Of The Biggest Fake News Hits On Facebook From 2016", *Buzzfeed,* December 30, 2016. Retrieved from: http://bzfd.it/2Ge4ZXo.

[4] Silverman, Craig, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook", *Buzzfeed,* November 16, 2016. Retrieved from: http://bzfd.it/2ssm265.

[5] Tynan, Dan, "How Facebook powers money machines for obscure political 'news' sites", *The Guardian,* August 24, 2016. Retrieved from: http://bit.ly/2bhzDzv; Silverman, Craig y Alexander, Lawrence "How Teens In The Balkans Are Duping Trump Supporters With Fake News", *BuzzFeed,* November 3, 2016. Retrieved from: http://bzfd.it/2EubVDu; Ohlheiser, Abby, "This is how Facebook's fake-news writers make money", *The Washington Post,* November 18, 2016. Retrieved from: http://wapo.st/2Bt9wGk.

[6] Silverman, Craig y Singer-Vine, Jeremy, "The True Story Behind The Biggest Fake News Hit Of The Election", *BuzzFeed,* December 16, 2016. Retrieved from: http://bzfd.it/2BZ7kHs.

[7] Rogers, Kathy and Bromwich, Jonah, "The Hoaxes, Fake News and Misinformation

engine to "final vote count 2016" was a site called 70News where it was falsely stated that Donald Trump had won both the electoral and popular votes.[8]

After Donald Trump's surprising victory, the discussion about fake news exploded.[9] Some people - including a fake news author- said that misinformation in social networks had directly influenced the outcome of the elections.[10] And although at the moment there is no study to measure this impact in a clear way, it is undeniable that Facebook, Twitter and Google did play an important role as the main source of information for many people.[11]

The services of these companies were not only exploited by astute young people to make profits out of advertising. According to the findings of American intelligence, the Russian government used these platforms to spread fake news and propaganda, seeking to influence public debate during the campaign and benefit Trump's candidacy to the detriment of Clinton's.[12] More recently, the Prime Minister of the United Kingdom, Theresa May, made similar accusations against the Russian government.[13]

We Saw on Election Day", *The New York Times,* November 8, 2016. Retrieved from: http://nyti.ms/2o4NhOA.

[8] Abrams, Dan, "Now Even Google Search Aiding in Scourge of Fake, Inaccurate News About Election 2016", *Mediate,* November 13, 2016. Retrieved from: http://bit.ly/2ssmbq9.

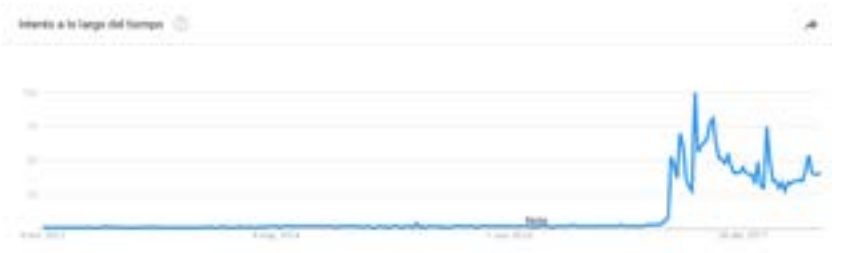
[9] A Google Trends search reveals that searches of the term *fake news* shot up in November, 2016, after being next to insignificant for the past few years. The peak of searches happened in the week of January 8 to 14, 2017, which is when the famous incident occurred in which Donald Trump refused to answer a question from a CNN journalist, saying " You are fake news! " Google Trends, http://bit.ly/2Gfn3QZ, last access: December 4, 2017.

[10] Read, Max "Donald Trump Won Because of Facebook", *New York Magazine,* November 9, 2016. Retrieved from: http://nymag.com/selectall/2016/11/donald-trump-won-because-of-facebook.html; Dewey, Caitlin, "Facebook fake-news writer: 'I think Donald Trump is in the White House because of me'", *The Washington Post,* November 17, 2016. Retrieved from: http://wapo.st/2f3NllC; Parkinson, Hannah, "Click and elect: how fake news helped Donald Trump win a real election", *The Guardian,* November 14, 2016. Retrieved from: http://bit.ly/2fSyaDH.

[11] According to the latest study on digital news from the Reuters Institute for the Study of Journalism at the University of Oxford, a survey conducted in 36 countries revealed that 54% of Internet users use social media as a source of news (Reuters Institute for the Study of Journalism, "Digital News Report ", http://bit.ly/SBryvD, last access: December 14, 2017). In the case of the United States, a study by the Pew Research Center shows that 67% of the population consumes news through these platforms. Living up to its name, Facebook's News Feed is the main source of news among respondents who used social media as a news source (Pew Research Center , "News Use Across Social Media Platforms 2017", http://pewrsr.ch/2vMCQWO, last access: December 14, 2017).

[12] Office of the National Director of National Intelligence of the National Intelligence Council of the United States of America, "Assessing Russian Activities and Intentions in Recent US Elections", January 6, 2017, retrieved from: http://bit.ly/2iRbS9b.

[13] Mason, Rowena, "Theresa May accuses Russia of interfering in elections and fake



Note. Google searches of the term "fake news" during the last five years (Google Trends).

Several US authorities initiated investigations into Russian interference in the election, for which they have requested information from Internet companies. Initially, Facebook wanted to downplay the problem: two days after the election, Mark Zuckerberg publicly argued that thinking that fake news articles had an impact on the election was a "pretty crazy" idea.[14] But later the company admitted before the United States Senate that millions of users had seen advertising produced by Russia, published on Facebook and Instagram.[15] Meanwhile, Twitter reported that it had found 2,752 accounts which were being controlled by Russians and that Russian bots had tweeted 1.4 million times during the election. Google found on YouTube more than a thousand videos on the subject.

This scenario became the perfect storm for Internet companies, which for a great variety of issues and on different fronts face pressure from governments and civil society to make their practices transparent.[16] Misinformation is not an isolated problem nor is it new. In the background is the question

news", *The Guardian,* November 14, 2017. Retrieved from: http://bit.ly/2mnQaML.

[14] Interview conducted by journalist David Kirkpatrick, founder of Techonomy Media, on November 10, 2016, retrieved from: http://bit.ly/2o6NMrv.

[15] Shaban, Hamza, Timberg, Craig and Dwoskin, Elizabeth "Facebook, Google and Twitter testified on Capitol Hill. Here's what they said", *The Washington Post,* October 31, 2017, retrieved from: http://wapo.st/2gZM9fx.

[16] In June 2017, the German parliament passed a law that requires Internet companies with over two million users to remove hate speech and other illegal content from their platforms, such as defamatory fake news in less than 24 hours, on pain of fines of up to 50 million euros (Miller, Joe, "Germany votes for 50m euro social media ends", BBC, June 30, 2017, retrieved from: http://bbc.in/2C0rTna). More recently, a spokesperson for Theresa May reported that the UK is evaluating the role of Facebook and Google in the supply of news and what their responsibilities might be ("Britain looking at Google, Facebook role in news: PM May's Spokesman," Reuters, October 10, 2017, retrieved from: http://reut.rs/2swt75t).

of harmful content -defamatory, incendiary, which violates privacy, among others- and the companies' response to these. The moderation of content is, therefore, at the center of this discussion and its implications reach the exercise of online freedom of expression.

As it will be seen below, Facebook and Google are adopting measures to deal with the problem. Some of them focus on prevention, seeking to educate the citizen to make informed decisions about the content he or she consumes. Others have a direct effect on the information that is published on the platforms, either through invisible changes in the algorithm or visible warnings about the veracity of the information. Most of these measures are trials or have partial geographic coverage;[17] others are meant only for specific moments, such as elections. In any case, it is not easy to determine their scope and depth because they are essentially a series of announcements whose implementation is not entirely clear.

This document exposes the measures announced by Facebook and Google to combat misinformation. It also includes a brief allusion to YouTube and Twitter. The focus is set on those measures that have a direct effect on the platform and on the information received by its users, and not on preventive and educational measures that have been in development in parallel. Likewise, the document tries to determine the geographic coverage of these measures. Subsequently, as a conclusion, the possible problems behind the proposed solutions are divided in four points: i) scale and time, ii) impact, iii) the role of civil society, and iv) transparency.

II. Misinformation and manipulation: a tentative classification

In November of 2017, Collins Dictionary chose "fake news" as the word of the year.[18] The expression, which was used 365% more this year, will be included in the next edition of that dictionary as "false information, frequently sensational, disseminated under the guise of news reporting". However, fake news does not have one single connotation. The general

[17] The "Explore" Facebook experiment, developed in six countries, separates friend posts and advertising into one tab and public content from other Facebook accounts into another. Facebook has announced that it has no plans to make this measure definitive or extensive to everyone. Facebook, "Clarifying Recent Tests", http://bit.ly/2zwW0NF, last access: December 14, 2017.

[18] Flood, Alison, "Fake news is 'very real' word of the year for 2017", *The Guardian,* November 2, 2017, retrieved from: http://bit.ly/2iTWYk4.

public does not only use it to refer to false reports, but in general to express a discontent with the misinformation, especially online.[19] Without having to look too far for an example, President Donald Trump uses it to disqualify any information he does not agree with.

The truth is that, beyond the political use of the term, fake news is also related to extremist opinions, propaganda and manipulation. Many might consider a politician's uninformed and alarmist opinion on Twitter to be fake news, to the same degree as a news story which reports falsely about the death of a world leader in deliberate bad faith. In both cases we find differences in content (an opinion versus a piece of news), format (a tweet versus a web page) and possibly motivations (the politician wants to rally his or her base while the website wants clicks).

Below there is a classification to help understand how content is produced and how it reaches the reader. This classification does not set mutually exclusive categories. For example, content can be fake news and have, at the same time, a propagandistic approach. It is not fully comprehensive either, it specially excludes contents of satirical journalism and errors of reporting committed in good faith. The latter could be part of a debate about fake news, but are not related to the purpose of this document.

1. Fake news

This refers to deliberately false content that is published on websites whose appearance tries to be formal and authentic.[20] Sometimes the design of the site and its URL take the place of a well-known news portal. The clear purpose is to deceive the user. Generally these contents proliferate in social networks through the own accounts of those portals, either in an organic way - with likes, retweets and shared by users- or with promoted actions, that is, paying for these contents to be advertised by the platforms.

[19] *"Our findings suggest that, from an audience perspective, fake news is only in part about fabricated news reports narrowly defined, and much more about a wider discontent with the information landscape— including news media and politicians as well as platform companies. Tackling false news narrowly speaking is important, but it will not address the broader issue that people feel much of the information they come across, especially online, consists of poor journalism, political propaganda, and misleading forms of advertising and sponsored content."* Nielsen, Rasmus and Graves, Lucas, "'News you don't believe': Audience perspectives on fake news", Reuters Institute for the Study of Journalism, October 2017, retrieved from: http://bit.ly/2o4Exb6.

[20] Media Matters, "Understanding The Fake News Universe", http://bit.ly/2EDq8NH.

Note. False story published in the abcnews.com.co site, created to mimic ABC News, whose web address is abcnews.go.com. According to this false news, some of the protesters against Donald Trump were paid to protest. Despite the falsehood of the news, days after his election, Trump himself suggested that these people were "professional demonstrators".[21]

Note. A picture of voting in Arizona was altered to include the image of an arrest. The Twitter user who posted it said that an undocumented immigrant had been arrested for trying to vote.[22]

[21] This news was shared on Twitter by Eric Trump, Donald Trump's son, by Corey Lewandowski and Kellyanne Conway, two of Trump's campaign leaders. Jacobson, Louis, "No, someone wasn't paid $3,500 to protest Donald Trump; it's fake news", *Politifact,* November 17, 2016, retrieved from: http://bit.ly/2flpuUR; Stahl, Lesley, "President-elect Trump speaks to a divided country", *CBS News,* November 13, 2016, retrieved from: http://cbsn.ws/2swFxKz.

[22] Wardle, Claire, "6 types of misinformation circulated this election season", *Columbia Journalism Review,* November 18, 2016, retrieved from: http://bit.ly/2CkRdjh.

Pieces of fake news in the strict sense can have economic or political motivations or a little of both. In the first case, they are commercial operations that seek to generate traffic from false contents and, above all, sensationalist headlines that people click on, but where the information does not make sense nor has any relevance. In the second case, they try to appear authentic not so much to generate traffic and profits but to manipulate the public debate in favor of certain political interests. An example of this category is the aforementioned false information on the support of Pope Francis to Donald Trump's candidacy in 2016.

Note. Fake news about the Pope's support for Donald Trump.

Political interest in the piece of fake news does not necessarily exclude the economic interest. While some pieces of fake news are created with either one of these motivations, in many cases both orbits can converge. In the case of misinformation around the presidential campaign in the United States, young Macedonians may have been indifferent to who won the election; but not the Russian operatives who also influenced it. In the latter case, the economic benefits derived from web traffic and interaction added to the underlying political agenda.[23]

[23] "These Macedonians on Facebook didn't care if Trump won or lost the White House. They only wanted pocket money to pay for things--a car, watches, better cell phones, more drinks at the bar". Subramian, Samanth, "*Welcome To Veles, Macedonia, Fake: News Factory To The World", Wired,* March, 2017, retrieved from: http://bit.ly/2o7BOxQ.

2. Propaganda

Jacques Ellul considers propaganda an elusive concept that develops around psychological and war action, re-education, brainwashing and public relations between humans. In that sense, the French sociologist thinks that propaganda is, above all, a technique to influence the actions of groups or individuals.[24]

Among other manifestations, propaganda may include false information or certain information presented with a deceptive approach.[25] For example, some facts are reported but others are omitted; information is out of context; content is manipulated; theories or opinions are presented as facts; highly disputable information is given as credible; information is denied in order to create confusion, or one statement is proclaimed as the only truth in opposition to the 'other' - the strategy of the nationalist movements–.[26]

Propaganda has been a part of politics and communications at least since the beginning of last century. Therefore, it is not a digital phenomenon. However, the reach of these contents online is especially significant: through the advertising tool -particularly in Facebook- propaganda is tailored to specific communities and groups based on tastes, political bias and friendship circles. In the case of the elections in the United States, this level of detail reached an alarming level:

Dark ads - as they are known due to the impossibility of knowing who sees them - allowed reaffirming political convictions, stirring up differences and, in general, polarizing the electorate. A white man from a republican state saw an advertisement against immigration or in defense of the use of guns; an African-American saw an advertisement that recalled the racial persecution of the Ku Klux Klan; a Catholic saw Hillary as the incarnation of the Devil in a fight against Jesus.[27]

---

[24] Ellul, Jacques, *Propaganda: The Formation of Men's Attitudes,* New York, Vintage Books, 1965.

[25] "*Propaganda is false or misleading information or ideas addressed to a mass audience by parties who thereby gain advantage. Propaganda is created and disseminated systematically and does not invite critical analysis or response".* Huckin, Thomas, "Propaganda Defined", in: Henderson, Gae Lyn and Braun, M.J (Eds), *Propaganda and Rhetoric in Democracy: History, Theory, Analysis*, Carbondale, Southern Illinois University Press, first ed, 2016, pp. 118-136.

[26] Powell, John A. "Us vs them: the sinister techniques of 'Othering' – and how to avoid them", *The Guardian,* November 8, 2017, retrieved from: http://bit.ly/2iGLAUX.

[27] Cortés, Carlos, "El algoritmo imposible, redes sociales y noticias falsas", *Revista Arcadia,* December, 2017, retrieved from: http://bit.ly/2CmcdpT.

Note. A television ad for Donald Trump's campaign stated that the then-candidate would stop illegal immigration on the "southern border" by building a wall that would be paid by Mexico, while showing a video of dozens of people crossing the border between Morocco and the Spanish city of Melilla, not between Mexico and the United States.[28]

3. Conspiracy theories

Conspiracy theories seek to explain a particular event as the result of a plan carefully coordinated by an individual or a group. Motivations are generally secret and malicious, and actions are carried out to the detriment of the general interest.[29] These theories swarm in video channels and Internet pages, and are often presented as news despite their scant factual foundation.

In Colombia, there has been a theory for some years according to which President Juan Manuel Santos was recruited to secretly work for the Cuban government.[30] In Argentina, after the disappearance of Santiago Maldonado, multiple conspiracy theories erupted on the Internet, including one that stated that the website santiagomaldonado.com, created to demand the appearance

[28] Edds, Carolyn, "Donald Trump's first TV ad shows migrants 'at the southern border,' but they're actually in Morocco", *Politifact,* June 4, 2016, retrieved from: http://bit.ly/1mvNQi9.

[29] Media Matters, "Understanding The Fake News Universe", http://bit.ly/2EsVDdZ, last access: December 14, 2017. According to Michael Barkun, conspiracy theories are characterized by following three principles, namely: i) nothing happens by accident, ii) nothing is what it seems and iii) everything is connected. Barkun, Michael, *A Culture of Conspiracy: Apocalyptic Visions in Contemporary America*, Berkeley, University of California Press, 2003, pp. 3-4, retrieved from: http://bit.ly/2Btakek.

[30] "Los Santos y su militancia castrocomunista", *Periodismo Sin Fronteras,* June 1, 2013, retrieved from: http://bit.ly/1hjVlzx.

of the activist alive, had been set up before his disappearance.[31]

The Pizzagate conspiracy theory, mentioned in the introduction of this paper, was examined by several media outlets in the United States with the purpose of identifying its origin.[32] In October 2016, a Twitter user posted a Facebook message in which a woman claimed that an anonymous source from the New York City Police Department had told her there was evidence that Hillary and Bill Clinton were involved in a child sex trafficking network. The tweet reached thousands of retweets and hours later an user of a discussion forum on the Internet posted a message saying that "internal sources" had confirmed the existence of the pedophilia ring, which would be exposed in a matter of hours.[33] The next day, the fake news site YourNewsWire.com posted a story based on the comments of a 4chan user - a trolling-heavy online board - where it was reported that an FBI source had confirmed the accusations.[34] The story was replicated and expanded by other fake news sites and shared on Facebook.

Among all the published versions, the story reached hundreds of thousands of interactions on Facebook and Twitter, and began to go viral under the #PizzaGate label.[35] Fake news articles were created with manipulated photos. Discussion forums and comments sections talked about a network of underground tunnels, torture chambers, Satanism and cannibalism in the basements of several restaurants. When the media refuted the theory they were accused by believers of wanting to hide the truth. Even weeks after the arrest of the man who shot the Comet Ping Pong pizzeria - where, of course, no illegal network was operating - some people still hinted that the story was true.[36]

---

[31] "El sitio de Santiago Maldonado fue creado antes de su desaparición", *Data 24,* September 18, 2017, retrieved from: http://bit.ly/2w3l0ed.

[32] The following articles can be read on this: (i) Aisch, Gregor, Huang, Jon and Kang, Cecilia, "Dissecting the #PizzaGate Conspiracy Theories", *The New York Times,* December 10, 2016, retrieved from: http://nyti.ms/2jcCzlu. (ii) Silverman, Craig, "How The Bizarre Conspiracy Theory Behind 'Pizzagate' Was Spread", *Buzzfeed,* December 5, 2016, retrieved from: http://bzfd.it/2CjSQOl. (iii) LaCapria, Kim, "Chuck E. Sleaze", *Snopes,* November 21, 2016, retrieved from: http://bit.ly/2xX7xta.

[33] The discussion forum "Breaking: Its worse then classified emails. Political Pedophile Sex Ring exposed" retrieved from: http://bit.ly/2ErmerY, last access: December 14, 2017.

[34] "FBI Insider: Clinton Emails Linked To Political Pedophile Sex Ring", *YourNewsWire.com*, October 31, 2016, retrieved from: http://bit.ly/2fesMdl.

[35] Silverman, Craig, "How The Bizarre Conspiracy Theory Behind 'Pizzagate' Was Spread", BuzzFeed, December 5, 2016, retrieved from: http://bzfd.it/2CjSQOl.

