UNIVERSITAS
Revista de Ciencias Sociales y Humanas de la Universidad Politécnica Salesiana

https://doi.org/10.17163/uni.n34.2021.08

Artificial Intelligence, algorithms and freedom of expression

Inteligencia Artificial, algoritmos y libertad de expresión

Manuel Ernesto Larrondo
Universidad Nacional de La Plata
larrondomanuel@gmail.com
https://orcid.org/0000-0002-0569-502X

Nicolás Mario Grandi
Universidad Nacional de La Plata
drgrandinicolas@hotmail.com
https://orcid.org/0000-0003-4191-8849

Abstract
Artificial Intelligence can be presented as an ally when moderating violent content or apparent news, but its use without human intervention that contextualizes and adequately translates the expression leaves open the risk of prior censorship.
At present this is under debate within the international arena given that, since Artificial Intelligence lacks the ability to contextualize what it moderates, it is presented more as a tool for indiscriminate prior censorship, than as a moderation in order to protect the freedom of expression.
Therefore, after analyzing international legislation, reports from international organizations and the terms and conditions of Twitter and Facebook, we suggest five proposals aimed at improving algorithmic content moderation.
In the first place, we propose that the States reconcile their internal laws while respecting international standards of freedom of expression. We also urge that they develop public policies consistent with implementing legislation that protects the working conditions of human supervisors on automated content removal decisions.
For its part, we understand that social networks must present clear and consistent terms and conditions, adopt internal policies of transparency and accountability about how AI operates in the dissemination and removal of online content and, finally, they must carry out prior evaluations impact of your AI on human rights.

Keywords
Artificial Intelligence, automatic content moderation, fake news, freedom of expression, social networks.

Suggested citation: Larrondo, E., & Grandi, N. (2021). Artificial Intelligence, algorithms and freedom of expression. *Universitas, 34,* pp. 169-186.

Resumen
La Inteligencia Artificial puede presentarse como un aliado al momento de moderar contenidos violentos o de noticias aparentes, pero su utilización sin intervención humana que contextualice y traduzca adecuadamente la expresión deja abierto el riesgo de que se genere censura previa
En la actualidad esto se encuentra en debate dentro del ámbito internacional dado que, al carecer la Inteligencia Artificial de la capacidad para contextualizar lo que modera, se está presentando más como una herramienta de censura previa indiscriminada, que como una moderación en busca de proteger la libertad de expresión.
Por ello luego de analizar la legislación internacional, informes de organismos internacionales y los términos y condiciones de Twitter y Facebook, sugerimos cinco propuestas tendientes a mejorar la moderación algorítmica de contenidos.
En primer término, proponemos que los Estados compatibilicen sus legislaciones internas respetando los estándares internacionales de libertad de expresión. También instamos a que desarrollen políticas públicas consistentes en implementar legislaciones protectoras de las condiciones laborales de supervisores humanos sobre las decisiones automatizadas de remoción de contenido.
Por su parte, entendemos que las redes sociales deben presentar términos y condiciones claros y consistentes, adoptar políticas internas de transparencia y rendición de cuentas acerca de cómo opera la IA en la difusión y remoción de contenido en línea y, finalmente, deben realizar evaluaciones previas de impacto de su IA a los derechos humanos.

Palabras clave
Inteligencia Artificial, moderación automática de contenidos, *fake news,* libertad de expresión, redes sociales.

Artificial Intelligence and freedom of expression. State of the art and preventive proposals

Freedom of thought and human expression is the fundamental basis of any democratic society. This is recognized by art. 19 of the Universal Declaration of Human Rights, as well as art 18 (freedom of thought) and art 19, 1) and 2) of the International Covenant on Civil and Political Rights (ICCPR) by providing that "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and

ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice" (UN, 1966).

However, at the same time, said article recognizes that this right may be subject to restrictions that must be established by law necessary to: a) ensure respect for the rights or reputation of others and b) the protection of national security, the public order or public health or morals.