[36] Allam, Hannah, "Conspiracy peddlers continue pushing debunked 'pizzagate' tale", *Miami Herald,* December 5, 2016, retrieved from: http://hrld.us/2CFYGKw.

4. False information, rumors, chains, memes

A broader category of misinformation, which includes different forms of fake news, propaganda and conspiracy theories, consists of the contents mixed and spread through instant messaging services, mainly WhatsApp. On this platform, information moves from hand to hand in the form of images, videos or memes, without an identified or identifiable author. In the end, the contact who shared it gives legitimacy and authority to the content.

In countries like Colombia this kind of content had a seemingly far reach last year. During the campaigns for and against the Havana peace agreements, prior to the October 2016 plebiscite, false, inaccurate and decontextualized content was disseminated through WhatsApp, aimed at capturing the vote against the agreements. Like one of the leaders of the "No" campaign admitted, the goal was to generate anger and outrage.[37]

Note. Messages with fake information in social networks and WhatsApp during the campaign prior to the October 2016 plebiscite.[38]

Whether as eminently false news, propaganda, conspiracy theories or rumors, fake news is not an isolated phenomenon of social and political real-

[37] Ramírez, Juliana, "El No ha sido la campaña más barata y más efectiva de la historia", *Asuntos Legales,* October 4, 2016, retrieved from: http://bit.ly/2EHxlwc.

[38] "¿Qué tan ciertos son algunos memes de la campaña del "NO" en el plebiscito?", *Pacifista,* August 30, 2016, retrieved from: http://bit.ly/2bP16MN.

ity, and in no way is it an externality of technology. "Today, when we talk about people's relationship with the Internet, we tend to adopt the uncritical language of computational science. 'Fake news' is described as a 'virus' among users 'exposed' to online misinformation" explains North American sociologist Katherine Cross.[39] Misinformation originates and feeds back in and from human actions and it is in this relationship that we can locate the dimension of the problem and the limitations of the proposed solutions.

III. Affordances and spaces

Just as the phenomenon of misinformation responds to different social contexts, it does not manifest itself in the same way in all digital media: a space such as WhatsApp enables the exchange of anonymous content -without author or apparent source-; Facebook gives greater relevance to content that is massively shared and on Twitter the user chooses the voices he or she wants to hear. In this dialog between the platform and the user the production and exchange of information acquires particular connotations. Fake news do not take form in the same way in all these spaces.

The affordances are the properties that arise from the relationship between an object and a person. The concept of affordances, introduced by the psychologist James Gibson, refers to the possibilities of action in a given context, and is used to talk about how users interact with objects, environments or technologies to achieve certain results.[40] Identifying the offerings of an object allows the understanding of the different ways in which the object can be used for various purposes. Some of those offerings are inscribed in the design of the object; others are "discovered" by the individual. A broom is used to sweep, but also to hit a distant object, like an orange in a tree; a

[39] Cros, Katherine, "The Art of the Real: Disinformation vs. democracy", *The Baffler,* June, 017, retrieved from: http://bit.ly/2Er1KLR.

[40] Gibson, James, *The Ecological Approach to Visual Perception,* Hillsdale, Cornell University, 1986. Evans, Pearce and others define offerings as the multifaceted relational structure between an object or technology and the user, which allows or restricts potential behavioral outcomes in a particular context. To understand the concept they propose the following example: the built-in camera of a smartphone is a function of the phone; an offering is the phone's ability to record (e.g., images or videos of a person) and a possible result is the documentation of a human rights violation with that camera. They explain that functions or tools are static, while offerings are dynamic. Evans, Sandra K. and others , "Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research", in: *Journal of Computer-Mediated Communication,* Vol. 22, No. 1, Blackwell Publishing Ltd, 2017, retrieved from: http://bit.ly/2EEfYfM.

table is used to put objects or sit on it; a courier service is used to spread your own messages or those from strangers. French intellectual Bruno Latour defines design itself as a process of inscription of modes of use: "The resulting products carry with them the 'scripts' that inhibit or preclude certain actions while inviting or demanding others."[41]

When those objects, environments or technologies allow social actions, we are faced with a social affordance[42] In the context of social technologies, different studies have focused, for example, on how social networks are used by people to organize their private lives; by governments to have direct contact with their citizens or to monitor them; by companies to stimulate teamwork; by educational institutions to promote pedagogical purposes, or by political organizations to motivate citizen participation.[43] The affordances of the service are also influenced by the use policies and algorithms, which allow these interactions and access to content or users (through the recommendations of the News Feed, for example). This adoption of the product by the user and the communities helps identify how misinformation is designed, disseminated and consumed. In many cases, it is simply the common use of the service: a piece of fake news is an informative unit like any other. In other cases, it is an unintended consequence that the product offers: using a deceptive headline to generate clicks and viralize a lie. One way or another, these are uses that are not alien to the product.

Take the aforementioned Pizzagate example. The genesis of the story

[41] Dow Schüll, Natasha, *Addiction by Design: Machine Gambling in Las Vegas,* Princeton, Princeton University Press, 2012 (Ed. Kindle). Informal translation.

[42] "Along the same lines as Gibson's, our hypothesis is that the richest and most elaborate environmental affordances are provided by other animals and other people". Kaufmann, Laurence and Clément, Fabrice, "How Culture Comes to Mind: From Social Affordances to Cultural Analogies", *Intellectiva,* No. 46-47, 2007, retrieved from: http://bit.ly/2sv5lqD.

[43] See: Kursat Ozenc, Fatih and Farnham, Shelly "Life "modes" in social media": *CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Vancouver, May 7 to 12, 2012; Mergel, Inés, "Implementing Social Media in the Public Sector", October, 2013, retrieved from: http://bit.ly/2o6K0hT; Zeiller, Michael and Schauer, Bettina "Adoption, motivation and success factors of social media for team collaboration in SMEs",: *i-KNOW '11 Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies,* Graz, September 7 to 9, 2011; Wang, Qiyun, Lit Woo, Huay y Lang Quek, Choon, "Exploring the Affordances of Facebook for Teaching and Learning",: *International Review of Contemporary Learning Research*, No. 23-31, 2012, retrieved from: http://bit.ly/2Gd7ol3; Blegind Jensen, Tina and Dyrby, Signe "Exploring Affordances Of Facebook As A Social Media Platform In Political Campaigning",: *Proceedings of the 21st European Conference on Information Systems,* ECIS 2013 Completed Research, 2013, retrieved from: http://bit.ly/2ElGgxm.

was in social networks: a rumor shared on Facebook ("an anonymous source of the NYPD said ...") and replicated on Twitter. This false content arises organically, that is to say, through the normal information-sharing mechanisms of the service. The configuration of both platforms, which rewards interaction in various ways, allows consuming this content massively. Its veracity is irrelevant. At the same time, a website that produces fake news presents that rumor as an article and, in addition to publishing it organically, promotes it. That is to say, it pays a platform like Facebook to display that information as advertisement to get more interaction and visibility. At this point, the product offers a tool to make that advertisement reach a defined audience. It is logical to assume that the inscribed use -the desired use - of the service is to promote truthful products, but its design offers - affords - the possibility of promoting fraudulent information.

Let's look at Twitter. The design of the network allows and promotes the creation of open and decentralized conversations. This offer has allowed planning social protests, rights movements and countless joint actions. But the incorporation of the product allows these joint actions to be concerted, which opens the possibility for a group of apparently authentic accounts to devote themselves to promoting false content.[44] And in this open and participatory environment, many users consider that this volume of exchanges is authentic and that, therefore, the information conveyed is accurate.

Let's finish with WhatsApp. As a messaging service, it is designed to preserve the privacy of communications. This closed configuration makes it impossible, in principle, to monitor the contents, either by the platform itself or by third parties (hence it is known as a "dark social" environment).[45] Furthermore, the only people who can share this content are the contacts that are part of a conversation -individual or collective-.[46] As individuals are the ones sharing videos, images, audios or texts, the information does not necessarily include a source, but it is still shared in a context of intimacy and

[44] A study carried out by the University of Edinburgh identified hundreds of false accounts operated from Russia with the aim of influencing the referendum on the permanence of the United Kingdom in the European Union (Brexit). Booth, Robert and others, "Russia used hundreds of fake accounts to tweet about Brexit, data shows", The Guardian, November 14, 2017, retrieved from: http://bit.ly/2iUPF8b.

[45] Communications surveillance schemes are not ruled out, but the central point is that the product is not designed to generate and measure interaction and consumption of content such as Facebook or Twitter.

[46] If an unknown person gets one's phone and shares something on WhatsApp, it will possibly have an alienating effect. The recipient will reject the message or will not give relevance to the information.

with a known interlocutor, all of which gives it legitimacy. The object has the designed use of connecting known people in a closed environment, which does not mean that this channel is not used to share any type of information.

Being aware of these differences is important for two reasons. On the one hand, it allows the understanding of the proposed solutions and identifying their inherent limitations. On the other, it places the problem of fake news in the orbit of the social incorporation of a technology. The use of technology - and therefore, the questions that arise from using it - is a process mediated by people and not an isolated equation: "Technologies are not inserted in everyday life, causing a revolution or a radical break as people say; on the contrary, this insertion usually entails a gradual evolution, a negotiation between the inherited practices and the desire for the transformation of societies".[47]

IV. The solution to the problem

Long before a debate arose around misinformation, social networks already faced the general challenge of moderating content online. Arbitrating the flow of information is perhaps the greatest sign of the power of these platforms in their condition of intermediaries:[48] They do not produce the content, but they make important decisions about that content: what is going to be distributed and to whom, how are users going to be connected and how are their interactions going to be managed, and what is going to be rejected".[49] By moderating content, these actors try to apply their community rules to build user loyalty with the service and maintain an operation free of undesired interference.

Hate speech and terrorism, discrimination against minorities, harassment against women and toxic content, in general, forced these companies to seek a complicated balance between the free circulation of content and timely restriction. In this practice, they increasingly face questions about the transparency and accountability of these processes. The role of algorithms in content decisions, the suspension of accounts and the appeal mechanism,

[47] Gómez, Rocío and others (comp.), *Facebook como obra mundana. Poetizar la vida y recrear vínculos personales*, Universidad del Valle, Editorial Program, 2016, pp. 66.

[48] *See* Cortés, Carlos "Las llaves del ama de llaves: la estrategia de los intermediarios en Internet y el impacto en el entorno digital", in: Bertoni, Eduardo (comp.), *Internet y Derechos Humanos. Aportes para la discusión en América Latina.* Buenos Aires, CELE, Universidad de Palermo, 2014.

[49] Gillespie, T. Cited in: Myers West, S. 'Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms', Media and Communication (ISSN: 2183–2439) 2017, Volume 5, Issue 3, p. 28 (informal translation).

among others, are part of an agenda of demands that both governments and civil society ask from these companies.[50]

This is the context in which companies like Facebook, Google and Twitter try to respond to the problem of fake news. And although it would be a topic to develop in another opportunity, it is relevant to locate those answers within the content moderation policies already in use by these platforms -and not as a separate issue-. For example: while Facebook has a policy of real names, Twitter does not prohibit pseudonyms or parody accounts. This starting point delineates different developments regarding the moderation of misinformation.

1.Facebook

In mid-2017 the English newspaper The Guardian, through the project "The Facebook Files", offered a panorama of the company's moderation of content practices.[51] Instances of fake news in particular are part of what they call "information operations", which Facebook interprets as "the actions taken by actors organized to distort the national or foreign political sentiment."[52] Information operations are divided into fake news, misinformation and false accounts. Note how the company recognizes the joint and organized nature of these actions, understanding that it is an adapted and particular use of the product.

With that understanding, Facebook has announced solutions to face fake news with a focus on two areas. On the one hand, it has announced measures to promote literacy in news (known as news literacy) in its users, to help them make informed decisions about news and sources that can be trusted. To this end, it has developed two projects: the Facebook Journalism Project and New Integrity Initiative. [53] These preventive and educational measures

[50] See, among other initiatives, www.onlinecensorship.org and http://responsible-tech.org/. The Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of the United Nations, for example, will make a report in June 2018 on the regulation of content in the digital age.

[51] See, "The Facebook Files", *The Guardian*, retrieved from: http://bit.ly/2r6h5yb.

[52] Facebook, "Information Operations and Facebook", April 27, 2017, retrieved from: http://bit.ly/2oOOS9s.

[53] Through this initiative, Facebook offers training tools for journalists. "Facebook Journalism Project", http://bit.ly/2ioDPAO, last access: December 14, 2017. The News Integrity Initiative is Facebook's big bet on the subject of news literacy. This project led to the creation of a consortium of technology industry leaders, academic institutions and other organizations, which will aim to help people make informed decisions about the news they read and share on the Internet. "News Integrity Initiative", http://bit.ly/2D5udGi, last access: December 14, 2017.

are not addressed in this document. On the other hand, it announced the adoption of technical measures that have an impact on the contents received by users on the platform. As published by the company, these are the characteristics of said measures:

1.1. Report, fact-checking and flagging

In December 2016, Facebook announced that it was testing different mechanisms to facilitate its users the reporting of possible fake news.[54] Community reports and "other signs" -not specified by Facebook- would be considered to select stories that they would send to independent organizations to do fact-checking. Facebook does not detail how many reports are necessary to generate an alert.

If the organizations dispute the content after the verification process, Facebook will display warnings (flagging) indicating that the article has been disputed. A link to get more detailed information will be included next to the flag. If despite these messages a user wants to share the content, Facebook can show a new message in which it warns the user that they will share content that has been disputed. Moreover, according to Facebook, these stories may have less prevalence in the News Feed, and may not be promoted or converted into advertisements.

Mark Zuckerberg has said that Facebook does not want to be an arbitrator of the truth.[55] The task of verifying the facts was entrusted to external organizations that are signatories of Poynter's International Fact-Checking Network's code of principles.[56] Initially, the company is working with ABC News, FactCheck.org, Associated Press, Snopes and Politifact.[57] Between the end of 2016 and 2017, these measures have been announced as trials in countries such as the United States, Germany, France and the Netherlands, and are not available permanently.[58]

---

[54] "Enfrentando engaños y noticias falsas", http://bit.ly/2HgPW0k, last access: December 14, 2017.

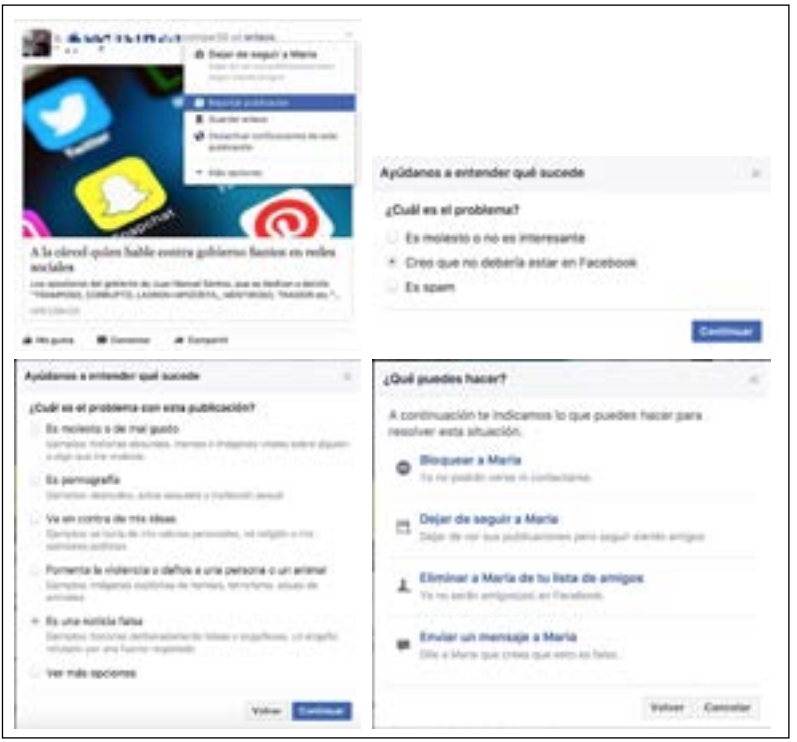
[55] http://bit.ly/2fOsvha

[56] International Fact-Checking Network, "International Fact-Checking Network fact-checkers' code of principles", retrieved from: http://bit.ly/2BwtakU.

[57] Mullin, Benjamin, "Facebook rolls out plan to fight fake news", *Poynter,* December 15, 2016, retrieved from: http://bit.ly/2h5bPm5.

[58] See: Facebook, "News Feed FYI: Addressing Hoaxes and Fake News", December 15, 2016, retrieved from: http://bit.ly/2gFFvVw; Facebook, "Umgang mit Falschmeldungen", January 15, 2017, retrieved from: http://bit.ly/2gFveyn; Facebook, "Facebook annonce le lancement de son outil de fact-checking en France", February 6, 2017, retrieved from:

When making this announcement, Facebook published some images of the reporting scheme. A test made from a connection with IP in the United States confirmed the mechanism. However, the same exercise carried out in August of 2017 from an IP in Colombia showed that although users could point to the option "It is fake news", it was not possible to send a report to Facebook. The only options available were to block, stop following, delete or send a message to whomever had shared the disputed information. A new test carried out in December 2017 shows that the reporting option is disabled in both countries, at least with our Facebook user.



Note: screenshots of the fake news reporting scheme.

http://bit.ly/2swl4Ws; Facebook, "Addressing Fake News Globally", January 15, 2017, retrieved from: http://bit.ly/2jyZ1FV.

1.2. Changes in the news feed and measures to counter the false amplification.

As stated above, the content that the verifiers mark as false may have less relevance in the News Feed. Additionally, there are other signs that can lead to Facebook giving less relevance to these contents: for example, if reading an article people are less inclined to share it, it can be a sign that the story is misleading.[59] Or when users regularly post many publications per day with external content that can be considered of low quality -deceptive or sensationalist- the publication with an external link can be "punished".[60] On the other hand, additional measures have been announced in the news feed:

- According to an announcement made in April 2017, Facebook is trying to make it so that articles related to a publication appear before the user reads the desired content, so that they have quick access to other perspectives on the same subject, including articles that have passed the fact checking filter.[61]

Note: Related articles in Facebook

[59] Facebook, "Enfrentando engaños y noticias falsas", December 15, 2016, retrieved from: http://bit.ly/2HgPW0k.

[60] Facebook, "News Feed FYI: Showing More Informative Links in News Feed", June 30, 2017, retrieved from: http://bit.ly/2uzG8HM.

[61] Facebook, "News Feed FYI: New Test With Related Articles", April 25, 2017, retrieved from: http://bit.ly/2q2zxHe.

- In an August 2017 update, Facebook announced that it will start using automated learning (machine learning) to detect more deceptive content. Facebook will be able to show the fact-checking stories below the original articles.
- Since July 2017, Facebook has eliminated the option to customize the image, title or preview description of the links that are published in Facebook profiles or groups. These previews will depend solely on the metadata of the website that is linked.[62] This prevents modifying this information to attract clicks through headlines or misleading images.
- Since 2013, Facebook has announced the adoption of a new algorithm to detect and give more relevance to "high quality" content. At the time, for example, whether the content was relevant or if the sources were reliable were both taken into consideration.[63] On this matter, in 2016 Facebook established a policy to prevent advertisers with low quality content pages from advertising on the platform.[64]

Facebook has also announced measures to counter "false amplification", which it defines as the coordinated activity that seeks to manipulate the political discussion.[65] The creation of false accounts (often executed on a large scale) is part of this irregular practice; the coordinated distribution of content or repeated messages; the likes or coordinated reactions; the creation of groups with the purpose of distributing sensationalist news or headlines, and the creation of memes, videos or manipulated photos; among others.

To this end, the company announced the development of technological tools that allow it to identify false likes and false comments that come from false accounts, malware or so-called "click farms" (groups of accounts that have an apparently organic activity that are created to generate a false interaction).[66]

[62] Facebook, "API Change Log: Modifying Link Previews", retrieved from: http://bit.ly/2u1ndq6.

[63] Facebook, "News Feed FYI: Showing More High Quality Content", retrieved from: http://bit.ly/2BZYU2U.

[64] See: "Contenido molesto o de baja calidad", retrieved from: http://bit.ly/2q9hkFk, last access: December 14, 2017.

[65] Facebook, "Information Operations and Facebook", April 27, 2017, retrieved from: http://bit.ly/2oOOS9s.

[66] See: Facebook, "Breaking new Ground in the Fight Against Fake Likes", April 17, 2015, retrieved from: http://bit.ly/1yCsKDM; Facebook, "Disrupting a major spam operation", April 14, 2017, retrieved from: http://bit.ly/2oQJZQX.

2. Google

According to the New York Times, last October fake news ads used Google's advertising system and even appeared in fact-checking portals such as Snopes and Politifact. A fake Vogue website announced, for example, that the first lady of the United States, Melania Trump, had refused to live in the White House.[67] However, this was not Google's only front of misinformation. Fake news portals have appeared in its search engine results, and videos of conspiracy theories are frequently recommended in YouTube.[68]

Some measures taken by the company to deal with misinformation are focused on digital literacy about news consumption, such as the Internet Citizens program on YouTube, which gives workshops to young people between the ages of 13 and 18. On the other hand, Google finances hundreds of projects in Europe to produce original journalism and to guide citizens on reliable content.[69] Beyond this, the response of the service has focused on two fronts: search services (Google Search and Google News) and advertising services (Google AdWords, Google AdSense and YouTube).

2.1. Changes in search services: Google Search and Google News

Google has announced several changes that could affect the experience of users of the search service: fact checking for search results, changes in the search algorithm and user feedback. Moreover, Google has announced that it will make the way searches work more transparent.

- Fact checking

Since 2009, Google has implemented a labeling system to mark some of the results shown in Google News searches.[70] For example, the content opinion, Blog or Satire labels help the users to identify the type of content they will encounter when opening an article. In October 2016, Google announced that it would implement the Fact checking label, with which it intends to label articles that have gone through a fact-checking process.

[67] Wakabayashi, Daisuke and Qiu, Linda, "Google Serves Fake News Ads in an Unlikely Place: Fact-Checking Sites", *The New York Times,* October 17, 2017, retrieved from: http://nyti.ms/2hO00Ca.

[68] Abrams, Dan, "Now Even Google Search Aiding in Scourge of Fake, Inaccurate News About Election 2016", *Mediaite,* November 13, 2016, retrieved from: http://bit.ly/2C0uaic.

[69] In this regard, see the Cross Check project and the Digital News Initiative.

[70] Google, "¿Qué significa cada etiqueta (p. ej., "blog")?", http://bit.ly/2F2tiYW, last access: December 14, 2017.

For example, for months President Trump has said that his government will make the largest tax cut in history. A search of this topic in Google News (November 2017) shows among the results an article of FactCheck.org marked with the label Fact-Check, where it is explained to which extent is that statement truthful and plausible.

Note. The FactCheck.org story appears under the subtitle "Related Coverage". Taken from Google News.

Originally, this function was available in the United States and the United Kingdom. However, on April 2017, Google announced that the label would be available worldwide and that it would be extended to its general Google Search system in all languages.[71] Thus, when a user does a Google search and the results show contents that have been verified, Google will display this information indicating what was said, who said it, who checked the information and the result of that checking.[72] The result of the checking can be not only true or false, but also mostly true or partially true. To index this information, Google starts with criteria such as the reliability of the source, checked facts, sources and quotes, and the conclusions reached through this review.[73]

[71] Google, "Fact Check now available in Google Search and News around the world", April 7, 2017, retrieved from: http://bit.ly/2tZyRSf.

[72] Different verifiers can reach different conclusions. Several conclusions can be presented in the results.

[73] See data verifications in the search results, http://bit.ly/2CjpcJ5, last access: December 14, 2017.

Publishers who want their fact checks to show up in search results can point them to Google in two ways: i) using the ClaimReview label[74] in the publication code, based on Google's guidelines for data verifiers; ii) using the Share The Facts widget, which acts as a quality seal for participating organizations and can be embedded - as if it were a tweet - on a web page.[75] We see then how Google limits itself to highlighting the checks made by third parties based on certain criteria, even if they have reached different conclusions.[76] In contrast, Facebook incorporates the checking of some organizations in its own conclusions, either to decide the prominence of the story or to make a particular warning.

Note. Examples of how checked information appears in Google searches.

[74] See: Schema.org, "ClaimReview", retrieved from: http://bit.ly/2EpwIDP, last access: December 14 2017; Google, "Verificaciones de datos", retrieved from: http://bit.ly/2CmGPrd, last access: December 14, 2017.

[75] Share the Facts, "Share the Facts", retrieved from: http://bit.ly/2srvQxi, last access: December 14, 2017. To date, it works together with twelve organizations, none of them in Latin America: PolitiFact, The Washington Post, FactCheck.org, Gossip Cop, Pagella Politica, AGI, La Stampa, The Ferret, Climate Feedback, Demagog, Newsweek and VERA Files Fact Check (http://www.sharethefacts.org/about/).

[76] Google, "Fact Check now available in Google Search and News around the world", April 7, 2017, retrieved from:http://bit.ly/2tZyRSf.

– Improvements in the search ranking

Although Google states that only 0.25% of searches show offensive or deceptive content, in April 2017 it announced changes in Google Search so that the results reflect more reliable content.[77] This includes updating the guidelines used by teams that evaluate the quality of search results.[78]

– Feedback from users

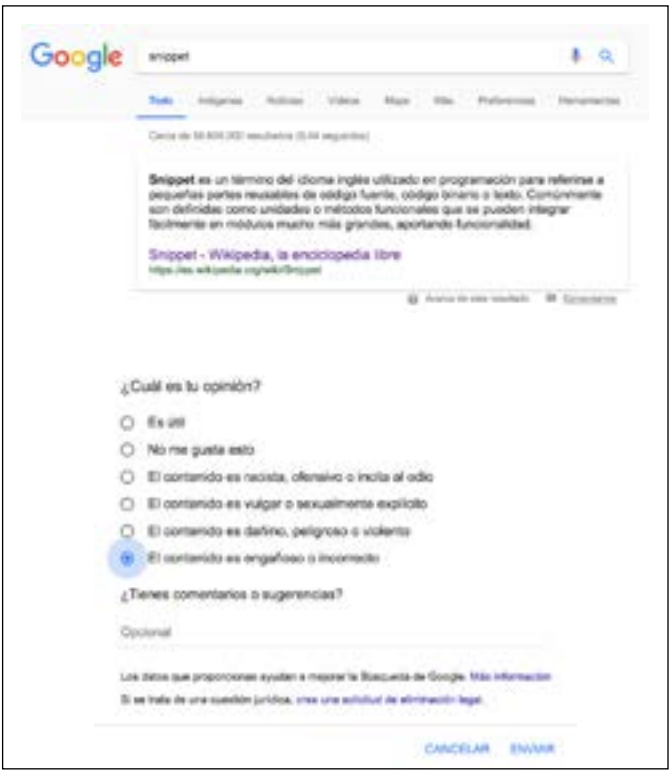
In April 2017, Google announced that it would improve the users' reporting mechanisms in relation to the terms suggested in the autocomplete function and the contents highlighted in the search results (featured snippets). According to Google, new mechanisms for obtaining feedback include pre-established categories of offensive content that facilitate reporting. These reporting mechanisms not only work to signal fake news, but also other problematic content (sexually explicit, violent, dangerous, etc.).[79]

Note. Users can send comments on the autocomplete function by clicking on the option 'Report offensive query', which displays several reporting options.

[77] Google, "Our latest quality improvements for Search", April 25, 2017, retrieved from: http://bit.ly/2sya4ro.

[78] Google, "General Guidelines", July 27, 2017, retrieved from: http://bit.ly/1ZdLFyy.

[79] Google, "Our latest quality improvements for Search", April 25, 2017, retrieved from: http://bit.ly/2sya4ro. Before the announcement, it was possible for a user to report the terms suggested by the autocomplete function if they violated Google's AutoComplete Policy, however, it was necessary to go to the help page, find and fill out a form (http://bit.ly/2C0rj95).



Note. Reports on outstanding results in the form of snippets can be made by clicking on the 'Comments' option and then pointing to 'The content is deceptive or incorrect'.

According to Google, this feedback is used to evaluate if the changes introduced in its search system are successful and if they should be applied to other users. Moreover, the company gave more information on how Google Search works and disclosed a few content policies for its autocomplete service.[80]

2.2. Changes in advertising services: Google Adwords, Google Adsense and YouTube

In March of 2017, an investigation by The Times revealed that YouTube ads of L'Oréal, The Guardian, Nissan, the BBC, among many others, appeared while playing videos with extremist, anti-Semitic or homophobic

[80] Google, "Políticas de autocompletar", http://bit.ly/2EqUeRb, last access: December 14, 2017; Cómo ayudar a mejorar los resultados de búsqueda de Google", http://bit.ly/2o9oqci, last access: December 14, 2017; Google, "How Search works", http://bit.ly/2rudrP8, last access: December 14, 2017.

messages.[81] According to the research, those who post videos on YouTube generally receive up to $ 7.6 per thousand reproductions, and some videos had millions of views. After the disclosure, major advertisers such as PepsiCo, Starbucks, Walmart, AT&T, Verizon, Volkswagen, Johnson and Johnson, decided to withdraw their ads, in an event that was known in the media as the YouTube Ad Boycott.[82]

In response, Google published a statement acknowledging that, despite all efforts, its technological tools did not always detect bad ads or advertisers that violate its policies. A few days later, the company announced, among several measures, the strengthening of its policies on hate or offensive content, the creation of mechanisms for advertisers to control where their advertising appears, and the development of new tools to check questionable content.[83]

The problem of ads with undesirable content is the same as with fake news: advertisers do not want their products associated with misinformation. Before the controversy broke out, Google was already working to identify "bad ads": for example, products that falsely promise to help lose weight, illegal or counterfeit products or ads that carry malicious code or viruses.[84] Along the same lines, on November 2016, Google introduced changes to its Google AdSense policy - its ad network that appears on different Internet sites - and prohibited the publication of ads on websites that distort or conceal information about the publisher, its contents or the main purpose of the site.[85] This policy, however, does not refer to fake news but to false representation.[86]

In the latest report on the implementation of its policies to combat prohibited advertisements and sites, Google reported that after introducing

[81] Mostrous, Alexi, "YouTube hate preachers share screens with household names", *The Times,* March 17, 2017, retrieved from: http://bit.ly/2HfmUON.

[82] Solon, Olivia, "Google's bad week: YouTube loses millions as advertising row reaches US", *The Guardian,* March 25, 2017, retrieved from: http://bit.ly/2n41Ccw.

[83] Google, "Improving our brand safety controls", March 17, 2017, retrieved from: http://bit.ly/2uPlDdc; Google, "Expanded safeguards for advertisers", March 21, 2017, retrieved from: http://bit.ly/2EJTCJT; YouTube, "YouTube Partner Program overview", http://bit.ly/1bQpskt, last access: December 14, 2017.

[84] Google, "How we fought bad ads, sites and scammers in 2016", June 25, 2017, retrieved from: http://bit.ly/2jpTTWP; Google, "How we fought bad ads, sites and scammers in 2016", June 25, 2017, retrieved from: http://bit.ly/2jpTTWP.

[85] Google, "Prohibited content", http://bit.ly/2nZvWYe, last access: December 14, 2017.

[86] In fact, in January 2017, the Media Matters organization for America reported that Google removed this expression which was included before as an example of forbidden content. "Google Quietly Removes "Fake News" Language From Its Advertising Policy", Media Matters, January 12, 2017, retrieved from: http://bit.ly/2BZZY6U.

said changes, actions were taken against 340 sites for false representation and other offenses.[87] Two hundred of these sites were definitively expelled from the company's advertising network. However, Google did not clarify how many of them were cases of fake news. The closest thing to the issue appears in an excerpt in the report where it reveals the blocking of more than 1,300 accounts of tabloid cloakers, portals that publish ads claiming to be news headlines but directing them to sites with advertising.

3. Twitter

Despite having 16% of the users Facebook has, Twitter has also become part of the controversy surrounding misinformation. According to the company's own statements at the hearings with the US Congress, more than 2,700 accounts associated with Russian agents moved 1.4 million tweets between September and November 2016.[88]

This is not a completely new problem for Twitter, which is constantly questioned due to false accounts and bots in their platform. According to the company, this type of accounts does not represent more than 5% of its users, but external sources assure that there are many more. For example, according to a study conducted by Alessandro Bessi and Emilio Ferrara, bots were responsible for a fifth of the tweets related to the US presidential elections.[89] Nonetheless, it is important to remember that Twitter allows automated accounts, and many of them fulfill a relevant service in terms of information (news channels, government services) and are not considered spam.

Twitter has acknowledged that it is impossible to determine the veracity of the published tweets and, like Facebook, has maintained that it does not want to be an arbiter of the truth.[90] If Twitter starts to evaluate information it would not only be undesirable, but impossible to implement in practice: at least one billion tweets per day pass through the platform. Thus, the focus of the company's response is to detect and remove accounts that, in an automated or manual manner, disseminate malicious content (spam, falsehoods or attacks, among others). These are some of the concrete actions

[87] Google, "How we fought bad ads, sites and scammers in 2016", January 25, 2017, retrieved from: http://bit.ly/2kclNnZ.

[88] Solon, Olivia and Siddiqui, Sabrina. "Russia-backed Facebook posts 'reached 126m Americans' during US election", October 31, 2017, retrieved from: http://bit.ly/2hoiMRe.

[89] Bessi, Alessandro and Ferrara, Emilio. "Social bots distort the 2016 US presidential election online discussion", November 7, 2016, retrieved from: http://bit.ly/2oZNst8.

[90] Twitter, "Our Approach to Bots & Misinformation", June 14, 2017, retrieved from: http://bit.ly/2HhmFCC.

announced by the company:

- Reducing the visibility of tweets and possible spam accounts while investigating whether a violation actually occurred.
- Suspension of accounts once prohibited activity has been detected.
- Implementing measures to detect applications that abuse Twitter's public API.[91]

Twitter considers that any automatic detection system carries a high risk of false positives; that is, accounts that the algorithm considers may be violating the policies but in reality have a legitimate behavior. For example, an account of an activist who is tweeting constantly can be confused with a bot or with a person who is deliberately sending spam. On the other hand, misinformation in the platform is carried out through joint actions between groups. These are coordinated operations that comply with the rules and restrictions of the platform to achieve their purpose. "It is much more complicated to detect non-automated coordination", explains the company. "The risks of inadvertently silencing legitimate activity are much higher".[92]

V. Conclusion: the problem of the solution

The main objective of this document was to expose the measures that some intermediaries -mainly Facebook and Google- have set up to combat misinformation. After explaining the problem and placing it in the territory of each platform, we describe the proposed solutions, most of which are in the preliminary phase. In this final part of the text we detail some of the problems that these solutions entail and we make some recommendations. To carry out this analysis, we propose four points: i) the scale and time of the solution; ii) the impact; iii) the role of civil society, and iv) transparency. This analysis starts from a basic assumption that we described at the beginning and we reiterate now: it is neither possible nor desirable that the solution to fake news be automatic. It is a phenomenon with technological incorporation, which is social by definition.

This analysis focuses on organic content and not on the promoted one. The pieces of fake news that are promoted commercially present some challenges and different demands for the platforms. It would be reasonable to demand a higher level of monitoring and control in that case, since they

[91] API is an application programming interface that allows an external software to use several functionalities of a service, in this case, Twitter

[92] Op. Cit., Twitter, "Our Approach to Bots & Misinformation". Informal translation.

receive a direct profit for advertising information for commercial purposes. Organic content is the one which users disseminate among their contact networks, and in contrast to commercial advertisements it includes spontaneous user information, manipulation and coordinated actions in the same degree. This is where we find the greatest challenges for freedom of expression and, therefore, where we want to direct our attention.

1. Scale and time

To analyze these two factors, it is necessary to return to the concept of affordances. As explained above, the response to misinformation must necessarily start with the configuration of the space where the problem occurs. This demonstrates a structural issue: some elements that enable misinformation could only be eliminated in the design of the service itself. These are limitations inherent to the architecture of the space. For example, if Facebook introduced a system of prior review of all content: if it were possible, it would eliminate in equal parts misinformation and a lot of legitimate content. But as Facebook is not going to modify that substantial aspect of its structure and space, the "offering" of this platform limits the scope of the solution.

Facebook and Google, YouTube and Twitter, "offer" a space where content can be disclosed without prior review. This means that, in the vast majority of cases, the strategies to deal with misinformation will be some form of subsequent control: labels, related articles, less visibility, etc. That type of control, as we saw, is not susceptible to total automation, and to that extent there is a challenge in terms of scale and time.

The scale refers both to the company's teams and to the work with external actors. Can the company replicate the initiatives against misinformation for all its users (more than two billion in the case of Facebook)? Does it have enough internal human teams to review all the questionable content? Are there enough fact-checkers for all the available information? The answer will invariably be negative. If the strategy to deal with misinformation requires a human effort - which is also desirable- it will not be possible to meet all the demand of the problem.

Last September, Facebook carried out several preventive actions to avoid misinformation during the German elections: suspension of dozens of accounts, direct work with local authorities and educational campaigns.[93] With

[93] Facebook, "Update on German Elections". September 27, 2017, retrieved from: http://bit.ly/2Buursp. See also: Locklear, Mallory. "Facebook axed 'tens of thousands' of accounts before German election". Engadget, September 27, 2017 retrieved from:

the antecedent of the North American elections, and this time in the midst of the tense European regulatory context, the company could not sit idly by.

This type of efforts could hardly be deployed in markets with less relevance, as are the vast majority of Latin American countries. The scale of the solution, then, is mainly focused on key countries, and although some of the answers can be extended to other markets, they are thought and located for those priority scenarios (language, problem approach, solutions, and context). In the case of Colombia, as we have seen, the tool to report fake news has apparently not been available at any time. The fact-checking schemes, on the other hand, could be done in countries of the region (Share the Facts, we repeat, does not include Latin American organizations at this time), but they require local joint efforts that will not be easily applied to all of Latin America.

The limitation in scale is related to the time variable. The risk of misinformation in the public debate is even higher during election time. A false news or a rumor spread as truth can affect the outcome of an election. To that extent, a timely response to address this problem is desirable. A post-mortem action is relevant to understand the phenomenon, but not to face its immediate effects.

At key moments in a campaign, the responses described in this document may be too late. While building a lie is quick, verifying a fact requires time. It is essential to be aware of this limitation, not so much as an argument to justify more restrictive measures for public debate, but to understand the reality we face and the scope of the proposed solutions.

2. Impact

The scale and time of the solution influence its impact. On the one hand, a set of partial and isolated actions, many of them late, will hardly serve to combat misinformation in the public debate. On the other, these actions may have an effect contrary to what is sought.

On the possible undesired effects, the evidence is still unreliable. Nonetheless, some studies indicate that verification tools and warnings can have a negative impact. According to an investigation by a group of academics -also under review-, labeling fake stories as such does not necessarily change the user's perception of them. Furthermore, which is even more serious, if a user begins to see stories where there are warnings of possible falsehood, they may conclude that all those that do not have a warning are true, which, of course, is a wrong generalization.[94]

http://engt.co/2k1Etdw.