For its part, the Inter-American Human Rights System establishes that right in the same sense and scope of broad protection in Article 13 of the American Convention, with the particularity that it expressly prohibits censorship in any form and only considers it, in a prior manner, to protect the rights of children and adolescents. In the same sense, it emphasizes that whoever exercises this right is subject to the subsequent responsibilities that must be established by law, respecting their need, legitimacy, and proportionality.

Regarding the possible restrictions and responsibilities subsequent to the exercise of the right, we note that the UN Human Rights Committee - interpreting the scope of art. 19 of the ICCPR — is inclined towards an even more protective position, considering that freedom of opinion "does not authorize any exception or restriction" to its exercise, either "by law or by another power" (UN, 2011).

It is evident that the limits and scope of the exercise of this right are the center of analysis and complementary interpretation on the part of the main international organizations.

Without going any further, it is enough to cite the validity of OC 5/85 issued by the Inter-American Court which indicates that freedom of expression is not exhausted on the individual but also includes the collective dimension, underlining that free thought and its dissemination are inseparable, in such a way that a prior limitation —state or private— to any of them would be incompatible with the inter-American standards that protect this right (Sec. Gral. OEA, 2017).

In Argentina, the Supreme Court of Justice of the Nation (CSJN), has followed the same line, noting that freedom of expression is one of the most important freedoms, while, without its proper protection, the democratic system would function only in an apparent fashion.

These intercontinental standards were agreed throughout most of the 20th century and the beginning of the present 21st century. With the emergence of social networks and other online intermediaries, the moment came when practically half of the world population exercises the triple action of

disseminating, investigating, and disseminating information through the Internet through the main platforms that use "Artificial Intelligence" (Hereinafter AI) with which we interact every day.

Now, what is AI? At the moment, it has not been possible to agree on a single universal definition. On this occasion, we will be inclined to cite the one provided by the Office of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of the UN when saying that "it is a 'constellation' of processes and technologies that allow computers to complement or replace specific tasks that would otherwise be performed by human beings, such as making decisions and solving problems" (UN, 2018).

The Rapporteurship adds that at the base of AI are the "algorithms" that are computer codes designed and written by human beings. All kinds of data that an algorithm processes are translated and produces a specific result such as inferences, suggestions, or predictions. Thus, the flow of infinite data generated by a person per second when interacting on the network, leads to the necessary development of AI in the face of the material impossibility that a person can do it on their own in a short time and efficiently.

As a proof of this, it will suffice to note that the volume of online data generation grows exponentially every second, to the point that in just one minute of browsing the Internet on Google, more than three million searches are carried out, on Facebook more than thirty million messages and more than two million videos are viewed, more than four hundred and fifty thousand tweets are published on Twitter, more than forty-six thousand photos are posted on Instagram, more than four million hours of videos are uploaded on YouTube and almost double on Netflix. This large amount of information has been called big data, and it arises from the interrelation of our electronic devices connected to the Web. Storage capacity is no longer measured in kilobytes composed of a four-digit number, but evolution has led us to alhella bytes, which has twenty-seven figures (DAUS, 2019), that is, the information is six hundred times greater.

This immense volume of data that we generate through social networks and intermediaries makes up an eccentric virtual place where we converge with other people, as well as with "bots" and other automated systems based on AI.

Although the latter contribute to the human exercise of free expression, at the same time, concrete evidence has emerged that alerts us to the serious

risk that, little by little, AI "usurps" the human right to receive, investigate and disseminate content in line when deciding, in an automated way, which content remains and which is removed according to the "terms and conditions".

A first approach to the use of AI by platforms could be considered appropriate for the removal of violent content, disinformation, or that which incites hatred, for example.