[94] See, Pennycook, Gordon, et. al. "Prior exposure increases perceived accuracy of fake

Regarding the first point, this document does not intend to disqualify actions such as fact-checking, using warnings or context articles. These are answers that seek to ponder the problem of misinformation guaranteeing the freedom of expression of users. Discarding restrictive or openly arbitrary measures hinders the solution of the problem, but prevents the creation of worse ones. From this perspective, it is also relevant to analyze the decisions to hide or remove contents under potentially arbitrary categories, such as "low quality" or not being a "reliable" source. These schemes based on reputations run the risk of favoring mainly mass and commercial media, to the detriment of voices that cannot access this type of alliance, but whose content is legitimate and relevant.

The measures to deal with misinformation are located in the context of the terms of service of the platform and, in particular, the moderation of content. Therefore, following the proposal of the Special Rapporteurship for Freedom of Expression of the Inter-American Commission on Human Rights, "in the design and setting of their terms of service and community rules, companies should not limit or restrict the freedom of expression disproportionately or unnecessarily".[95]

3. The role of civil society

Most strategies to deal with misinformation include the participation of civil society. Whereas companies work with organizations to verify information, make warnings or provide more context. This participation is essential, as long as these initiatives are set regionally, they should include organizations that have legitimacy and knowledge to weigh information about the public debate. The user, however, is less important in this discussion. Except for Google feedback, it is not clear how the user feedback is taken into account to address misinformation in these services.

Finally, at this point it is necessary to take into account that civil society is also an actor in the production of misinformation: political parties, communication agencies and different interest groups are part of the problem which is currently expected to be solved directly by platforms. Concerted

news". Version of August 26, 2017; Pennycook, Gordon, et. al, "Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy". Version of September 15, 2017. These texts have not had peer review.

[95] Special Rapporteurship for Freedom of Expression of the Inter-American Commission on Human Rights, Standards for a Free, Open and Inclusive Internet OAS OEA/Ser.L/V/II CIDH/RELE/INF.17/17 March 15, 2017, Original: Spanish, Para. 98

actions, for example, show an organized and systematic process to exploit the services for the benefit of an individual purpose and to the detriment of the general interest.

While this does not remove liability from intermediaries, it does highlight the need to seek answers in addition to those of a preventive and educational nature, in different economic, political and social sectors. In other words, the solution to this problem is not limited to the traditional actors of Internet governance.

4. Transparency

The development of this document had a constant difficulty: to understand if the large number of decisions and actions covered in the media and announced by the companies were being effectively implemented and to what extent. In fact, as these words are being written, measures that would reverse the changes announced recently and referenced in this text are being reported.[96] Ultimately, the conclusion on this point is that many of the measures are just announcements, and those that are implemented reach only a partial and temporary degree of application. In any case, it is not possible to determine it accurately, which shows a serious problem of transparency.

It is understandable and desirable that companies experiment with possible solutions. Misinformation must be faced with creative and innovative actions, and strategies cannot be set in stone. However, the lack of clarity for users and the general public obscures the understanding of the problem and prevents civil society from making prudent feedback.

Technology produces informative returns as its adoption becomes more widespread. Through this use - and the network effect - data is obtained which serves to improve and adapt this technology.[97] That social return must also be widespread, or at least have some degree of openness. Knowledge of how misinformation works and measures to address it are controlled by companies that are unwilling to share it with civil society. There is no clarity in the diagnosis or in the responses adopted, all of which adds to the existing questions against intermediaries for their lack of transparency in the way they provide their service. Transparency, to this extent, is not a subsequent unilateral action, but a joint process that is part of the solution.

[96] Cohen, David. "Facebook Is Opting for Context Over Labeling in Its Battle Against Fake News". *Adweek*, December 21, 2017. Retrieved from: http://bit.ly/2Ero77P.

[97] See MacKenzie, Donal. Knowing Machines. Essays on Techincal Change. MIT Press, 1998.

Chapter Four

Over-the-Top Services: fundamental principles for discussing their regulation in Argentina

Maia Levy Daniel*

I. Introduction

Communication services and the media are subject to different regulations; either because of their huge infrastructure, because of the use they make of limited resources often owned by the State, for economic reasons, or in order to protect human rights potentially affected by these activities. In Argentina, as in many other countries, there are laws to protect competition, avoid monopolies and oligopolies, and to guarantee access and non-discrimination of minority populations, among others. In this sense, the Law on Audiovisual Communication Services [Ley de Servicios de Comunicación Audiovisual] — with specific derogations of December 2015 — regulates different aspects of the media market to avoid the existence of monopolies that constitute indirect violations to freedom of expression and access to information, and establishes guidelines around the contents. Furthermore, the Argentine Digital Law [Ley Argentina Digital] regulates the broadband Internet access market in the country. This law has different aspects and different purposes. On the one hand, it regulates infrastructure and, on the other, it regulates content, in order to guarantee the plurality of voices that a democratic society requires. However, neither law took fully into account the growth and expansion of the Internet.

---

* Maia is a lawyer who graduated from the Universidad Torcuato Di Tella, Master in Public Affairs in the same university and Master of Laws (LL.M.) at Harvard University. She worked as a researcher at CELE until March 2018. This document was written under the direction and with comments of Agustina Del Campo, director of CELE.

** This document was drafted in November 2016.

Internet has stimulated the distribution of content, creativity and entrepreneurial spirit that led to the incorporation of new services and players equivalent or complementary to existing ones. By creating new services, in many ways different from those traditionally offered by companies specialized in communications and information transmission, business models have also changed. High-speed networks and the increase of smart devices, including phones, tablets, laptops, etc., contribute to the rapid expansion of these new services, blurring the boundaries between digital devices and causing a shift for many communication services from the analog to the digital space.[1]

Although technically there is no unanimous definition of the term, certain sectors agree to use "over the Top" services or "OTTs" for *"online services that can replace traditional media and telecommunications services at some level."*[2]

Some authors define OTTs as a group of actors,[3] while others consider it a term to qualify a category of services.[4] On some occasions, they have been defined as *"any content, service or application that is provided to the end-user on the open Internet."*[5] However, the common denominator among the different definitions is that OTTs are characterized by the method used to provide the service — i.e. the Internet — and not by the type of service offered.[6] The OTTs fully depend on the Internet to provide the service.[7]

The rise of these new industries presents great challenges for traditional service companies and the State regarding their regulation.[8] Netflix, YouTube, Skype, Airbnb, and WhatsApp, among many others, are typical examples of OTT services offered around the world. Many of these services have replaced traditional forms of communication and telecommunications

[1] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Una Agenda Digital para Europa* [Digital Agenda for Europe], COM(2010) 245 final/2, Brussels, August 26, 2010, p. 4.

[2] European Parliament, Policy Department A: Economic and Scientific Policy, Directorate-general for internal policies, *Over-the-top players (OTTs)*, Study for the IMCO Committee, 2015, p. 11.

[3] Body of European Regulators for Electronic Communications, *Report on OTT services*, BoR (16) 35, January 2016, p. 6.

[4] *Idem*.

[5] Body of European Regulators for Electronic Communications, *supra* note 3, p. 3.

[6] European Parliament, *supra* note 2, p. 20.

[7] *Idem*.

[8] Ganuza, Juan José and Viecens, María Fernanda, *Over-the-top (OTT) applications, services, and content: implications for broadband infrastructure*, Work Document No. 13, Centro de Tecnología y Sociedad, Universidad de San Andrés, February 2013, p. 9.

in different aspects, offering users new messaging, audio and video services, among others. Moreover, they have given users the possibility of interacting with the content, giving them a much more active role than they had in traditional communication services. OTT services offer the added value of "being able to choose", compared to the traditional linear television model, for example.[9]

There are different types of applications in the digital field that refer to very diverse services. As Internet services grow in local markets worldwide, different sectors raise their voices to demand an overview of their regulations and/or their compatibility with human rights potentially affected by these new actors.

This article seeks to identify the fundamental principles that must be taken into account when regulating the Internet, and the basis shared by OTT services, applicable to all in equal measure. For a better understanding of the subject, this paper offers a description of the current regulatory situation in Argentina, as well as arguments that are normally used for and against the regulation of OTTs. Finally, it includes a section with conclusions and useful recommendations for legislative debate.

II. OTT services: what they are and what their presence in the market means

OTT services appeared on the scene recently. Skype, for example, was designed in 2003 and WhatsApp was created in 2009. Netflix, in turn, did not enter Latin America until September 2011 and has less than two decades in the US market where it was created. Despite its short history, the importance of OTTs has grown rapidly in recent years, becoming a very important part of the information and communication technology industries, as well as for consumers and businesses.[10] So much so that, in 2015, seven out of ten Internet users in the United States accessed online videos through OTT providers.[11]

[9] Conseil de l' Audiovisual de Catalunya, *Las Plataformas OTT para la distribución de contenidos audiovisuales: ¿una amenaza para el duopolio de la televisión en abierto en España?* [OTT Platforms for the distribution of audiovisual content: A threat to the open television duopoly in Spain?], in *Audiovisual OTT, nuevas fronteras y desafíos*, Quaderns del CAC No. 42, Vol. XIX, July 2016, p. 24.

[10] Body of European Regulators for Electronic Communications, *supra* note 3, p. 3.

[11] Conseil de l' Audiovisual de Catalunya, *supra* note 9, p. 24. See also *Seven in 10 US Internet Users Watch OTT Video*, in http://bit.ly/1ORVChp

Three different and equally necessary actors converge in the operation of an OTT service: OTT service providers, device manufacturers and, finally, Internet service providers (ISP).[12] There is no OTT service without the gadgets and devices (for example, a modem, cell phone or computer) that make an Internet connection possible, or without an Internet service provider to connect to the network. Therefore, the consumer considers these three services complementary.[13]

Broadband networks have enabled the existence of these services. It is reasonable to think that, as these networks increase in scope and quality, companies offering OTT services will expand into new sectors. Specifically, it is reasonable to expect an increase in the number of consumers as download speed improves over time and can offer a better experience using these services. Currently, Latin America does not have the infrastructure and technology necessary to provide quality OTT services as in other regions, so that even today the companies providing traditional services (for example, telephone, television, etc.) do not feel really threatened by OTT[14] services as competition. The case of messaging services, in particular, is different, such as WhatsApp, which has already expanded greatly along with smartphones.

III.Current situation in Argentina

In Argentina, there are no specific laws on the regulation of OTT services. The Argentine Digital Law (Law 27.078),[15] enacted in 2014, regulates Information and Communication Technologies (ICTs), Telecommunications and related resources. This law was voted based on *"(...) technological innovations, the needs of the population, population growth and the expansion of the telecommunications sector"*. In article 2, the law establishes that its provisions *"(...) are intended to guarantee the human right to communications and telecommunications, to recognize Information and Communication Technologies (ICT) as a dominant factor in the technological and productive independence of our Nation, to promote the role of the State as administrator, encouraging the social function of these technologies, as well as competition and creating employment by establishing clear and transparent guidelines that favor the sustainable development of the sector, seeking the accessibil-*

[12] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 7.
[13] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 7.
[14] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 27.
[15] See Ley Argentina Digital [Argentine Digital Law]: http://bit.ly/2fDLTfS

ity and affordability of information and communication technologies for the people." However, it explicitly excludes in its article 1 "*(...) any type of content regulation, whatever its means of transmission*," so it does not stipulate any regulation regarding OTT services.

At the same time, the Law on Audiovisual Communication Services is in force, as amended by Decree No. 267/2015. Likewise, there is currently a debate around the drafting of a new communication services law that will also contain regulation on convergence.

In 2015, the Federal Authority for Audiovisual Communication Services (AFSCA) and the Federal Audiovisual Communication Council, created by Law 26,522, were dissolved through Decree No. 267/2015[16] along with the Federal Authority on Information Technology and Communications (AFTIC) and the Federal Council of Telecommunications and Digitalization Technologies. Similarly, the National Communications Agency (ENACOM) was created, under the umbrella of the Ministry of Communications, as the Authority for the Application of the Argentine Digital and Audiovisual Communication Services law (Law 26,522), and continuing the responsibilities of the aforementioned bodies. ENACOM created, within the scope of the Ministry of Communications, the Committee for Drafting the Bill for the Reform, Update and Unification of Laws 26,522 and 27,078. The purpose of this Committee is to draft the bill, which will then be submitted to Congress for debate and eventual enactment. Regarding its methods, the Committee intends to convene "*(...) representatives of consumers, unions, organizations, journalists and various intellectuals and specialists for the drafting of a plural and modern regulatory framework that takes into account the new Information and Communication Technologies (ICT), the Internet, telecommunications networks and audiovisual media in an integral and unifying (convergent) way.*"

At the time of writing this article, the Committee is still working on drafting a bill of law in this regard. In July 2016, after meetings with different sectors that presented their views and recommendations on the subject, the Committee announced the principles that will guide the drafting of said bill.[17] Principle No. 13 of the published document could be applicable to OTT services, as it states that "*Intermediate applications in the provision of services in the field of Convergent Communications, regardless of the means used, must respect pertinent local regulations, and are liable for any damages that the intermediation activity produces if once notified they do*

[16] See Executive Order No. 267/2015: http://bit.ly/2eLG5mX

[17] See the 17 principles: http://bit.ly/29PiYUA

not cease the harmful action. They must be registered as established by the enforcement authority respecting intellectual property rights." However, until there is a greater degree of detail regarding the specific content of the bill, it is difficult to make an adequate analysis.

IV. Arguments for and against regulation

Conflicts such as the one that arose with the arrival of Uber to Argentina[18] or predicaments about how to collect taxes on services provided on the Internet[19] illustrate the need to promote a discussion about the regulation of OTTs. The government's intention to regulate convergent communication services, including traditional media and services offered on the Internet, increased the urgency of the debate.[20] The discussion, however, does not yet lie in the type of regulation applicable; it also covers the advisability of regulating this type of service, a prior issue that deserves some reflection.

The first point to analyze is the need to regulate OTT services. The sectors that are in favor of regulation argue that:

- OTT services offer the same services as traditional communication companies and, therefore, should be regulated in the same way in order to provide a balance to the regulatory situation.
- OTT services are free-riders of Internet service providers. At the moment they do not share with them the costs of obligations, like paying providers for the use of their networks. Those who argue in this line say that traditional communication companies have invested in infrastructure that now OTT companies use without making any contribution.
- OTT services have a negative economic impact on Internet providers, which hinders investment.[21]

[18] See, for example: http://bit.ly/2ahlAOq and http://bit.ly/2ahlLcJ

[19] The discussion that took place in 2014 in the City of Buenos Aires about the possibility of collecting local taxes from providers of OTT services such as Netflix, Spotify and others is another example of the need to encourage analysis on these issues. See, for example: http://bit.ly/2cwkblR and http://bit.ly/2cwiOnl

[20] At the time of writing this article, the text of the Convergent Communications Bill of Law was not yet known. However, the 17 Principles proposed by the Committee for Drafting the Bill for the Reform of the New Communications Law of the National Communications Authority (ENACOM) that will govern the drafting of the law can be accessed: http://bit.ly/29PiYUA

Comments submitted by CELE to the Principles are available: http://bit.ly/2fPPjBk

[21] Asia Internet Coalition, *Smart Regulation for OTT Growth*, October 2015, p. 5. Retrieved from: http://bit.ly/1mT2e3K

- Finally, it should be noted that "(…) the public interest demands (…) from the State the promotion of public policies that stimulate a diversity of actors, diversity of perspectives and points of view, cultural diversity, geographic diversity, diversity of services and applications. When public policy does not promote diversity, business logic disregards the needs of the most vulnerable sectors (like access and content)."[22] Therefore, the participation of the State by implementing policies that guarantee diversity and safeguard human rights is necessary, including on the Internet.

These arguments generally are supported by the need to level all conditions (the concept of "level playing field"). According to this idea, services that have the same function and compete with each other should be subject to the same regulatory treatment.[23] This concept is highlighted by traditional network operators who are affected by the expansion of companies that offer their services online. They emphasize that the services are the same, beyond the different means used, highlighting the nature of the services and not the technologies.[24]

On the other hand, and advocating against regulating OTTs or regulating them differently from traditional media and telecommunications players, the argument is that while the service may be similar, the technology through which it is provided is radically different and its integrity depends on respect for this characteristic. Some of the most significant arguments in this line are:

- The Joint Declaration on Freedom of Expression and the Internet signed by representatives of different international human rights organizations states that "*Regulatory approaches developed for other means of communication — such as telephone or radio and television — cannot simply be transferred to the Internet, but they must be designed specifically for this medium, taking into account its particularities*."[25] The regulation of the services offered on the

[22] *See* "La convergencia es más que un proyecto de ley" [Convergence is more than a bill] by Martín Becerra. Retrieved from: http://bit.ly/23X2YDF

[23] Body of European Regulators for Electronic Communications, *supra* note 3, p. 22.

[24] See document presented by the Inter-American Association of Telecommunications Companies (ASIET) before the Editorial Committee for the New Communications Law dated May 18, 2016, p. 21. Document retrieved from: http://bit.ly/1sxiXg6

[25] The Special Rapporteur of the United Nations (UN) on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Representative on Freedom of the Media of the Organization for Security and Co-operation in Europe (OSCE), Special Rapporteur on Freedom of Expression of the Organization of American States (OAS), and Special Rapporteur on Freedom of Expression and Access to Information of the African Commission on Human

Internet must consider the different levels of this technology, and respect its open, dynamic and decentralized nature, characteristics that other traditional communication means do not have.[26] In the case of television, the differences are also essential when trying to regulate this service. For example, until a few years ago it was only possible to see on a particular screen specific content and products that were distributed through exclusive technology and identified actors. Nowadays, thanks to OTT services this has changed radically: television is available on multiple devices with multiscreen and multichannel access, and the contents are available in multiple forms.[27] The user can interact with the content and, furthermore, the content offer is no longer the exclusive domain of the traditional broadcasting company. Any content or services producer can offer contents on equal terms.[28]

- OTT companies, unlike traditional operators, do not control "critical infrastructure", have low market entry costs and a lot of competition. Traditional operators, on the other hand, have a very high market entry cost and very little competition. Therefore, the concept of level playing field fails from its conception.[29]
- An OTT service is part of the Internet as another application, so it is not necessary to regulate it, just like with applications in general. It would not make sense to try to impose old regulatory structures on the new broadband networks. According to Vinton Cerf, "The Internet was designed as a neutral platform for the exchange of packages on which everything that could be packaged could be transmitted and received." Thus, the "notion of OTT is simply a misunderstanding of this system's design, since any application is made over this high-speed network of underlying packet exchange."[30]

---

and People's Rights (CADHP)), *Declaración Conjunta sobre Libertad de Expresión e Internet* [Joint Declaration on Freedom of Expression and the Internet], June 1, 2011. Par. 1) c.

[26] See "Internet plural y abierta – Principios fundamentales para la regulación" [Plural and open Internet - Fundamental principles for regulation], in http://bit.ly/1XksnlW

[27] Conseil de l' Audiovisual de Catalunya, *Estrategias y normativas de los servicios OTT en el marco de los EE.UU. (2005-2015)* [Strategies and regulations for OTT services in the US framework. (2005-2015)], in *Audiovisual OTT, nuevas fronteras y desafíos*, Quaderns del CAC No. 42, Vol. XlX, July 2016, p. 16.

[28] *Idem*.

[29] Taking into account the costs that regulation entails, for certain sectors the solution should be eliminating or restructuring the regulation of network operators in order to promote investment, growth, and access to bandwidth. Therefore, a level playing field would be achieved through the deregulation of all suppliers.

[30] Vinton Cerf in the panel "¿Somos todos OTT? Los peligros de regular un concepto

- Regarding the argument that states that OTT companies are free-riders some authors point out that access and applications are complementary of each other. More complex applications create greater demand and willingness to pay for better network access, and better quality access coverage allows greater use of messaging services and other applications. Therefore, the free-rider problem does not exist.[31]
- Finally, "many of the regulations affecting the telecommunications sector made sense when there were monopolies, but with the emergence of new technologies and new providers, there is less need for regulation to promote competition."[32]

Traditional services are offered locally in each country, while Internet services are offered globally and across territorial boundaries.[33] The regulator has to take into account that OTTs are not only subject to local regulations, but also to the laws of all the other countries where they operate, balancing the benefits of technology and global innovation with the need to uphold specific local regulations. In many cases, OTT services are offered by companies that do not have a physical presence in the country that imposes the regulation, in which case important problems arise around the implementation of the law.

The global nature of the Internet and the possibility of causing fragmentations that harm users are also arguments offered for not regulating OTT services.

V. OTTs and freedom of expression

What makes OTT services different from traditional communication services is the fact that they are offered through the Internet. As the United Nations states human rights, including freedom of expression, are fully in

indefinido" [Are we all OTT? The dangers of regulating an indefinite concept] organized by the Asociación Latinoamericana de Internet [Latin American Internet Association] (ALAI) at the 2016 Internet Governance Forum (IGF) in Guadalajara, Mexico. Notes from the session can be accessed here: http://bit.ly/2hx09MX

[31] Williamson, Brian, *Next Generation communications & the level playing field – what should be done?* June 2016, p. 17. Retrieved from: http://bit.ly/2guYerk

[32] Alexander Riobó, Regional Director Regulatory Affairs of Telefónica, in the panel "¿Somos todos OTT? Los peligros de regular un concepto indefinido" organized by the Asociación Latinoamericana de Internet (ALAI) at the 2016 Internet Governance Forum (IGF) in Guadalajara, Mexico. Notes from the session can be accessed here: http://bit.ly/2hx09MX

[33] Asia Internet Coalition, *supra* note 21, p. 7.

force on the Internet.[34] The Internet has great potential to allow people to express themselves, participate in public discussions, as well as to have access to goods and services, which fosters the basic principles of a democratic society. Internet and the services available in it make communications more economical, accessible and fast, and increase the audience for those who express themselves, extending the borders of the city, province, state, region, etc. The Joint Declaration on Freedom of Expression and the Internet, bearing in mind precisely this distinctive factor and considering the crucial importance of the Internet for the expression and participation in public discussions, establishes as a general principle the positive obligation of States to facilitate universal access to the Internet.[35]

Precisely because of its nature and potential, the regulation of this medium — the Internet — should be evaluated from a human rights perspective and, particularly, from the perspective of freedom of expression. Any limitation to freedom of expression must be assessed around the three-part test, i.e.: the limitations must be established clearly by the law; they must be necessary for a democratic society and proportional to the legitimate purposes pursued by the policy. The measures to remove content or block access to certain applications on the Internet must be used exceptionally and adopted in processes that guarantee due process and adequate notification to the user. [36]

In every case, the regulation of OTTS must avoid becoming a direct State interference of the contents offered on the Internet, which could have serious consequences for citizens' right to freedom of expression, both in their individual and social aspect, and regarding the exercise of other fundamental human rights, such as the right of association and assembly, and of political participation, among others.[37]

---

[34] UN, Human Rights Council, *Promoción, protección y disfrute de los derechos humanos en internet*, [The promotion, protection, and enjoyment of human rights on the Internet] A/HRC/20/L.13, June 29, 2012, par. 1. Retrieved from: http://bit.ly/1XksnIW

[35] The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, par. 6) e.

[36] OAS, IACHR, Office of the Special Rapporteur for Freedom of Expression, *Libertad de expresión e Internet* [Freedom of Expression and the Internet], OEA/Ser.L/V/II., CIDH/RELE/INF. 11/13, December 31, 2013, Par. 55.

[37] These issues have recently been brought to attention in Argentina by different civil society organizations as a result of the presentation of a bill to the Buenos Aires City Legislature (bill D-2298) that amended the city's Law of Contravention Procedure. Said bill

The link between the individual and social dimensions of freedom of expression was summarized by the Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights, saying that freedom of expression is "*the means that enable the community, when exercising its options, to be sufficiently informed. Consequently, it can be said that a society that is not well informed is not a society that is truly free."*[38] The Inter-American Court of Human Rights added that "*The Internet has not only made it easier for citizens to express themselves freely and openly, but has also provided ideal conditions for innovation and the exercise of other fundamental rights such as the right to education and free association."*[39] Social media platforms are a clear example of this. People from all over the world have found there a space that they did not previously have to express themselves, inform themselves and connect with other people easily and quickly. Facebook, Twitter, and even WhatsApp, among other social media platforms and instant messaging services, have become essential in the exercise of fundamental rights, in journalism, and as spaces for meeting, association, social protest, activism, and to make public statements. For groups that had never had access to traditional media, social media platforms have come to provide an affordable and easy-to-access alternative that allows them, perhaps for the first time, to express their opinions and ideas, and to disseminate them.

Online media are a special instrument that allows and promotes, in a way never before experienced, the exercise of the right to freedom of expression and other fundamental human rights. Aware of their importance, the Rapporteurs for the United Nations and the Inter-American Commission for Freedom of Expression have argued that the decision to block people's access to websites or filter certain Internet content should be an extreme measure, only justified by international standards (for example, when necessary to protect minors from sexual abuse).[40] Because it refers to the Internet,

---

allowed, as a prevention, a prosecutor or preventive authority "*(…) to order the companies that provide "Internet" service to block, or deny access to the domain or application in question partially or totally depending on the illegal conduct it generates within the City of Buenos Aires or causes consequences within it."*

[38] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85*, La colegiación obligatoria de periodistas (arts. 13 y 29 de la Convención Americana sobre Derechos Humanos)* [Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Arts. 13 and 29 American Convention on Human Rights)], November 13 1985, Par. 70.

[39] OAS, IACHR, Office of the Special Rapporteur for Freedom of Expression, Freedom of *supra* note 36, Par. 2.

[40] The United Nations (UN) Special Rapporteur on the Promotion and Protection of

this restriction also applies to any platform or application that offers OTT services. Blocking a service provided through the Internet must be a last-resort measure and must comply with the aforementioned standards since it could lead to a violation of people's right to freedom of expression and access to information, as well as being an infringement of human rights in general.

VI. Fundamental principles to take into account when drafting regulations

Each OTT service has certain characteristics that distinguish it from the rest: Spotify does not offer the same service as Netflix (Netflix has audiovisual content, while Spotify offers only audio). Similarly, the specific regulation of telephone companies could not be directly used for the service offered by Facebook. Another relevant difference is in the business models that currently operate on the Internet and that require specific distinctions regarding their regulation. The best-known models use micropayments,[41] transactional payment,[42] subscription,[43] membership,[44] freemium/premium services,[45] advertising[46] and

the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, par. 3).

[41] Micropayments / fractional content: a transaction is made to access some type of content (for example, an article on a website, a song or access to the next level in a video game).

[42] Transactional: a model that was born in the field of television ("pay-per-view") and then reached other areas. It is a payment system in which the user pays only for what they see.

[43] Subscription: there is a fixed customer base in a given period (weekly, monthly, and annual) and also a fixed income stream. For example, this is the case of Netflix, Hulu or Spotify.

[44] Membership: it consists of belonging to a group that can also be any type of company that offers services or content. A user can be a member of a fan club or a community of readers, but the subscription would imply a payment for some type of service or access. Being a member means belonging to something (a club, newspaper, insurance, community, etc.). A good example is the case of the video game Second Life.

[45] Freemium / Premium Services: it is a model that offers some product or content for free, while to have access to another part of the content users have to pay (the Premium model). Many times this model includes advertising or marketing added to the contents (or only in Freemium models) so that they support the business together with Premium users. This is the case of Spotify, Skype, iCloud, and Dropbox, among many others.

[46] Ad insertion: it is another version of the Freemium / Premium model. It consists of offering free content but with advertising inserted in the eBooks, compared to the Premium service, which has no advertising and other advantages.

open access.[47] Each business model is different and affects each type of OTT service differently, and these differences matter at the time of regulating. For example, if Internet transactions and not subscriptions were taxed, it would harm a specific type of OTT services and not others that would not have to pay. Despite the differences, which can be approached from various perspectives for their possible regulation, there are fundamental general aspects shared by all companies that offer different OTT services. The State, when deciding to regulate the provision of OTT services, must consider a series of basic principles that arise directly from the characteristics of the Internet, its openness, decentralization, and universality that will guide legislative work in this area.

Taking into account the analysis developed in the previous sections, and based on the bibliography consulted, the opinion of experts who addressed different aspects of the subject and the consensus on the subject, below we present the general principles that we consider fundamental when studying these issues:

6.1 Define precisely the concept of "Over-the-Top service" before regulating it

Currently, the definition of OTT services covers a large number of varied and radically different services in terms of procedures, benefits, and economic models. Using the various current definitions, services such as those offered by Netflix or Facebook could be regulated in the same way, despite their radical differences.

The lack of clarity in the definition of the concept gives the State discretion to decide, even arbitrarily, what rules apply to certain contents or services and not others which are similar or even the same. This power should be as restricted as possible, due to the obvious negative consequences in terms of freedom.

To make-up for the terminology deficiency, some countries have chosen to group certain OTT services according to their function, for example. Thus, in countries such as Hong Kong, Singapore, and the United States, regulators have specifically differentiated the services which are substitutes of traditional telephone from emerging services,[48] drafting specific regula-

[47] Open Access: refers to all types of access without prior subscription or payment. It is usually used to offer educational, scientific or academic material that is directly related to the management of purchases and loans in libraries. New business models in the digital era, LIBER, 2014. Retrieved from: http://bit.ly/1tdtTlv

[48] Asia Internet Coalition, *Smart Regulation for OTT Growth*, October 2015, p. 7. Retrieved from: http://bit.ly/1mT2e3K

tions for them. In the case of Singapore, for example, the regulation does not require all OTT service providers to offer access to emergency services but requires that those who do not provide it to inform the user about it.[49]

6.2 Promoting universal Internet access

The State must ensure that, in any case, the regulation does not have detrimental consequences for citizens' access to the Internet, not only referring to infrastructure access, but also to the set of skills required to fully access the network, including accessibility of content, digital literacy, plurality of languages, affordability, etc.

States must promote universal Internet access in order to guarantee the effective exercise of the right to freedom of expression and other fundamental human rights, such as the right to education, health care and work, the right to freedom of assembly and association, and the right to free elections.[50] In the words of the OAS General Assembly, "*telecommunication and information and communication technologies (ICT) and their applications are essential to political, economic, social, and cultural development and are also an essential factor in poverty reduction, job creation, environmental protection, and natural disaster prevention and mitigation.*[51]"

The regulation of OTT services must consider the fundamental role of the Internet today, both for communications and public participation, and the consequent obligation to promote universal access. The obligations the States imposed on to OTT services should not be a barrier for the universal access to the Internet, but they should lean towards and contribute to fulfilling the objective of universal access. Establishing disproportionate or unnecessary regulatory obligations to OTT services could imply the imposition of barriers that may impact the open nature of the Internet, fragmenting it and making

[49] *Idem*.

[50] The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, Point 6. a).

[51] OAS, General Assembly, *Utilización de las Telecomunicaciones/Tecnologías de la Información y la Comunicación para crear una sociedad de la información integradora*, [Telecommunications/Information Technologies and Communication to Build an inclusive Information Society] AG/RES. 2702 (XLII-O/12), June 4, 2012.

it impossible for users to access valuable content.[52]

In this regard, the IACHR has expressed, for example, that if a State decides to impose registration obligations or other requirements on Internet service providers, they may only be considered legitimate if they pass the test established by international law for restrictions to freedom of expression, summarized in the tripartite test,[53] previously mentioned.

Given the Internet's special characteristics, the regulation of OTT services can influence and impact negatively or positively on innovation, economic growth and the availability of services in the country. For example, traditional regulation, in general, establishes high entry costs and specific local requirements.[54] If this model were followed, discouraged by the Offices of the Special Rapporteur on Freedom of Expression of the UN and the IACHR, each country could establish its own rules to protect its operators or users, putting up barriers to innovative products.[55] Hence, the objective of universal access and the nature of the Internet should be addressed when discussing regulatory issues regarding OTTs.

6.3 Respecting the principle of net neutrality

Net neutrality is "*the principle according to which Internet traffic shall be treated equally, without discrimination, restriction or interference regardless of its sender, recipient, type or content, so that Internet users' freedom of choice is not restricted by favoring or disfavoring the transmission of Internet traffic associated with particular content, services, applications, or devices."*[56] Even when there is no univocal definition of the concept of net neutrality, there is at least consensus on a common issue: net neutrality includes non-discrimination of content, guarantees regarding not-blocking and freedom of use of devices.[57]

---

[52] Asia Internet Coalition, *supra* note 21, p. 4.

[53] The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, Point 6. d).

[54] Asia Internet Coalition, *supra* note 21, p. 4.

[55] *Idem*.

[56] Dynamic Coalition on Network Neutrality, "Model Framework on Network Neutrality", retrieved in English from: http://bit.ly/2fY6Oi0, and in Spanish from http://bit.ly/1QFWbaC

[57] Cortés Castillo, Carlos, *La neutralidad de la red: la tensión entre la no discriminación*

Contents must be treated in the same way by the infrastructure providers, regardless of the origin or type of content in question.[58] Furthermore, consumers should be able to access services and content of their choice for free via public Internet.[59] Therefore, "*Traffic over the Internet should not be discriminated against, restricted, blocked or interfered with unless strictly necessary and proportional in order to preserve the integrity and security of the network.*"[60] The goal is to have an open, fast and fair Internet.

The principle of net neutrality "*shall apply to all Internet access services and Internet transit services offered by ISPs, regardless of the underlying technology used to transmit signals*."[61] This principle constitutes a guarantee so that the companies that enter a new market in the Internet can compete fairly with big companies that are already in operation.[62] This type of measure could cut down innovation in certain sectors because there are no incentives to create new products or services without adequate means to guarantee access to them online. If net neutrality is not defined and established as a principle in the legislation, fixed and mobile Internet operators, who provide the users with the Internet connection through which online services are offered, could have the incentives and the possibility to regulate or block online services that constitute a commercial threat to their own services, such as messaging or VoIP.[63]

The Joint Declaration on freedom of expression on the Internet, alerted to these risks, provides that "*Internet intermediaries should be required to be transparent about any traffic or information management practices they employ, and relevant information on such practices should be made available in a form that is accessible to all stakeholders*."[64] Failure to comply

y la gestión [Net neutrality: the tension between non-discrimination and management], in "Internet y Derechos Humanos: aportes para la discusión en América Latina" [Internet and Human Rights: contributions for discussion in Latin America], Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE), Buenos Aires, February 2014, p. 30 and 31.

58 Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 10.

59 Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 74.

60 OAS, IACHR, Office of the Special Rapporteur for Freedom of Expression, *supra* note 36, Par. 30.

61 Dynamic Coalition on Network Neutrality, *supra* note 56.

62 Mahanagar Doorsanchar Bhawan, Jawahar Lal Nehru Marg; Telecommunications Regulatory Authority of India, *Consultation Paper on Regulatory Framework for Over-the-Top (OTT) services*, Consultation document No. 2/2015, March 27, 2015, New Delhi, p. 78.

63 Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 75.

64 The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur

with the principle of net neutrality could give ISPs a high level of control and discretion to discriminate certain types of content and opinions, whether for economic, political or other reasons. This discretion could prejudice the entire society, depriving it of opinions and ideas, and users of their individuality, while the dissemination of expressions contrary to general, mass or majority opinions may be impeded.

Some Internet providers have expressed their intention to create an additional price scheme for Internet access and differential rates to give priority to certain content[65] (usually, because content producers take advantage of the infrastructure without paying for it and their services threaten those traditionally provided by them[66]), but this type of scheme could harm society, [67] precisely because it would enable the prioritization or degradation of certain services in clear contrast to the principle of neutrality.[68]

Countries such as the United States, Brazil and Europe in general have enacted laws that safeguard net neutrality.[69] In the United States, for example, the Federal Communications Commission (FCC) launched new Internet rules in March 2015 based on Title II of the Communications Act of 1934 and Section 706 of the Communications Act of 1996. These rules strengthen the concept of net neutrality in those countries, and they apply to both fixed and mobile Internet services.[70] These rules are:

- No blocking: service providers may not block access to legal content, applications, services or non-harmful devices.
- No regulation: service providers may not affect or degrade legal Internet traffic based on non-harmful content, applications, services or devices.
- No paid prioritization: service providers may not favor any legal Internet traffic over another in exchange for any type of consideration

on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, Point 5. d).

[65] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, p. 11.

[66] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, pp. 10 and 11.

[67] See EU Regulation 2015/2120 recently passed by the European Union on net neutrality: http://bit.ly/2ckebgv

[68] UN, *La nueva regulación digital: de la Internet del consumo a la Internet de la producción* [The new digital revolution: from the consumer Internet to the industrial Internet], eLAC 2018, United Nations Economic Commission for Latin America and the Caribbean (ECLAC), July 2015, p. 85.

[69] Ganuza, Juan José and Viecens, María Fernanda, *supra* note 8, pp. 75 and 76.

[70] Mahanagar Doorsanchar Bhawan, Jawahar Lal Nehru Marg, *supra* note 62, p. 64.

— namely, there should be no "fast lanes". This rule also prohibits ISPs from prioritizing content and services from their affiliates.[71]

In Argentina, net neutrality is enshrined in the Argentina Digital Law of 2014 and is vaguely mentioned in the Principles published by the Drafting Committee for the new convergent communications law. Net neutrality must be expressly established and guaranteed to protect and promote freedom of expression on the Internet.

6.4 Respecting privacy and promoting transparency

The UN has foreseen that "*(...) any capture of communications data is potentially an interference with privacy and, further, that the collection and retention of communications data amounts to an interference with privacy whether or not those data are subsequently consulted or used.*[72]" Any regulation of OTT services must respect people's privacy and promote the transparency of State actions.

Based on the different reasons to regulate OTT services, including reasons for internal security and preventing and investigating crime, it is worth highlighting the importance of guaranteeing the right to privacy both online and offline. Intercepting Internet communications should only be allowed with a court order from a competent authority, in compliance with the guarantees of due process. In this regard, and as established by the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, *"(...) The United Nations High Commissioner for Human Rights has also urged information and communications companies to disclose risks and government demands transparently. (See A/HRC/27/37) (...) Corporate transparency obligations may also include a duty to disclose processes and reporting relating to terms of service enforcement and private requests for content regulation and user data."*[73] In other words, according to the corporate obligations in the field of human rights, those that provide

[71] Mahanagar Doorsanchar Bhawan, Jawahar Lal Nehru Marg, *supra* note 62, p. 64.

[72] UN, General Assembly, *El derecho a la privacidad en la era digital*, [The right to privacy in the digital age] Report of the Office of the United Nations High Commissioner for Human Rights A/HRC/27/37, June 30, 2014, Par. 20.

[73] UN, General Assembly, *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión* [Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression], A/HRC/32/38, May 11, 2016, Par. 12.

OTT services should publish statistics on interception requests, if any, in order to achieve greater transparency.

Additionally, when regulating these services, it is appropriate to legally establish an obligation of transparency on the part of the State and the OTT services companies regarding the processing of personal data, so that they are not used for purposes other than those that users consented.

6.5 Promoting models that benefit freedom of expression and do not restrict it

It is important to keep in mind that every OTT service is offered on the Internet, which has special characteristics and is different from other traditional media. The Office of the Special Rapporteur for Freedom of Expression of the IACHR argued that limitations on freedom of expression on the Internet must consider the consequences that such limitation entails for cyberspace as a whole.[74] Regulatory models that enable blocking certain domains and applications, for example, could be detrimental to the exercise of the right to freedom of expression. The Internet must be an open space that promotes the dissemination of information and ideas, and the regulatory solution must take these particular characteristics into account.