However, anticipating the development and conclusion of our proposal, we consider that without human moderation that contextualizes and adequately translates the expression line, there may be a serious risk that the platforms will give precedence to AI as an automated moderator of online content, thus breaching with the aforementioned international standards in relation to the human right to receive, investigate and disseminate, which would be limited not by a necessary law, with a legitimate and proportional purpose, but by a de facto "constellation" made up of inhumane algorithms.

Therefore, our work will begin by explaining how international organizations conceptualize and diagnose the use of AI in the automated moderation of online content, as well as pointing out its main advantages and disadvantages. Next, we will analyze what are the implications of AI in the exercise of freedom of expression through platforms such as Facebook and Twitter in order to assess whether its implementation has contributed, in recent times, to restrict expression or not. Finally, and based on the factual and legal framework analyzed, we postulate the necessary intervention of human supervision when the AI suggests the removal of online content of public interest.

AI and automated online content moderation. Reasons for its implementation and the necessary human supervision

AI is usually considered as a set of automatic and impartial technological systems aimed at facilitating the effectiveness in the moderation of content in search of mitigating possible hateful, discriminatory, terrorist, etc., discourses and thus improving the experience of its users and the Citizenship Construction.

However, the UN has remarked that, in the field of content moderation, although AI has its positive aspects, the negative ones are also significant.

Among the benefits of using AI, the UN highlights that the personalized selection of content enhances the online experience of each person allowing them to quickly find the requested information, even in different languages. However, this initial virtue has as a disadvantageous element the limitation that each person faces to access different points of view, thus interfering with the personal possibility of delving into and confronting different ideas and opinions with individuals who have another ideological, political, religious or social position. In this way, this content segmentation that appears to be very useful and effective, could, at the same time, reinforce individual beliefs and lead to the exacerbation of violent content or misinformation with the sole purpose of maintaining the user's online participation (UN, 2018).

Sandra Álvaro explains that algorithms are already part of our daily lives, using as an example Facebook who has an algorithm called Edgerank that analyzes our browsing data — the "likes" we grant, the friends we have, and the comments we make — and with this, it profiles us in order to show us those stories that we like and hide those that bore us and show us new friends that match our profile and ideology (Álvaro, 2014).

This situation, which generates a kind of information bubble, has aroused the interest of the European Union as it warns that human beings who interact with AI systems must be able to maintain full and effective self-determination about themselves and be able to participate in the democratic process. That is why it urges that AI systems must not coerce, manipulate, infer or unreasonably group human beings.

At the discretion of the European body, AI should then be designed to increase, complement, and enhance human cognitive, social and cultural skills, thus following human-centered design principles (Eur. Comm., 2019).

Faced with this new reality in which information of all kinds overflows on the network, already in March 2018 the European Commission urged internet platforms to use automatic filters to verify and, where appropriate, remove extremist content, although —at the same time— suggested that human review be used in order to avoid errors that come from automated systems.

This is so, since the use of AI in the automated moderation of content can affect the exercise of freedom of expression since, for the moment, its limitations include the impossibility of being able to evaluate the context, the idiomatic uses, and cultural aspects of human beings.

Although in recent times AI has exponentially improved in Natural Language Processing (NLP), it has not yet achieved such a development that

allows it to understand all the linguistic and cultural nuances by which humans express themselves.

This has led to the fact that, when moderating content automatically, the algorithm used by the platforms has also eliminated images of nudity with historical, cultural, or educational value, historical and documentary accounts of conflicts, evidence of war crimes, interventions in against groups that promote hatred or efforts to challenge or report racist, homophobic or xenophobic language.

This would show that, in this face of AI development, we still find weak automated systems that need human supervision to be able to carry out their actions without affecting other rights.

It is precisely in this context that AI loses its "magic power" to solve the removal of online abusive content, hate speech, or the eventual misinformation. For this reason, internet companies have urged users to refine the content observed with different contextual elements, although, it should be clarified, the viability and effectiveness of these guidelines are not clear (UN, 2018).