As mentioned above, the Joint Declaration on Freedom of Expression and the Internet establishes that, given the plural and open nature of the Internet, a measure that determines the mandatory blocking of websites, IP addresses, ports, network protocols or certain types of uses must be only used in extreme situations.[75] Therefore, the regulation of OTT services may

[74] OAS, IACHR, Office of the Special Rapporteur for Freedom of Expression, *supra* note 36, Par. 53. The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, par. 3). a.

[75] Here it is important to mention again the recent case of the bill presented in the City of Buenos Aires (bill D-2298) that granted various local authorities the power to block domain names and applications. In the case in question, civil society organizations argued that "*(…) the Presti bill puts this tool [the Internet] at risk by allowing administrative authorities — including the police force — to censor content to prevent local contraventions. It also grants this power to Court officers and prosecutors, creating great opportunities for Internet fragmentation*." Likewise, among other points, the letter presented by civil society organizations argues that "*(…) the Human Rights pacts that protect Freedom of Expression have constitutional status and therefore, the content restriction must conform*

consider blocking only for extremely exceptional cases, since it implies the denying of an essential means for the exercise of the right to freedom of expression and other fundamental rights.

It is preferable that when choosing communication regulation models the priority is focused on the user of the services. The less expensive the service, the greater the access for consumers and thus they can exercise their right to freedom of expression, fully applicable to communications, ideas, and information that are disseminated and accessed through the Internet.[76] In general, OTT services have few entry barriers and lower costs, so the prices that consumers must pay are usually lower than those offered by traditional communication services.

6.6 Promoting investment and technological progress

Even though the emergence of OTT services could have negative consequences in some sectors, "*throughout history, every technological revolution has had "winners" and "losers", and what must finally be taken into account is the effect on the welfare of society as a whole. Therefore, governments should facilitate this process and not implement measures that could hinder it (...) A good economic and regulatory policy should undoubtedly ensure and strengthen these possibilities in society.*"[77]

Therefore, a regulation with this perspective must take into account, essentially, the impact that the policy may have on the country's investment and the development of technological advances. For example, realizing that measures such as the need to register companies that provide OTT services could con-

---

to these standards. As we explained in our Letter, this bill does not comply with any of them, and it is especially dangerous in relation to the investigation of the activities that all users carry out connected to the network." See: http://bit.ly/2bSdRHU

See "Internet plural y abierta – Principios fundamentales para la regulación", in http://bit.ly/1XksnlW

[76] OAS, IACHR, Office of the Special Rapporteur for Freedom of Expression, supra note 36, Par. 2. The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, a Representative on Freedom of the Media for the Organization for Security and Co-operation in Europe (OSCE), the Special Rapporteur on Freedom of Expression of the Organization of American States (OAS) and the Special Rapporteur for Freedom of Expression and Access to Information of the African Commission on Human and People's Rights (CADHP), *supra* note 25, par. 1). a.

[77] Ganuza, Juan José and Viecens, María Fernanda, *El desafío de los contenidos y servicios over-the-top*, en Banda Ancha en América Latina: más allá de la conectividad [The challenge of over-the-top content and services in Broadband in Latin America: beyond connectivity], CEPAL, February 2013, p. 349. Retrieved from: http://bit.ly/2fCFiUH

stitute a hurdle in the country's technological development and, subsequently, harm consumers, something that is not recommended as a regulation.

VII. Conclusions

This paper has mainly focused on systematizing and developing the principles that should be taken into account when drafting regulations on the subject, particularly in light of national and international human rights laws. The proposed principles have a general nature, notwithstanding the current local priorities and particularities of the country and those being considered for the future at the time of regulating.

As highlighted throughout this article, there are currently no clear regulations regarding OTT services. In fact, at the time, there is no exact definition of the term either from the technical or from the normative perspective. In this context, it is essential to prioritize the definition of the object to be regulated, taking into account that OTT services are offered on the Internet, a means and platform for the exercise of human rights, with unique and particular characteristics, such as openness, decentralization, and pluralism, among others. The regulation of OTT services must take into account the nature of the Internet, its importance for the development and full exercise of human rights, and its technical peculiarities, especially when assessing the proportionality of the proposed regulatory measures regarding the legitimate objectives that such measures can pursue.

Annexed I

Content Moderation and private censorship: standards drawn from the jurisprudence of the Inter-American Human Rights system*

I. Introduction

The Center for Studies on Freedom of Expression and Access to Information (CELE) is a research Center hosted at Universidad de Palermo law school in Buenos Aires, Argentina. The Center devotes its work to promoting and enhancing the protection of freedom of speech and expression through cutting-edge research capable of shaping and changing public debate on key policy issues, and capacity building. The Center's work is regional in scope and has a special interest in Inter-American law and standards that it seeks to promote and enhance region-wide.

This submission seeks to bring some of the standards that could be drawn from the Inter-American Human Rights System to the questions posed by the Rapporteur in his call for submissions on Private content regulation in the digital age.

In the words of the Rapporteur, "Private companies facilitate an unprecedented global sharing of information and ideas. Social and search platforms in particular have become primary sources of news and information (and disinformation) for hundreds of millions of people. With that role they have also become gatekeepers of expression that may excite passions and knowledge – or incite hatred, discrimination, violence, harassment, and abuse." It specifically asks "What steps should platforms, government actors, and others take to ensure that these processes establish adequate safeguards for

* Submission to the United Nations Special Rapporteur on the Protection and Promotion of Freedom of Opinion and Expression, David Kaye, by the Center for Studies on Freedom of Expression and Access to Information (CELE). December, 2017. This submission was written by Agustina Del Campo (adelca9@palermo.edu), Director of CELE.

freedom of expression?" This is the question we seek to address and try to establish whether the Inter-American system for the protection of human rights offers any guidance, either for companies and/or for States, to address content regulation in the digital age, and if so, what those standards look like. This submission concludes that the Inter-American System does provide some standards that could serve as a baseline for private actors, and offers concrete recommendations for States and private companies to further enhance the protection of freedom of expression in the digital age.

II. Expression, dissemination and censorship

The protection of Freedom of expression is widely recognized throughout the different international human rights instruments across regions as well as the Universal Declaration and the international Covenant on Civil and Political Rights.

There is common understanding among human rights and free speech organisms, advocates and experts that expression and dissemination go hand in hand and one cannot exist without the other. As stated by the Inter-American Court on Human Rights 30 years ago, "[f]reedom of expression goes further than the theoretical recognition of the right to speak or to write. It also includes and cannot be separated from the right to use whatever medium is deemed appropriate to impart ideas and to have them reach as wide an audience as possible. (…) [r]estrictions that are imposed on dissemination represent, in equal measure, a direct limitation on the right to express oneself freely."[1]

This very same notion that dissemination and expression are indivisible and contemplate any medium underlies several joint declarations by the Special Rapporteurs on Freedom of Expression at the Organization of American States, United Nations, Organization for Security and Cooperation in Europe, and the African Commission on Human and Peoples' Rights,[2] particularly when they stated that the right to freedom of expression

[1] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85, November 13, 1985, Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Articles 13 and 29 American Convention on Human Rights), par. 31.

[2] See, for example, UN, OSCE, OAS, and ACHPR, "Joint Declaration on Freedom of Expression and the Internet," June 1, 2011. Available at: http://bit.ly/1wnld8U; and "Joint Declaration on Freedom of Expression and Responses to Conflict Situations", May 4, 2015. Available at: http://bit.ly/2yYwRhE

applies fully to the internet, and online limitations are only acceptable if they comply with international standards.

While there is wide agreement as to the indissoluble nature of expression and dissemination, there is also wide agreement that free speech is not an absolute right and that certain expressions constitute abuses. Still, there is no universal agreement as to what "abuses" mean, or even as to the means to address such abuses. What may be deemed abusive in one country, may not be so in another. And what could be understood as a legal means to deal with such abuse in one region, may not be in another.

Private companies mediating content and expression across borders have an enormous (and ever increasing) power to affect public discourse, impact free speech and access to information. They are also undergoing increasing State and civil society pressure to exercise their powers to mediate content, while establishing terms of services (ToS) that prohibit certain types of content deemed abusive. Some of those abusive contents may be illegal. Some may be short of illegal but problematic. Some others may be illegal in some countries but not in others. An example of each may help illustrate: a) child pornography is illegal everywhere; b) aggressive or disrespectful language short of harassment, although unwanted, not illegal; c) blasphemy may be an offense punishable by law in some countries, and not be considered abusive at all in another.

On the means to address "abusive" content, legally tolerated means also vary from one region to another. Prior censorship, for example, understood as a preventative measure to impede the dissemination of information or ideas, may be understood as a legitimate means to redress abusive content in some regions (e.g. Europe) and not in others (e.g. the Americas).

While governments and laws vary from one State to another, some companies, particularly big social media companies, are a single unit and need common norms to function across borders through the different legal regimes. In doing so, however, they must not only bear in mind the substantive differences between the legal systems but also the different approaches that they can legitimately take or enforce globally as potential means for redress.

III. Article 13 of the American Convention: How does it differ from other frameworks for the protection of freedom of expression worldwide?

Article 13 of the American Convention on Human Rights (ACHR), although apparently similar to the ICCPR's Article 19, sets standards that differ slightly from it and bring about certain specificities that could be particularly relevant to the study that is being conducted. As established in Advisory Opinion 5/85, "The form in which the American Convention is drafted differs very significantly from Article 10 of the European Convention, which is formulated in very general terms. (…) The Covenant, in turn, is more restrictive than the American Convention, if only because it does not expressly prohibit prior censorship."[3]

Article 13 in its relevant parts states that:

1. *Everyone has the right to freedom of thought and expression. This right includes freedom of seek, receive, and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing, in print, in the form of art, or through any other medium of ones' choice.*
2. *The exercise of the right provided for in the foregoing paragraph shall not be subject to prior censorship but shall be subject to subsequent imposition of liability, which shall be expressly established by law to the extent necessary to ensure:*
   a. *respect for the rights or reputation of others; and*
   b. *the protection of national security, public order, or public health or morals.*
3. *The right of expression may not be restricted by indirect methods or means, such as abuse of government or private controls over newsprint, radio broadcasting frequencies, or equipment used in the dissemination of information, or by any other means tending to impede the communication and circulation of ideas and opinions.*

The text of the ACHR expressly prohibits prior censorship and addresses not only government indirect restrictions on freedom of expression, but also private restrictions of these rights where those restrictions could lead to similar results as government controls. As expressed by the Inter-American

[3] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85, November 13, 1985, Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Articles 13 and 29 American Convention on Human Rights), par. 45.

Court, "Neither the European Convention nor the Covenant contains a comparable clause" to Article 13 (3). Guarantees in the ACHR "(...) were designed to be more generous and to reduce to a bare minimum restrictions impeding the free circulation of ideas."[4]

The prohibition of prior censorship is wide and broad. The only exception is that established in article 13 (4), which allows for prior censorship of public entertainments with the sole purpose of regulating children's access to them.[5] The Court had multiple opportunities to address prior censorship through its jurisprudence and it consistently confirmed the understanding that censorship, whether prohibiting expression or its dissemination, constituted an unacceptable state measure.[6]

There are different views as to what constitutes prior censorship and how it applies *vis a vis* States/non-state actors.[7] The Inter-American Rapporteur for Freedom of Expression has asserted that government filtering and blocking prior to judicial review of its legality constitutes prior censorship. And following this line, Professor Nunziato argues that this will be so regardless of whether the content be removed before it is made publicly available or after being published but before judicial determination of illegality, and cites to the U.S. Supreme Court and the ACHR in reaching this conclusion.[8]

The other main difference between the ACHR and other instruments is an express prohibition of indirect restrictions including those generated by abusive government or private controls (Article 13(3)). These indirect restrictions may adopt different and multiple forms and Article 13(3) is open ended as to the examples it cites to. The Inter-American system has applied this clause to different cases as free speech restrictions in the region moved

[4] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85, November 13, 1985, Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Articles 13 and 29 American Convention on Human Rights), par. 50.

[5] ACHR, Art. 13(4):4. Notwithstanding the provisions of paragraph 2 above, public entertainments may be subject by law to prior censorship for the sole purpose of regulating access to them for the moral protection of childhood and adolescence.

[6] IACtHR, case of The Last Temptation of Christ (Chile), February 5, 2001, Merits and Reparations, Series C N. 73, available at http://www.corteidh.or.cr/docs/casos/articulos/Seriec_73_esp.pdf; And Palamara Iribarne vs. Chile, Nov. 22, 2005, Preliminary Objections, Merit & Reparations, Series C, N. 135, available at: http://www.corteidh.or.cr/docs/casos/articulos/seriec_135_esp.pdf.

[7] Nunziato, Dawn C., Preservar la libertad de expresion en America Latina, CELE, Towards an internet free of censorship, 2012. pag. 29-30.

[8] Nunziato, Dawn C., Preservar la libertad de expresion en America Latina, CELE, Towards an internet free of censorship, 2012. pag. 33 citing Bantam Books vs. Sullivan, 372 US 58 (1963).

from direct and manifest towards more subtle and indirect in nature (eg. license renewals, nationality processes, State publicity assignments, etc.). Under the same logic, the Special Rapporteurs' Office with the Inter-American Commission on Human Rights has repeatedly asserted that establishing intermediary liability for third party posted content could infringe upon this particular norm, and constitute an indirect restriction per the ACHR.[9]

Additionally, and following the above conceptualization of prior censorship, private content removals per terms of service could also be considered prior censorship under the American Convention. Per art. 13(3) this kind of private action could have a similar effect on the circulation of speech. As Bertoni puts it, States could be held internationally liable for leaving with private entities the ability to censor content, since those private entities would in fact be infringing upon the freedom of expression of their users.[10]

IV. Duty to Respect and Ensure in the ACHR, the ICCPR and the European Convention

When evaluating freedom of expression cases, the usual focus is on the obligations of States to respect free speech, understanding this as a negative obligation, to refrain or to not interfere illegitimately with freedom of expression, whether directly or indirectly. However, there are also positive obligations upon the States to guarantee the full exercise of this right for the people under their jurisdiction that are generally established both in the European Convention and the ICCPR as well as in the ACHR.

General Comment 34 of the Human Rights Committee sets from the outset in paragraph 7 that "The obligation [to respect and ensure] also requires States parties to ensure that persons are protected from any acts by private persons or entities that would impair the enjoyment of the freedoms of opinion and expression to the extent that these Covenant rights are amenable to application between private persons or entities." [11]

[9] IACHR, Freedom of Expression on the internet, RELE, 2013, available at: http://www.oas.org/en/iachr/expression/docs/reports/internet/foe_and_internet_report_2013.pdf

[10] E. Bertoni, OC 5-85: Su vigencia en la era digital, en Libertad de Expresión a 30 años de la Opinión Consultiva sobre la Colegiación Obligatoria de Periodistas, CIDH RELE, 2017. He exemplifies with two cases: Costeja, where the private censorship is being promoted by the State itself; and Delfi, where the private censorship draws from intermediary liability schemes.

[11] HRC, General comment No. 34: Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, available at http://www2.ohchr.org/english/bodies/

The European Convention has similar language although not as specific in article 1.[12] The European Court has said that "Genuine, effective exercise of this freedom does not depend merely on the State's duty not to interfere, but may require positive measures of protection, even in the sphere of relations between individuals […]".[13] Although the specifics of the positive obligations may be undefined, and the ECtHR endorses the principle of a wide margin of appreciation, authors sustain that under the positive obligation doctrine developed by the Court, States should, to "comply fully" with Article 10, ECHR, "ensure that they do not place intermediaries under such fear of liability claims that they come to impose on themselves filtering that is appropriate for making them immune to any subsequent accusation but is of a kind that threatens the freedom of expression of Internet users". [14]

Under Article 1 of the ACHR, States have a duty to guarantee or "ensure" the full exercise of rights, understood as a duty to adapt their entire structure so that the people under their jurisdiction may fully enjoy and peacefully exercise their human rights.[15] Certain specific positive obligations have been jurisprudentially developed within the right to freedom of expression. Standards towards guaranteeing pluralism and diversity in media can be set as examples of these positive obligations.

The Inter-American Court's Advisory Opinion 5/85 states that:

"48. Article 13(3) does not only deal with indirect governmental restrictions, it also expressly prohibits "private controls" producing the same result. This provision must be read together with the language of Article 1 of the Convention wherein the States Parties "undertake to respect the rights and

hrc/docs/gc34.pdf. This obligation drives from HRC General Comment No. 31 – "The Nature of the General Legal Obligation Imposed on States Parties to the Covenant".

[12] ECHR, Article 1.

[13] Özgür Gündem v. Turkey, no. 23144/93, ECHR 2000-III, para. 43. via Study of fundamental rights limitations for online enforcement through self- regulation conducted by the Institute for Information Law (IViR) Faculty of Law University of Amsterdam, pag. 35.

[14] Study of fundamental rights limitations for online enforcement through self- regulation conducted by the Institute for Information Law (IViR) Faculty of Law University of Amsterdam, pag. 38 citing E. Montero and Q. Van Enis, "Enabling freedom of expression in light of filtering measures imposed on Internet intermediaries: Squaring the circle", Computer Law & Security Review 27 (2011) 21-35, at 34.

[15] IACtHR, Case Velásquez Rodríguez vs. Honduras, Series C, N. 4. 1989. par. 166 available at http://www.corteidh.or.cr/docs/casos/articulos/seriec_04_ing.pdf. "The second obligation of the States Parties is to "ensure" the free and full exercise of the rights recognized by the Convention to every person subject to its jurisdiction. This obligation implies the duty of States Parties to organize the governmental apparatus and, in general, all the structures through which public power is exercised, so that they are capable of juridically ensuring the free and full enjoyment of human rights."

freedoms recognized (in the Convention)... and to ensure to all persons subject to their jurisdiction the free and full exercise of those rights and freedoms....” Hence, a violation of the Convention in this area can be the product not only of State imposed restrictions that impede “the communication and circulation of ideas and opinions,” but also from private controls. States have an obligation to ensure that the violation does not result from the “private controls” referred to in clause 3 of Article 13. [16]

The obligation to ensure implies a duty to act when States have knowledge of a human rights violation and a duty to take appropriate measures to prevent such violations from happening. In this sense, laws that condone human rights violations conducted by private actors are incompatible with the American Convention.

Unfortunately, there is still no jurisprudence on this issue in cases of internet and freedom of expression within the Inter-American system. However, the principles and standards that do exist suggest that current practices among internet companies could compromise the international responsibility of the State. As Bertoni pointed out, leaving private entities to censor may amount to a violation of the Inter-American system’s standards. We would contend that allowing intermediaries to establish, interpret and enforce ToS in an arbitrary, obscure or ambiguous way could also amount to a violation of States’ duties to guarantee the right to freedom of expression, including preventing through reasonable means any violation of this right.

The Inter-American Court framed the issue clearly regarding media: “If freedom of expression requires, in principle, that the communication media are potentially open to all without discrimination or, more precisely, that there be no individuals or groups that are excluded from access to such media, it must be recognized also that such media should, in practice, be true instruments of that freedom and not vehicles for its restriction (…)” [17] and therefore “the conditions for its use must conform to the requirements of this freedom”.

It is clear from the wording of the Court that the objective is to protect the main means for the exercise of free speech, which back in 1985 was mass media. Still, following the logic of the Court, in 2018, mass media is certainly one vehicle, but the internet has risen to be equally and even

[16] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85, November 13, 1985, Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Articles 13 and 29 American Convention on Human Rights), par. 48.

[17] OAS, Inter-American Court of Human Rights, Advisory Opinion OC-5/85, November 13, 1985, Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Articles 13 and 29 American Convention on Human Rights), par. 33.

more powerful a means or vehicle for regular people as well as journalists to exercise this important right. Following the Court's standard, conditions for the internet's use must also conform to the requirements of this freedom.