In this sense, the UN Human Rights Committee understands that, unlike people, algorithms lack corpus and mind, that is, they are not yet capable of understanding when an expression is ironic or is a parody, or to confirm with precision whether a certain demonstration can be described as praise of "terrorism". Therefore, the automation of its mathematical operability tends more to opt for a quick result consisting of limiting or removing a certain expression without taking into account that this results in considerably affecting the human right to receive, investigate and disseminate (UN, 2018).

In the same way, the use of AI when uploading files on the web, in order to protect the intellectual property rights of both the videos has raised doubts due to the large number of blocks that occur, which, added to Possible leaks from the content linked to terrorism or other extreme positions may arrive at the opposite, that is, instead of protecting rights, totalitarian regimes can be established by applying an automated prior censorship.

Indeed, while the use of cryptographic comparison algorithms is extremely useful to detect images of sexual abuse of minors, on the contrary, their application to "extremist" content —which generally requires contextual evaluation— is difficult without the existence of clear norms that define what "extremism" is (UN, 2018).

In this sense, the UN understands that the platforms should make transparent the way in which they use AI, explaining in detail with aggregated

data that illustrates examples of real cases or hypothetical cases in order to clarify how their interpretation and the application of specific norms are (UN, 2018).

Likewise, since it is the responsibility of companies to prevent and eventually reduce the negative effects on human rights with the use of AI, it is clear that part of their transparency policy should consist of beginning by recognizing the important limitations that automation suffers in moderation of content, such as those difficulties already mentioned about the interpretation of the context as well as the wide variation of idiomatic nuances and the meaning and the linguistic and cultural particularities. That is why, at a minimum, current and future technology to address issues related to large-scale data should be subject to a rigorous audit and, of course, have contributions from civil society tending to enrich the analysis.

To end this section, we want to refer to the last aspect related to our proposal that human supervision be guaranteed in the event of the possible automated removal of online content.

We refer specifically to the one by which the platforms are urged to strengthen and guarantee that the automated moderation of online content has the possibility of review and supervision by human beings trained in knowing international standards of freedom of expression.

To this end, the UN states that it is essential that adequate protection be provided to the working conditions in which they perform tasks since they must be compatible with human rights standards applicable to labor rights (UN, 2018).

Such an application has its basis, for example, in a specific case of "job insecurity" of moderators who worked for Facebook.

Indeed, in 2015 this company had less than 4,500 people as moderators of audiovisual content, but, due to COVID-19, it had to expand the workforce by hiring some 15,000 moderators, most of whom are under the modality of subcontractors in various cities around the world (Dublin, Berlin, Manila).

This is how the magazine "The New Yorker" reports that moderators often work odd hours in different time zones in the world, to which is added the lack of sleep and the strong psychological impact they suffer from absorbing everything they see in their screens without having a standardized "protocol" to indicate what content should stay online and what shouldn't.

As a result, in May 2020 thousands of moderators joined a class-action lawsuit against Facebook alleging psychological disorders and, for this reason, agreed with the company a settlement of USD 52,000,000 (Marantz, 2020).

The human supervision that we postulate for the review of the automated content decision, although it would not be able to absolutely prevent online censorship, it is possible to anticipate that it would contribute to making up for the serious defects of the AI that cannot - yet - interpret contexts, linguistic terms, irony, satirical humor, artistic images of nudity, etc.

Let's see below certain specific cases that, according to our position, accompany this proposal to implement human supervision in the face of misinformation and automated content removal.

How Twitter and Facebook operate

Twitter guidelines

The social network Twitter has a series of rules entitled "General policies and guidelines" that must be respected in order to use the platform. One section of those guidelines is linked, as far as our analysis is concerned, to online content that relates to topics of public interest.