V. Conclusions and Recommendations

1. Standards are similar yet not equal across different regions and vis a vis universal ones

Although freedom of expression is recognized universally as a human right, linked directly to democratic governance and values, the definition and scope of the right vary if so slightly from one international instrument to another. Such differences generate different obligations for States across regions within the different frameworks.

As other instruments do, the American Convention states a duty among States to respect and guarantee the rights and duties contained in the ACHR. And it also states that when faced with different standards or obligations based on different instruments, States should enforce those most protective of the right in question. In the case of States within the Americas, the ACHR established the most protective standard on the right to freedom of expression and States should abide by those.

Therefore,

1. Companies must respect human rights in all the different jurisdictions where they operate and States must take all reasonable measures to guarantee compliance;
2. In determining global ToS, and verifying their compliance with human rights standards, companies should test against the most protective standards rather than the least protective ones and adjust regionally where the case rises. And
3. Companies should get better acquainted with universal and regional human rights standards and how they interact and dialogue amongst each other.

2. Duties to respect free speech imply obligations not to run illegitimate interference, whether directly or indirectly

States must not infringe upon free speech rights of people under their jurisdiction, neither directly, through government censorship, nor indirectly, through other regulations, including tax, nationality, monetary incentives like state publicity, nor through impositions of undue regimes for internet intermediary liability.

States must refrain from imposing direct or indirect restrictions on freedom of expression online and offline and must refrain from pressuring, suggesting, indirectly imposing restrictions and particularly censorship obligations upon internet intermediaries that would be otherwise incompatible with their human rights obligations.

3. Under Inter-American Standards, States could be liable for a company's abusive Terms of Service if they infringe illegitimately on freedom of expression

Under the standards set forth in the ACHR as explained above, States have a duty not only to protect free speech from government abuse or controls, but also from private abuse and controls where they don't conform to the American Convention for allowable restrictions.

States must also guarantee freedom of expression for the people under their jurisdiction through positive measures including enacting legislation that protects freedom of expression and clearly establishes any allowable restriction, following the criteria set by regional and universal instruments. Finally, States must guarantee that private actors including internet companies do not infringe arbitrarily upon the freedom of expression of their users, including through moderation, filtering, blocking, suspending or canceling measures.

This conclusion has two main ramifications:

a. States must refrain from pressuring, proposing or regulating terms of service that would otherwise be deemed abusive or incompatible with freedom of expression standards within their own countries and under regional or global standards.
b. States must regulate so that companies functioning within their jurisdictions abide by freedom of expression standards and don't abusively control the circulation of opinions and ideas nor do they

exclude specific groups or ideas from the debate while of course respecting the ability of companies to conduct business. This entails:

a. States must ensure that ToS are clear and transparent;
b. States must ensure that the application and enforcement of ToS are transparent and respectful of human rights, including freedom of expression, non-discrimination and due process.

c. The above mentioned should not be construed as a right to access a certain forum or determine specific terms of service for internet companies. While no one has a right to access gmail, Facebook or Twitter without agreeing to their terms of Service and complying with them, States cannot grant internet companies a right to arbitrarily and obscurely apply such ToS to the detriment of basic human rights such as those listed above.

4. Companies need to abide by international human rights standards and be transparent about their ToS, rules and processes.

Whether companies establish local, global or mixed ToS within different jurisdictions, they must make sure that ToSs respect human rights, not only in their theoretical notion but in their application, execution and enforcement.

a. Companies must abide by human rights standards in every jurisdiction where they operate.
b. Companies must clearly establish and publish their ToS unambiguously.
c. Interpretations and enforcement of ToS need to be transparent and subject to user's control and understanding. Arbitrary implementation or enforcement of ToS could amount to freedom of expression violations.

Annexed II

Considering Facebook Oversight Board: turning on expectations*

I. Introduction

The International Criminal Court is popularly referred to as the "world Court". With Facebook's announcement to create an appeals, independent body to review the application and enforcement of its terms of service over its 2.7 billion users all around the globe, the term "World Court" probably gained at least one more, new-found meaning.

Since the creation of this body launched there have been advocates for and against it. The common lines among both sides is that 1) this is a self-regulation initiative and cannot replace a judicial stance of control; 2) as it stands it only addresses some of the issues that users have with content moderation; 3) there are more questions than answers regarding nature, scope, design, expectations, etc.; 4) the creation of this Board, although an interesting proposal, should not divert the company's attention from due process for all content restrictions and curation, and for transparency in rule and decision-making as well as implementation of standards, whether legally imposed or self regulated.

II. Background and implications to the Oversight Board

After a couple of years of scandal after scandal and a growing set of rules, Facebook is seeking legitimacy in governing what the US Supreme Court

* This article was written in May 2019 by Agustina Del Campo, CELE's Director, with the comments of Franco Serra and Paula Roko, researchers at CELE. It was presented as a contribution to Facebook's open global consultation in the context of the creation of the Oversight Board.

has ruled to be "the new public square". The rules (http://www.facebook.com/communitystandards) have gained in complexity and ambiguity over the years and with every addendum (internal guidelines are being changed every week or so, and an overall 2000 changes are included per year approx. in total, according to information provided by FB). Content "flagged" by users to be against those rules raises to a meaningful daily amount.

Per user and government pressures of all sorts, the company has moved content moderation from a responsive to a proactive mode, having automated most of its detection practices (yet unclear how much of its removal practices) for unwanted content and facing extremely complex debates over ethics, corporate social responsibility, liability, damages, and free speech, among others.

The decision to create an independent board or oversight board for Facebook, with an open process and consultations is maybe the most dramatic decision that we have seen from FB in the last few years, at least regarding its content policy. And a first step among internet companies to bring in external actors to their decision making processes. After having shifted the rhetoric from a free-speech oriented discourse to a safety discourse over the last few years and having been on a defensive and ever expanding curating role for increasingly complex typologies of content, Zuckerberg's Blue Print for Content Governance and Enforcement (Nov. 2018) seems to mark a new departing point for the company's approach to content regulation.

For one, Zuckerberg's note appeals directly to the need for improved legitimacy over governance and decision-making. Second, there is a commitment to create an external independent review board, whose decisions are binding and public; every other detail, still unknown. Third, it expressly invites States to define what they expect from a content moderation regime and details at least two concrete initiatives that FB has already engaged with in Europe: 1) the agreement signed with President Macron to work on a new content regulation; 2) their workings towards a new European framework for content moderation and regulation within the next two years. Although FB had anticipated their change in view *vis a vis* legislation during hearings before the US Congress in 2018, this is the first communication that FB sends to its own community officially welcoming and even compelling States to regulate. As Zuckerberg states: "At the end of the day, services must respect local content laws, and I think everyone would benefit from greater clarity on how local governments expect content moderation to work in their countries."

The recognition that "a full system requires addressing both governance and enforcement" is without a doubt a positive step forward in content moderation debates, particularly those related to transparency. Civil society

and academics around the globe had voiced their concerns as to the lack of transparency and the levels of discretion that internet companies enjoyed and have been working with them through the years to improve transparency first over the rules and later onto processes.[1] With the new approach, FB may be able to take procedural and enforcement transparency further while making decisions binding and public.

But maybe what is even more interesting is the policy change *vis a vis* content governance on the platform and what seems to be a new found willingness to "share" the responsibility (or blame) over rule-making. "As I've thought about these content issues, I've increasingly come to believe that Facebook should not make so many important decisions about free expression and safety on our own" wrote Zuckerberg in November. Human Rights activists have been saying so for years. What changed? And how much did it change?

The creation of a board to serve as an audit to this platform's decisions on content moderation could strengthen the exercise of free speech online as much as it could hurt it. If the Board is understood as an internal process, intended to serve the company in dealing with complex issues of free speech and its balance with other rights, unify criteria, and help the system converse better with international human rights standards, the results could be more positive for the entire ecosystem. If understood as a replacement system for already weak due process mechanisms within the platform's decision making structure, and as a body that would create and interpret privately legislated law alone (contract law), it will most probably damage the ecosystem as well as the company. It would also fail the purpose for which it was created. If legitimacy is what FB wants and needs, it can only be built from the dialogue between the private self-regulation norms and existing international human rights standards.

Whether this Board can accomplish legitimacy or not in many ways depends on how the body is structured, the goals that are set for it, the requirements that candidates should meet to be a part of this body, the nature of its decisions (if it is an appeals body in fact), and the transparency and publicity of its decisions and reasoning.

[1] RDR, GNI, to name a few.

III. Legitimacy in adjudicative bodies:

The concept of legitimacy may be approached from a sociological and a normative dimension and the distinction could be useful to illustrate this point. On the one hand, the more positive the public's attitude towards an institution`s right to govern, the greater its popular legitimacy.[2] However, this legitimacy is fragile. Particularly for FB's Board, which will not have a reservoir of legitimacy accumulated over a long history to draw upon. On the other hand, legitimacy can also have a normative meaning, referring to whether the claim for authority is well founded.[3] Given the global challenge that FB faces, the legitimacy of FB Oversight Board must be strengthened on both fronts, but even more so in the latter. Building strong normative legitimacy could provide a standard for judging the Board and deciding if it deserves support. Also, normative legitimacy can influence sociological legitimacy, or perceptions of justified authority, and thereby, the extent to which it will undergird or undercut the work of the FB's Board.

There is rich literature on what elements contribute to the legitimacy of adjudicative bodies. There are even concrete writings on the legitimacy of international adjudicative bodies. What determines their legitimacy? Scholars and practitioners have identified three key elements to the legitimacy of international adjudicative mechanisms: 1) fair and unbiased decisions; 2) an interpretation and application of rules consistent with its scope and purpose; and 3) that the body be transparent, independent and infused with democratic norms.[4]

Fair and unbiased decisions need to be at the heart of any adjudicative system, not only international ones, in order for them to be legitimate. "Unbiased" has traditionally been related to independence and hence has focused on nomination and selection processes, quality and soundness of reports, decisions and recommendations, the public discourse and the writings of the tribunal members. All these elements have concrete and relevant definitions dispersed among a vast international and comparative jurisprudence that should not be ignored.

Fairness, on the other hand, according to Prof N. Grossman, "need not entail an equal ratio of rulings in favor or against any given party –in fact,

[2] Richard H. Fallon, Legitimacy and the Constitution, 118 HARV. L. REV. 1787 (2004–2005)

[3] Daniel Bodansky, "The Legitimacy of International Governance: A Coming Challenge for International Environmental Law?", 93 AM. J. INT'L L. 596, 601 (1999).

[4] N. Grossman, Legitimacy and International Adjudicative Bodies, 41 Geo. Wash. Int'l L. Rev. 107 (2009-2010)

most human rights adjudicative mechanisms issue much more rulings against States than they do against petitioners and they may still be legitimate for both parties-" but necessarily requires equality in arms, due process and consistent application and interpretation of the law.

As to the interpretation and application of the rules consistent with its scope and purpose, this is probably among the key elements to making FB Oversight Board legitimate. Community standards have a scope and a purpose and they daily dialogue with other norms, including national, regional and international human rights norms. The decisions and rationale behind individual case solutions (if FB so chooses) should follow the same logic and dialogue between community standards and human rights norms. Additionally, that logic should be transparent and published. The independent and objective reasoning behind each decision is what an experts' board adds to a closed system like the existing legal team at FB.

Finally, transparency in this context may be defined as a quality: an office or a body, whether judicial or otherwise, is transparent "when it creates the conditions that allow society to fully and clearly understand how they act, the reasons behind their acts, as well as the costs and resources associated with those actions."[5] This factor affects the previous two as well in that without some degree of transparency (manifested through some or all of the following: publishing decisions or having them publicly available, having reasoned decisions, identifying the decision makers, their dissents and concurrences, etc.) there is no way of evaluating whether a body is biased or not, or whether its decisions and interpretations of the norms are within the reasonable scope and purpose of the law.

While the legitimacy of international adjudicating bodies traditionally derives from the state's consent to its jurisdiction, the legitimacy of FB's Board would derive from: i) building a solid foundation on fair decisions; ii) a consistent and persuasive interpretation --persuasion is one of the legitimacy function-- of the community's rules in dialogue with international human rights standards; iii) and holistically transparent mechanisms.

[5] M. Pulido Jiménez, M. González Armijo, M. Sánchez de Tagle, S. Ruiz Cervantes, J. Sáenz Andujo, "Hacia un modelo de transparencia y acceso a la información en el Sistema Interamericano de Derechos Humanos", en *Desafíos del Sistema Interamericano de Derechos Humanos. Nuevos tiempos, viejos retos*, capítulo 3. Colección DeJusticia, 2015, p. 112. Disponible en: https://bit.ly/2hoG7jt.

IV. Representation vs diversity

One of the main goals of FB's Board is bringing greater legitimacy to its content moderation system. Diversity is often mentioned as one dimension that adds towards the legitimacy of a body, particularly when the constituency that the body governs is diverse. Diversity has many meanings though and should not be confused with representation.

Obviously representation could enhance the legitimacy of the board's decisions in the eyes of the broader community, particularly those that make it to being represented within the Board. However, a 40-member Board like the one that is being proposed could hardly represent a 2.7-billion-member community. It is practically impossible to provide representation for all. Not direct, proportional, or even asymmetrical representation could be accomplished in such an uneven ratio and because of the scale that FB has, there probably couldn't be a viable ratio to work with. This issue should be acknowledged and incorporated into the design of the Oversight Board so as to create realistic expectations. Ignoring it would probably mislead users, creators and bystanders and will severely undermine the ultimate goal.

Lacking representation, which could provide for diversity of voices, technical expertise could be the next best thing. Not a representative Board but a technical body capable of overseeing FB's implementation of their own policy, in dialogue with human rights norms. Such technical membership should consider cultural, geographic and gender diversity towards its conformation to allow for an actual dialogue among different experts and avoid a single voice/view approach. This diversity is key to guaranteeing closeness and cultural understanding across borders; deal with novel inter-jurisdictional issues, and serving the vast community that FB serves. There is no unique model to fit every need. Deciding what FB's Board will be necessarily implies defining and accepting what it will not be.

V. What is expected of the Board broadly?

The draft charter suggests that "The board will be a body of independent experts who **will review Facebook's most challenging content decisions** - focusing on important and disputed cases. Among the first questions that may be raised is what for? What is the ultimate goal in having the Board review these decisions? This question impacts directly on the type of struc-

ture that FB is trying to create. Is this a Supreme Court, a Court of appeals, a peoples' court? Or is it a panel of peers' approach? Can it be an advisory body or are we set on a reviewing nature? How will this Board interact with other existing structures within Facebook (security and safety, policy and outreach, to name a few?)

If it is in fact a "Supreme Court-like approach", which seems to be the rationale for it, Yale Professors Klonick and Kadri argue that "what this really means for free speech and fair process on the internet will depend on the answer to one key question: How much will the "Supreme Court of Facebook" be like the Supreme Court of the United States?"[6] As they argue in their New York Times piece, a key element to the US Supreme Court (like every other supreme court) is that it is bound by a set of rules that remain unchanged through time: The Constitution. But, as described in brief background section to this paper, FB policies change every week and so far there has not been a formal adoption of any specific standard or rule to enlighten that process. What should the equivalent be for FB's Board?

UN Special Rapporteur David Kaye argues that international human rights norms should be the ultimate rules to govern online content moderation.[7] The adoption of such standards provides a universal basis and a somewhat common understanding on what free speech means and what guarantees should be considered when limiting it. It would also provide common language to define and understand some restrictions, avoiding contradictory and ever-expanding terms (like hate speech). Ultimately the adoption of universal rules on human rights, particularly free speech, would also guarantee some certainty against discrimination and abuse, whether these arise from governments, users or advocacy groups, and would contribute to ensuring that no one is "a priori" excluded from public debate, which is the standard set by the Inter-American Court of Human Rights in Advisory Opinion Nº5 of 1985, one of the most progressive and protective international interpretations of freedom of expression and access to information.

Having international human rights norms be the "Constitution equivalent" has many advantages. However, it also has certain limits that cannot be ignored. Key among those is the freedom of the platform to develop and protect its business and to tailor it to different audiences. Example: adult entertainment is not illegal and is protected under international human rights

[6] "How to Make Facebook's 'Supreme Court' Work?", en *The New York Times*, 17/11/18, disponible en: https://www.nytimes.com/2018/11/17/opinion/facebook-supreme-court-speech.html.

[7] https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx.

standards. Following a direct application of international human rights law, no company could prohibit the distribution or uploading of pornography or violent content to its platforms. However, every company must respect and abide by the UN Principle on Business and Human Rights, which creates concrete expectations *vis a vis* company engagement with human rights, duty to mitigate and prevent violations, duty to provide redress, etc.[8]

The dialogue between international free speech standards and Facebook's content rules should be promoted, developed and expanded. Should/could the Board be the body to do that? The answer to this question will probably contribute to defining the goals for this Board.

VI. Should it be an adjudicatory body, what are the models out there and what can be imported to this new structure: arbitral systems; judicial tribunals; international tribunals; media councils; and more.

One of the main goals of FB Oversight Board is to bring greater legitimacy to the content moderation system; but creating a body legitimate enough to do so is among the main challenges.

Although this private board to oversee and unify content moderation decisions is a first of its kind, there are already numerous different models that can and should inform the process that FB is undergoing. First, the company should take advantage of the best practices and lessons learnt from over 100 years of international adjudicatory mechanisms of different sorts if the idea is in fact to create an adjudicative body; Second, there is a legitimacy to these bodies that justify their baring in deciding how to move forward with this particular initiative; third, as opposed to national decision-making bodies, international adjudication was specially designed to deal with cultural and national differences, having gained an expertise on the matter that should be acknowledged and learnt from.

International adjudication mechanisms vary from area to area and from region to region. There are different models that could be looked at, including the International Court of Justice; the different arbitration mechanisms that were created to deal with bilateral investment treaties (i.e. ICSID); the Human Rights Committee created under the International Covenant on

[8] UN Guiding Principles on Business and Human Rights, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

Civil and Political Rights; the different regional human rights adjudication mechanisms that exist (i.e. European Court of Human Rights, Inter-American Court of Human Rights, African Commission on Human and Peoples' Rights). These are but a few examples of international adjudicatory bodies that already exist and that have faced already some of the most challenging issues that FB's board will most probably face -i.e. cultural differences, language, nationality, global norms and standards; international sources of law.

The design chosen for the Board will of course determine a number of other things, including the nature and standing of its members, the dynamics that the Board is expected to have, the way and tools it will have to deal with diversity in all its shapes and sizes (criteria, language, legal culture, etc.). Arbitration panels for example are case specific, diverse, made up of individuals selected from an existing list of accredited arbitrators. They rely upon a strong secretariat to maintain the process on track and guarantee minimum procedural cohesiveness and some institutional memory that is helpful in processes that are often times confidential. The ability to choose arbitrators helps the parties build trust around the process, providing for an opportunity for each party to choose an arbitrator and having the third usually being appointed by the Secretariat. This also contributes to guaranteeing some familiarity between the decision-makers and the issues, context, language and culture that the case has arisen in. The diversity of the pool of arbitrators is particularly relevant to these structures. Still, the decentralized nature of the body itself can attempt against cohesiveness of the decisions arising from the body. The fact that the panel is case specific can be problematic when trying to define through interpretation the meaning and implications of an otherwise general or broad international standard.

International tribunals like the Inter-American Court of Human Rights or the European Court of Human Rights are radically different from arbitration panels. They have a consistent membership through the years, which guarantees some consistency and legal certainty as to the decision-making criteria. It also guarantees a greater level of equality for their users. Most of the times there is cohesiveness to their decisions and their permanent nature provides more transparency and accountability of the panel itself. The downside of these structures is that guaranteeing diversity and representation of every potential party is not possible and they are probably more constrained in the number of cases that they can review in a given year. The Inter-American Court for example always hears cases in full. The European Court has set up a system whereby the members of the Court sit in different chambers, each having 3 judges appointed to them. Some decisions are brought to

the Grand Chamber and therefore they guarantee the cohesiveness of their decision-making more broadly.