Although this social network anticipates taking various kinds of measures on tweets that violate its rules, at the same time it recognizes that on certain occasions —without specifying which ones, at least as an example— they keep certain tweets online that may be useful to society, since otherwise they would be erased. When would a tweet be considered in the public interest? The platform reports that it qualifies as such when it is presented as "a direct contribution to the understanding or debate of an issue that concerns the entire public" (Twitter, 2020).

Thus, this social network highlights that those tweets issued by government officials are of public interest because it is important to know what they do in order to debate their actions or omissions. Twitter thus anticipates that it will give prevalence to the dissemination of content of public interest based on the following four criteria that make up an exception to the direct removal of content, specifically:

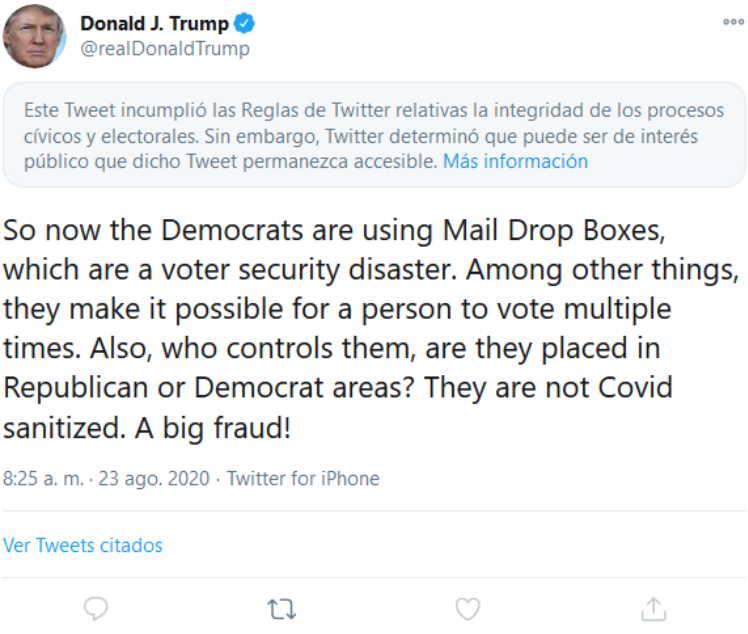
- The tweet violates one or more Twitter rules.
- The author of the tweet is a verified account.
- The account has more than 100,000 followers.

- The account represents a current or potential member of the local, national or supranational government or legislative branch: i) current holders of a leadership position elected or appointed by a Government or legislative body, or candidates or nominees for political office.

It can happen, however, that a public official publishes a tweet violating the terms and conditions of Twitter. In that case, as an exception, the platform informs that one can choose to keep the tweet, which would otherwise be deleted. For this purpose, Twitter inserts behind it a notice that is intended to contextualize the breach of the rules and allow people to enter to see it if they wish.

Going to the use of AI, it expresses that, by placing that notice, the possibility of interacting with that tweet is also decreasing, through "Like", "Retweet" or by sharing it on that same social network to generate that the Twitter algorithm avoid recommending it. Thus, it is noted that through these actions an attempt would be made to restrict the scope of the tweet, at the same time, guaranteeing the public the possibility of viewing it and discussing the subject in question.

As a first observation to be made, we want to highlight the limited and restrictive framework that Twitter implements when it requires an account to have 100,000 followers in order to be included in the conditions of the public interest standard. The quantitative measurement based only on the number of followers - which could well be made up mainly of bot accounts - we believe that it would undermine a qualitative analysis of the discourse in question as long as it defines whether or not is of public interest, since it should be used Human supervision following jurisprudential standards such as that of the Inter-American Court of Human Rights that defines the public interest with those opinions or information on matters in which society has a legitimate interest in keeping informed about the operation of the State or general rights and interests (IACHR, 2009, 2011). Returning then to the analysis of the measures that Twitter implements on this point, we can see an example of this in particular in one of the many tweets that President Donald Trump issued on August 23, 2020, on the occasion of the presidential electoral contest.



As can be seen, in said tweet, President Trump alluded to possible electoral fraud that could be committed through the citizen vote-by-mail system. In that case, Twitter inserted a notice in the tweet that warned the public about the breach of the rules regarding the integrity of the civic electoral processes, although it was also decided that the tweet remains accessible. For more information, a link was attached to refer the user to reading the policies and general guidelines on public interest cited above.

For these particular cases, we note that Twitter informs that its "Trust & Safety Team", which is made up of professionals who are experts in various fields, will implement a second analysis, in order to analyze the tweet and give an opinion to keep or not its visibility based on public interest criteria. Subsequently, the first recommendations made by this team will be made known to a group of internal referents of the social network with extensive knowledge on the subject and in the cultural context in which the tweet was circumscribed so that, after they are issued, the Trust & Safety leaders finally make the decision whether to apply the notice or delete the tweet.

However, this mode of personalized review would not appear to be applied by Twitter in a uniform manner for situations of public interest. An example of this can be seen when in October 2020 Twitter prevented users from sharing an article in the New York Post newspaper linked to presidential candidate Joe Biden and his eventual contacts with a Ukrainian businessman. Why did it stop it? The notice stated the following rationale: "Your Tweet could not be sent because Twitter or our partners identified this link as potentially harmful" (Cox, 2020). No additional information was provided about whether a team of "Trust & Safety" professionals could have intervened in such a decision, as it would appear that they did when referring to their "public interest" policies in the notice inserted in President Trump's tweet.

Going to the analysis of the general policies of this platform, it is worth referring to the case of dissemination of multimedia content. Thus, Twitter anticipates that it will focus its attention on content that is significantly altered or falsified with the deliberate intention of deceiving. However, it does not explain how it would arrive at such a conclusion, that is, how it would determine that certain audiovisual content has been altered or falsified. To this end, Twitter alerts that it has the power to apply its own technology — it does not specify or report it— or to collect a complaint through its collaborators or external partners. Only in those cases in which it is impossible to determine with certainty whether what is exposed in multimedia content was modified or is a copy, it may be — it does not guarantee— that it does not take any measure to restrict or reference it (Twitter, 2020).

Likewise, and always in relation to the dissemination of multimedia content on which it fails to provide details about how it concludes that it could lead to confusion or suggest a malicious intention to deceive, it reports that it analyzes the context of the tweet to determine if the content is modified or falsified, although it does not specify whether professionals are involved for this purpose, as expressly indicated by the content of public interest. Thus, the lack of precision inclines us to infer that Twitter would use AI for the purposes of reviewing:

- The text of the tweet that is attached to or included in the multimedia element.
- The metadata associated with the multimedia element.
- The profile information of the account that tweets the multimedia element.

- The websites linked in the tweet in the profile of the account that tweets the multimedia element.

In this sense, we observe that the automated measures that Twitter adopts in the face of content that the same platform qualifies as false or altered, since it prevents it from being shared on Twitter and, consequently, it could be deleted, at the same time, that the account from which the aforementioned content emanates may be permanently suspended.

Facebook

Facebook reports on its platform that its strategy to stop misinformation consists of three specific actions:

- Remove accounts and content that violate our community rules or advertising policies.
- Reduce the distribution of fake news and inauthentic content such as "click bait".
- Inform people by providing more context to the publications they view.

This three-pronged action would tend to weed out the "bad actors" who frequently spread fake stories and, it says, would dramatically decrease the reach of those stories by helping people stay informed without stifling public discourse.

It stands out that for this work it uses machine learning to help its teams detect fraud, enforce its anti-spam policies and block millions of fake accounts every day when they try to register (Facebook, 2020).

It reports that it takes "action" —although it does not explain what would it consist of— against entire pages and websites that repeatedly share fake news, which would reduce its overall news distribution. They highlight that because Facebook has no intention of making money from misinformation or helping its creators make a profit, those publishers are prevented from running ads and using its monetization features like Instant Articles.