As set out in the brief examples, there are different potential models that FB could adopt in designing their Oversight Board. Still, regardless of the one they choose, comparative and previous experiences should be consulted and taken into account when finalizing this project.

In taking these tribunals and its practices under consideration, particular attention should be paid to the lessons learned and the worst and best practices arising from them. One of the many problems that these tribunals face is the backlog that they generated. Cases take a long time to be litigated before these bodies, whether they require a lawyer or not, and access to international tribunals is not easy. Key among the questions that FB should ask itself is how is this "private quasi tribunal" going to work in scale? Will they be adding to a systematic crisis worldwide *vis a vis* access to justice, or will they be contributing to at least partly solving that issue. The goals and expectations for the Board are not a minor thing to consider here: if this is an appeals Court, a users resort of some sort, how will this body deal with 2.7 billion users and the amount of content they generate?

FB already faces criticism for their internal immediate processes for reviewing company decisions on content moderation. These include not only removal decisions, but also decisions to downgrade or foster the circulation of certain contents versus others. These issues will not be solved with an oversight board and in fact, the creation of an oversight board should not draw the attention away from them -regular appeals mechanisms within the company-, as these are the basis for any potential redress for wrongful or unfair content moderation practices.

A Supreme Court (or Constitutional or international Court) style approach could make a much more substantive contribution and fit more smoothly into the scaling issue. However, there should be a lot more clarity as to expectations, criteria, process, case selection, standing (for NGOs, users, consumer organizations?), etc. for this board to receive and select cases, deal with them and make them public. Decisions arising from the Board should also impact the resolution of similar cases within FB' regular content moderation operations and appeals processes, thereby turning the Board into an internal reference and authoritative body for the company to interpret ToS in dialogue with international human rights law. Otherwise, and because of issues of access, scale and relevance, the exercise will soon derive moot.

VII. Who will be part of this Board? Requirements to be members of the Board?

In thinking about adjudicative bodies, who decides is as important as how they decide. Still, the question that we propose is not literally who will be on the board but rather what will the criteria be for selecting those members. The question is very much linked to the expectations that one may have for this board. Who or what is it overseeing and what for?

Selection criteria is key to guaranteeing legitimacy as explained above. Comparative and existing bodies should also be brought to bare in defining the eligibility criteria for Board members. Knowledge and expertise in international human rights and particularly free speech should be among the qualities of any candidate. This recommendation should not be taken lightly. FB already has security and safety councils where third party experts participate and actively engage in designing the company's policy and terms of service. Reviewing content moderation and curation decisions necessarily implies balancing freedom of expression against other rights.

Most international human rights adjudicative bodies require that its members meet the requirements in their own countries to be judges and that they have demonstrated expertise in human rights. While not every judge will have the same education (some will be from common law countries, others from civil law traditions; some will be from the global north, others from the global south), they will all have some sort of legal education- Members of the Inter-American Commission or the UN Committee need not be lawyers but need a number of years of experience, sound knowledge of international human rights law and high moral and ethical standards. In order to be in an arbitration panel, there are certain requisites including being versed in law that arbitrators need to meet.

As opposed to other structures common to a number of different companies -like the security and trust committee, or the children's safety group- the Oversight Board that is being proposed is intended to deal with limits and restrictions over speech. While these other bodies are mostly made up of experts on children's issues, vulnerable populations, risk management, violence and abuse, there probably are not many free speech experts within those groups. Different areas require different skills and a key question that FB should ask in defining criteria for the Oversight Board is what their role will be and what their skills should be.

VIII. Conclusions:

First and foremost, in designing the Oversight Board, FB and its team should evaluate the impact of such a body on the human rights of its users, particularly freedom of expression, due process, access to justice, equality and non-discrimination. It should also evaluate how this Board, whatever the structure it ends up having, will contribute towards implementing the UN Business and Human Rights Principles.

Legitimacy is a key element that both FB and its users crave in its content moderation. Still, legitimacy may be defined in different and varied ways and legitimacy among decision-making bodies require certain key characteristics that should not be ignored if the goal is set for an adjudicative body. Scale and diversity in this particular case -with a 2.7 billion user constituency across 180 countries- pose additional and complex challenges to accomplishing legitimacy in traditional representational models. These challenges should be acknowledged and addressed systemically. FB should not target that which it cannot provide.

There are still more questions than answers surrounding the creation of FB Oversight Board. In defining and answering those questions, particular attention should be paid to the objectives and expectations for the Oversight Board. After many consultations, it is clear that different organizations and different people will have different expectations for this body. Clarifying what FB is thinking on these points is key and it will be important to provide new spaces for dialogue and comments after the company lands a concrete proposal for its Board and before it actually implements it.

Should FB decide to create a private adjudication model, it should bare in mind best practices and lessons learned from the many and diverse structures that adjudication mechanisms have adopted over the years, with varying results. International adjudication mechanisms are particularly relevant to look at since they are cross-borders, serve an inter-jurisdictional multicultural constituency, are courts and mechanisms of last resort, and usually resolve complex and intertwined legal issues.

Following FB's human rights obligations as well as the particular framework that speech restrictions enjoy internationally, particular attention should be paid to the technical expertise required for members of the Board to join. Developing concrete, specific and clear criteria for the selection of the Board is key to its founding, regardless of who chooses the first set of members. The selection criteria will depend on the objectives and the concrete expectations for this Board.

About CELE

The Center for Studies on Freedom of Expression and Access to Information is a research Center housed at Universidad de Palermo, in Buenos Aires, Argentina. The Center provides legal technical research to promote the understanding and development of freedom of expression and access to information, particularly in Latin America. Since 2012 we have an Initiative for Freedom of Expression online (iLEI- Spanish) under which we have studied and produced research pertaining to online free speech, access to information and privacy particularly under the framework of the Inter-American human rights system and standards. Our strategies to affect change include research; capacity building; and promoting spaces for high level reflection and debate. Please visit us at www.palermo.edu/cele and at www.observatoriolegislativocele.com.

Annexed III

Commentaries to Twitter's proposed change in rules regarding "Dehumanizing content"*

I. Introduction

At CELE we welcome the opportunity to submit comments and contribute towards an improved dialogue about the rules that govern spaces like Twitter, which are fundamental to today's exercise of freedom of expression, freedom of opinion and access to information. Collaborative processes in the creation of such rules provide an opportunity for users, the community, civil society and any interested stakeholder to take part in the discussion, shaping and framing of policies that will later on affect them.

II. Broadly, and of certain concern: A shifting approach to freedom of speech and expression. The approach towards "A healthy conversation"

Twitter's famous primary goal to provide a space to "speak truth to power" has been the force behind the company's expansion and increasing growth since its inception. The company's policies on limited intervention and free flow of information and ideas of all kinds has contributed to make it a unique space, particularly in Latin America, where there is an absence of other platforms and companies providing a similar space. Recently, however, the focus has shifted from "speaking truth to power" to contributing to a "healthy conversation".[1] The change in semantics seems particularly important to analyze the proposed policy and what it might imply. And the compatibility of

* This document was prepared by the Center for Studies on Freedom of Expression and Access to Information (CELE) at Universidad de Palermo in response to Twitter's invitation to comment on a proposed rule addressing "dehumanizing content".

[1] https://twitter.com/jack/status/969234275420655616.

such policies that foster "healthy conversations" with freedom of speech and expression as recognized internationally and regionally in the Americas will probably depend on the scope and means to implement them.

There are individual and social dimensions to the right to freedom of expression that contribute towards its radical importance for the development of one's own identity, as an individual and as a group; as an instrumental means towards the exercise of other rights (political, social, economic and cultural ones); and a necessary requirement for democratic societies.[2] Freedom of expression protects not only polite, politically correct expression but also shocking, offensive, and otherwise distasteful expression. And dissemination of information and expression of all kinds are linked and tied and are indissoluble. A restriction or interference with dissemination would necessarily affect the exercise of free speech and directly impact the social dimension of this right. It is therefore implied that a restriction or interference with one constitutes necessarily a violation of the other.[3]

Because of its importance both individually and socially, international human rights law treaties mandate that limitations to freedom of expression be clearly established by law, pursuant to a legitimate objective, be necessary in a democratic society and proportionate. The UN Special Rapporteur for Freedom of Expression, David Kaye in his report on June 2018 recommended that internet companies align their moderation policies to international human rights law and cited the UN Principles on Business and Human Rights as an authoritative source for such recommendation,[4] in order to foster and promote free speech in times of increasingly privatized spaces.

The reason behind these rules and these protections lies not on the need to empower those who abuse speech, but on the need to guarantee that no voices are excluded "a priori" from the social and political debates. And although shocking or distasteful statements may sometimes not seem to contribute towards healthy conversations, there are objective reasons that suggest otherwise: 1) what constitutes taste, shock or disrespect lies more often than not on the eyes of the beholder; 2) social discourse is heated, diverse and passionate, and not surgically clean, academically tested or scientifically proven; 3) speech and expression does not occur in a vacuum and responds to varying cultures, contexts and environments; 4) most importantly, one expression alone does not constitute a conversation, but rather

[2] OAS RELE, Inter-American framework for the right to freedom of expression, 2009.
[3] IACtHR, OC5/85 http://www.corteidh.or.cr/docs/opiniones/seriea_05_ing.pdf.
[4] https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx.

a statement which may give rise to a rich and important conversation, the health of which remains to be determined.

While we understand that the shift towards "fostering healthy conversations approach" may seek to address a valid concern, we are concerned that this policy change represents a major change in the company's understanding of freedom of opinion and expression and its role in contributing towards its full realization.

III. The proposed policy

"Twitter's Dehumanization Policy

"You may not dehumanize anyone based on membership in an identifiable group, as this speech can lead to offline harm.

Definitions:

Dehumanization: Language that treats others as less than human. Dehumanization can occur when others are denied of human qualities (animalistic dehumanization) or when others are denied of their human nature (mechanistic dehumanization). Examples can include comparing groups to animals and viruses (animalistic), or reducing groups to a tool for some other purpose (mechanistic).

Identifiable group: Any group of people that can be distinguished by their shared characteristics such as their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, serious disease, occupation, political beliefs, location, or social practices."

1. What constitutes "dehumanization" and how should it be addressed?

The proposed policy seeks to address "dehumanizing" content that may contribute to violence. We argue that although the goals are well intended, the policy may be too wide and too broad and could potentially become a tool for discretionary censorship. Everything from discriminatory or inflammatory to politically incorrect content could fall under the categorization of "dehumanizing content".

Following Professor Timothy Garton Ash, there should be a clear distinction between speech that incites to violence, and therefore causes damages; and

speech that is disrespectful or offensive. "If we don't allow any speech that may be offensive to anyone, there will be very little left out there to talk about."[5]

From the blog post that announces the change, the new policy seeks to: "expand our hateful conduct policy to include content that dehumanizes others based on their membership in an identifiable group, even when the material does not include a direct target." It goes on to say that "Many scholars have examined the relationship between dehumanization and violence. For example, Susan Benesch has described dehumanizing language as a hallmark of dangerous speech, because it can make violence seem acceptable, and Herbert Kelman has posited that dehumanization can reduce the strength of restraining forces against violence."

Still, as explained by the quoted authors, what constitutes dehumanizing content may depend upon context and culture. However, the policy does not refer to context nor does it qualify the prohibition to their relevance at all. As currently stated, the proposed policy selectively chooses to focus on what prof. Benesch describes as a "hallmark for dangerous speech" but ignores other rather important aspects of her argument: "Dangerous Speech cannot be identified solely by the hallmarks or by any aspect of its content, since its capacity to inspire violence depends so much on its context – on who spreads it, how, to whom, and in what social and historical context." And that "Such efforts (to combat it) must not infringe upon freedom of speech since that is a fundamental right – and when people are prevented from expressing their grievances, they are less likely to resolve them peacefully and more likely to resort to violence."[6]

Based on the language of the proposed policy, "dehumanizing" content will be prohibited in the platform. This blank prohibition without exceptions equals a ban on such content from the platform. The impact of this policy on public interest debates, colloquialisms, and speech may be even more far reaching and restrictive if the content is proactively sought through artificial intelligence and decisions are left to algorithms.

The examples provided as to what may constitute dehumanizing content includes:

[5] https://www.lanacion.com.ar/2176297-timothy-garton-ashlanzar-discursos-odio-no-deberia.

[6] https://dangerousspeech.org/the-dangerous-speech-project-preventing-mass-violence/.

If taken literally, under this policy any content that refers to Argentina's soccer team fans of River Plate as "gallinas" could be interpreted as dehumanizing; or when someone refers to Mexicans as "Guey" (from the Spanish word "buey", meaning "ox") that could also be interpreted as dehumanizing. In the same breath, when people are too strong in some cultures they are referred to as "bulls", or "Hoaxes", what to do with that?

In a more complex example, comments referring to women as "potras" or men as "potros" (horses, and meaning beautiful in a very 90's style) would also be flagged in violation of the policy. In these instances, although relating a person's beauty to a horse may not be desirable, regardless of how well intended, it probably should not be banned either. The example could be problematized further, if we consider references of women as "yeguas" (little horses), which in fact has rather different implications, is offensive and implies mean spirited or evil. Same animal, same "dehumanization", different words, uses, contexts, cultural understandings, different implications. Would they fit under this policy? Should these last references be banned?

But even if the content is dehumanizing in the sense that it could, if combined with other factors, naturalize violence, and could potentially (exceptionally) merit banning, we argue that some of this content should also be exempt from the banning rule and preserved for public access for reasons that vary, but should include: accountability; humor; and most importantly, positive consequences deriving from such speech: counter speech. Additionally, the exception of newsworthiness should be explicitly applicable to this policy. A couple of examples to illustrate them:

1) As recently as this past week, in October 2018, news broke that

a senior authority within the Trump government "liked" a meme circulating on Twitter about the Obamas staring at a banana; a clearly racist meme intended to portray African Americans as apes or monkeys. The Huffington Post published an interesting article about it and provided further background on past instances of executive officials within the Trump administration sharing, liking, disseminating racist, bigot content (https://www.huffingtonpost.com/entry/epa-andrew-wheeler-social-media-conspiracy-theorists_us_5bbcee58e4b01470d055fe56). This kind of public accountability is fundamental to democratic societies. It is fundamental that the people learn what their representatives think and stand for, and even more so when that is racist, segregationist, sexist, etc. Companies should not contribute to "hide" these instances as they would contribute to foster impunity and the exception of newswrthiness should be explicitly included in the rule.

2) Memes and humor: an important part of satire and parody implies mocking and ridiculing social elements that are common to our reality, religion, political situation or status. An image is worth a million words and very often these ridiculizations imply resorting to some sort of "dehumanization" as defined by this policy: either focusing on the mechanic nature of what the person or group does or using animals to create parallels. Here is an example of a caricature that was judicialized in Colombia for comparing Former President Uribe's supporters with pigs. The second drawing portrays the pigs making a claim against the artist for comparing them to politicians. The case was dismissed in court in favor of the caricaturist but what would have happened in Twitter under this policy?

3) Counter speech: The example comes from Argentina, where a famous rock star during an academic exercise with a group of journalism

students was asked his opinion about allegations of sexual abuse and misconduct involving minors within the national rock and roll scene. This person responded among other things that "some women need to be raped in order to enjoy sex" and that "16 year-olds should not be considered minors under this particular laws". The interview was uploaded to Facebook by one furious student and there was a push to have it blocked. However, the social condemnation and counter speech that the interview brought about (including shows being cancelled, national debates on television, critiques in the press, national uproar about such sexist comments) was so important and raised the issue to the agenda so effectively that having the content blocked would have meant putting unnecessary restrictions on healthy, productive, and interesting debates and chilling a necessary conversation on gender equality and sexism in our societies.

The examples provided are just a few to show that policies like these should not be taken lightly, should be carefully designed and redacted and very carefully applied so that it does not infringe upon free speech and curtails people's ability to debate the important things of our times. These are usually heated discussions, potentially offensive, shocking to some, sometimes very polarized discussions. However, they are needed and in times of increasing restrictions among social media, Twitter, which has a history for protecting speech, should provide some space for those to happen.

2. How does the policy define "vulnerable groups"?

Some internet companies, in designing their policies regarding hate speech, threats, violence, create their own definitions of what constitutes "vulnerable groups". The identification of these vulnerable in some instances overlap with what is traditionally under human rights law or practice a vulnerable group or population, but often times there are significant discrepancies. Facebook, for example, considers that public officials and heads of state are "vulnerable groups" for certain types of content.

Although the policy doesn't include these group within the vulnerable groups that it is intended to protect, the enumeration does not explicitly exclude it either. If public officials, politicians, etc are interpreted as being part of a vulnerable group, the policy could impact political speech disproportionately, generating an unwarranted restriction on free speech and

expression. Political humor, parody and satire could take a big blow to the detriment of public interest debate and democratic exchange.

3. How will this policy be implemented?

Because of the relevance that context, culture, as well as the who, how and when, play in defining content as "dehumanizing content" that could naturalize violence or become a "hallmark for dangerous speech", Twitter should not proactively scan content to detect it, but should maintain its traditional "flagging requirement" practice. Using mechanic or automatized means to monitor, detect and act upon this kind of content could lead to over-blocking and/or overreacting over content that is not perceived as offensive by anyone, or otherwise even if offensive, should not be banned.

If the proposed policy is adopted, it should be implemented upon user's flagging content. In the same breath, maybe not any one flagging should suffice. An alternative could be to require a minimum number of flaggers for these kinds of content that do not refer to any one person but groups; and there should be a requirement that the flagger be a member of the actual group that is "offended" by it, or an organization that represents the interests of such group. For example, questioned content referring to women should be attended only if flagged by women or organizations representing women.

articular attention should be paid, as stated earlier, to defining vulnerable groups and carefully distinguishing who the content is targeted to, rather than applying the policy in abstract. Minorities counter speech can be regarded as offensive to members of the majority groups. In the second example quoted in the blog it says: "[gender] is only good for sex." While filling in the blanks, the phrase would not mean the same if we fill in "women" than if we fill in "men".

In all cases, the policy should be interpreted restrictively, paying close attention to context, culture, and other aspects of the discourse itself that would make this kind of content "violent" or conducive to violence. The company should conduct appropriate oversight of how the policy is being interpreted and applied. And there should be some discussion as to who and how could there be recourse to an otherwise restrictive decision regarding content referred to a group.

Finally, transparency is fundamental to any content moderation and there should be mechanisms in play to account for statistics on content blocking, removal, account suspension, and any other measures adopted in relation to this policy; as well as transparency in how the policy is being interpreted through time to avoid excessively discretional calls banning content from

the platform. The content generator should be given a proper explanation as to why the content was removed and how the policy was interpreted in his/her particular case.

In the case of this policy, transparency may be even more important, as it would contribute to the goal set up by the policy itself. If the goal is to promote a "healthy conversation", people should understand clearly what that means and how their interactions run against the health of the exchange. Twitter is in a privileged position to do this in implementing this policy transparently and openly, thereby educating while enforcing.

IV. Conclusion:

The proposed policy clearly responds to a shifting approach towards freedom of expression and access to information that raises serious concerns regarding their compatibility with international and regional human rights standards. Specifically, the proposed policy is too broad and wide and fails to narrowly establish the limitations that will be imposed, allowing for excessive discretion on its interpretation and implementation.

If adopted, the policy should clearly establish exceptions to the rule and clarifications on how it will be read, interpreted and implemented and what recourse will be provided to question the decisions. Transparency is vital to guaranteeing sound moderation practices and, towards this policy specifically, in contributing towards the overall objective of the norm.

We would like to thank again the company for the opportunity to submit comments and reiterate our belief that these exercises bare the utmost importance in collaboratively shaping and modeling the rules that will guide future debates on the internet.