It also highlights that part of its strategy to combat misinformation is to partner with various countries with third-party data verifiers to review and rate the accuracy of articles and posts on Facebook. These fact-checkers

would be independent since, as it notes, they are certified through the non-partisan International Data Verification Network. Thus, when those organizations rate some content as fake, Facebook rates that story significantly lower in the News Feed. In this way, they claim that this reduces future views by more than 80% (Lions, 2018). In line with what Agustina del Campo observes, it is noted that Facebook has gone from "a system that depended almost entirely on its users for complaints of content that violated its rules, to a system of activation and proactive 'enforcement' of its terms and conditions of service". Regarding the so-called infodemic, this change implied that this social network automates the moderation of content that would be "possibly" false, and then directly forward that same content to other users or to so-called "verifiers", even before someone uploads an internal complaint about said content (Del Campo, 2020).

To close, the following graph prepared by Facebook is illustrative as a global sample of content removal from 2013 to 2019 in the last six years (Facebook Transparency, 2019):

Conclusion: The necessary human supervision as a rule and not as an exception in the final decision to remove online content

Throughout this paper, we have analyzed and briefly described the international legal framework related to the protection of freedom of thought

and expression as a human right that is exercised regardless of what medium or platform it is done through. With the exponential growth of the various online platforms and the volume of data that grows second by second due to the interaction of users, we have realized how international organizations highlight the use of AI in distribution and, also, in automated restriction of online content.

We have also exposed that, in general terms, the inhuman use of predictive algorithms in regard to the precise and automatic removal of online content violates international standards of freedom of expression, such as the prohibition of prior censorship.

With only the use of AI in the decision to remove online content, the first factual situation that contradicts that standard of prohibition of censorship is visualized: that a series of instructions programmed by humans with predictive functions and with the ability to read natural language, It simply has what information we receive through the social networks with which we interact on a daily basis. This situation openly contradicts the standard established by article 13 of the American Convention on Human Rights, Article 19 of the ICCPR, art. 10 of the European Treaty on Human Rights, among others, because, in general, "the restriction of free expression is only admissible through the enactment of a necessary law, which pursues a legitimate purpose and is proportional to the right that it is trying to protect."

With a view to ensuring that this standard does not become a dead letter and at the same time without affecting the use of AI in the moderation of online content, in order to achieve a balance between the two, we postulate that human supervision is transcendental for the necessary review of any automated decision to remove content. The illustrative examples referenced throughout this work allow us to infer that, although human supervision of the decision adopted by the AI would not be able to absolutely prevent online censorship, it is possible to anticipate that it would contribute to remedying the serious defects of AI which cannot —yet— interpret contexts, linguistic terms, irony, satirical humor, artistic images of nudity, etc.

To this end, it is imperative that international organizations such as the UN, OAS, European Commission, etc. continue the global study of this problem and, from there, persist in urging States to:

Make their internal laws compatible while respecting international standards of freedom of expression. Although each country is sovereign and has the power to regulate speech on Internet platforms more directly, there are

specific cases such as the one that happens in Germany where the NetzDG Law has been in force since 2018. This law requires social networks to quickly remove illegal speech, with a specific focus on hate speech and hate crimes, otherwise, they should pay fines of thousands of euros. The laudable end that the issuance of this regulation could have has to confront with an undeniable fact: that an alleged rapid elimination of supposedly illegal online content ignores relevant constitutional guarantees such as due process and the right to defense when the decision is delegated to private platforms to confirm what content deserves or does not to remain online when, where appropriate, such a resolution would correspond to be adopted by a natural Judge, at least with regard to those democratic States.

Also, it is necessary that they develop public policies consisting of implementing protective legislation for the working conditions (psychophysical aspects in particular) of the dependent personnel who carry out supervision tasks of all automated decisions to remove online content under the orders of platforms.

Likewise, it would be pertinent to require companies that their terms and conditions are clearly explained and consistent with the human rights standards established for freedom of expression.

On the other hand, it would also be convenient for those companies that operate physically or virtually in their territories to also adopt internal policies of transparency and accountability about how AI operates in the dissemination and removal of each online content that everyone receives when interacting with their platform. All this, of course, together with the necessary collaboration that these companies should provide in perfecting the current internal appeal mechanisms in the event of a possible automated and supervised decision that orders the blocking of an account or removal of online content (UN, 2018).

And finally, complementing the above, those same companies must carry out due diligence through impact assessments on human rights, that is, what their rules are, how they are applied and what measures they take to prevent them from being violated. While it is clear that the details of individual compliance actions should be kept private, transparency reports in turn provide essential information on how the company is addressing the challenges of the day (New America, 2020).

Bibliography

Álvaro S. (2014). El poder de los algoritmos: cómo el software formatea la cultura. *CCCBLAB. Investigación e Innovación en Cultura*. https://bit.ly/3tiAGrO

Corte Interamericana de Derechos Humanos (2009). Caso Tristán Donoso vs Panamá. Serie C No. 193.

Corte Interamericana de Derechos Humanos (2011). Caso Fontevecchia y D'amico Vs. Argentina. Serie C No. 238.

Corte Suprema de Justicia de la Nación Argentina (2013). Grupo Clarín S.A. y otros c/ Poder Ejecutivo Nacional y otros/ Acción meramente declarativa. Ver voto de la mayoría. Octubre 29.

Cox, J. (2020). Twitter Says It Blocked NY Post Hunter Biden Article Because It Contains Hacked Data. https://bit.ly/2MNSDxB

Daus, G. (2019). Entrevista: Walter Sosa Escudero y la big data: el experto ante el diluvio de datos. *Clarín*. https://bit.ly/39AzxnC

Del Campo, A. (2020). *¿La desinformación en democracia o la democracia de la desinformación?* Univ. de Palermo. Facultad de Derecho. Centro de Estudios en Libertad de Expresión y Acceso a Información. Septiembre.

European Commission (2019). Ethics guidelines for trust worthy AI. https://bit.ly/3alzqM1

Facebook (2020). Normas Comunitarias. Punto 21. https://bit.ly/3cwmyFm

Facebook Transparency (2019). Content Restrictions Based on Local Law. https://bit.ly/3ameWCW

Lions, T. (2018). Hard Questions: What's Facebook's Strategy for Stopping False News? https://bit.ly/3j6VGwZ

Marantz, A. (2020). Why Facebook Can't Fix Itself. *The New Yorker.* https://bit.ly/3r5oMQa

New America (2020). So What Should Companies Do? *New America*https://bit.ly/3je4DET

Organización de las Naciones Unidas, ONU (1966). Pacto Internacional de Derechos Civiles y Políticos. Diciembre 16.

Organización de las Naciones Unidas, ONU (2011). Comentario General n 34 del Comité de Derechos Humanos. Párrafo 9. Septiembre 12.

Organización de las Naciones Unidas, ONU (2018). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión. Abril 6.

Organización de las Naciones Unidas, ONU (2018). Promoción y protección del derecho a la libertad de opinión y expresión. Agosto 29.

Secretaría General de la Organización de los Estados Americanos (2017). Libertad de expresión: a 30 años de la Opinión Consultiva sobre la colegiación obligatoria de periodistas: Estudios sobre el derecho a la libertad de expresión en la doctrina del Sistema Interamericano de Derechos Humanos. Bogotá. Colombia. Noviembre.

Twitter (2020). Política relativa a los contenidos multimedia falsos y alterados. https://bit.ly/3raXE2f

Twitter (2020). Acerca de las excepciones de interés público en Twitter. https://bit.ly/36zKXGy

Submission date: 2020/10/31; Acceptance date: 2021/01/31;
Publishing date: 2021/03/01