ARTICLE INFO
Received: July 18, 2021 | Accepted: August 20, 2021 | Available online: September 1, 2021

CITATION

COPYRIGHT
Copyright © 2021 by author(s). Metaverse is published by Asia Pacific Academy of Science Pte. Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), permitting distribution and reproduction in any medium, provided the original work is cited.
subject to subsequent responsibilities that must be established by law, respecting its necessity, 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, whether “by law or by any other power”[2].

The limits and scope of the exercise of this right are the focus of analysis and complementary interpretation by the main international organizations.

Without going any further, it is sufficient to cite the validity of OC 5/85 issued by the Inter-American Court, which states that freedom of expression is not exhausted in its individual aspect, but also includes the collective dimension, stressing that free thought and its dissemination are inseparable, so that a prior limitation state or private to either of them, would be incompatible with the Inter-American standards that protect this right[3].

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

Indeed, freedom of expression, in its collective aspect, seeks to ensure public debate in order to provide opportunities for expression to the various strata of society so that all people can make their ideas and opinions known, thus preventing one group or group from predominating over others[4].

These intercontinental standards were agreed upon throughout almost the entire twentieth century and early twenty-first century. With the irruption of social networks and other online intermediaries came the moment when practically half of the world’s population exercises the triple action of disseminating, researching and spreading information through the Internet by means of the main platforms that employ “Artificial Intelligence” (hereinafter AI) with which we interact on a daily basis.

But what is AI? For the time being, there is no consensus on a single definition worldwide. On this occasion, we are inclined to cite the one provided by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression when he says that “it is a ‘constellation’ of processes and technologies that enable computers to complement or replace specific tasks that would otherwise be performed by humans, such as decision-making and problem-solving”[5].

The Rapporteur ship adds that at the basis of AI are “algorithms” which are computer codes designed and written by humans. Any kind of data processed by an algorithm is translated and yields a given result such as inferences, suggestions or predictions. Thus, the infinite amount of data that a person generates per second when interacting on the network leads to the necessary development of AI, given the material impossibility for a person to do it alone in a short time and efficiently.

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 Internet browsing on Google more than three million searches are carried out, on Facebook more than thirty million messages are sent and more than two million videos are watched, on Twitter more than four hundred and fifty thousand tweets are published, on Instagram more than forty-six thousand photos are posted, on YouTube more than four million hours of videos are uploaded and almost twice as many 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 made up of a four-digit number, but rather the evolution has led us to the term bytes, which has twenty-seven digits\(^6\), i.e., the information is six hundred times larger.

This immense flow of data that we generate through social networks and intermediaries makes up an eccentric virtual place where we converge with other people, just as we do with other people, just as we do with other people, just as we do with other people.

We are also interacting with other people, as well as with “bots” and other automated systems based on AI.

While 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 online content by deciding, in an automated way, which content will remain and which will be removed according to its “terms and conditions”.

A first approximation of the use of AI by platforms could be considered adequate for the removal of violent, disinformation or hateful content, for example.

However, anticipating the development and conclusion of our proposal, we consider that without human moderation that adequately contextualizes and translates online expression, there may be a serious risk that platforms may give prevalence to AI as an automated moderator of online content, thus failing to comply with the international standards referred to above in relation to the human right to receive, investigate and disseminate would be limited no longer by a necessary law, with a legitimate and proportional purpose but by a de facto “constellation” made up of inhuman algorithms.

Therefore, our work will begin by explaining how international bodies conceptualize and diagnose the use of AI in the automated moderation of online content, as well as pointing out its main advantages and disadvantages. We will then analyze the implications of AI on the exercise of freedom of expression through platforms such as Facebook and Twitter with a view to assessing whether its implementation has recently contributed to restricting expression. Finally, and based on the factual and legal framework analyzed, we postulate the necessary intervention of human supervision when AI suggests the removal of online content of public interest.

2. AI and automated moderation of online content: Reasons for its implementation and the necessary human oversight

AI is usually considered as a set of automatic and impartial technological systems aimed at facilitating the effectiveness of content moderation in order to mitigate possible hate speech, discriminatory, terrorist, etc., and thus improve the experience of its users and the construction of citizenship.

However, the UN has pointed out that, in the field of content moderation, although AI has its positive aspects, it also has negative ones.

Among the benefits of the use of 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 drawback the limitation that each person can 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 position. Political, religious or social. In this way, that segmentation of containment that appears to be very useful and effective, could at the same time reinforce individual beliefs and lead to
Artificial intelligence, algorithms and freedom of expression

the exacerbation of violent content or disinformation with the sole purpose of maintaining the user’s online participation\[7\].

Sandra Álvaro explains that algorithms are already part of our daily life, giving as an example Facebook, which has an algorithm called Edge rank that analyzes our browsing data the “likes” we post, the friends we have and the comments we make and thus profiles us in order to show us those stories we like and hide those that bore us and show us new friends that match our profile and ideology\[8\].

This situation, which generates a sort of information bubble, has aroused the interest of the European Union as it warns that humans interacting with AI systems must be able to maintain full and effective self-determination over themselves and be able to participate in the democratic process. It therefore urges that AI systems should not coerce, manipulate, infer or unjustifiably group human beings.

In the view of the European body, AI should then be designed to augment, complement and enhance human cognitive, social and cultural abilities, thus following human-centered design principles\[9\].

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

This is because the use of AI in the automated moderation of content may affect the exercise of freedom of expression since, for the moment, its limitations include the impossibility of assessing the context, idiomatic uses and cultural aspects of human beings.

Although in recent times AI has improved exponentially in Natural Language Processing (NLP), it has not yet achieved a level of development that allows it to understand all the linguistic and cultural nuances through 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 against hate groups or efforts to challenge or denounce racist, homophobic or xenophobic language.

This would demonstrate that, at this stage of AI development, we are still faced with 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 libelous content, hate speech or eventual disinformation online. For this reason, Internet companies have urged users to refine the content observed with different contextual elements, although, it should be clarified, the feasibility and effectiveness of these guidelines are not clear\[7\].

In this regard, the UN Human Rights Committee understands that, unlike people, algorithms lack corpus and mind, i.e., they are not yet able to understand when an expression is ironic, or is a parody or to confirm with precision whether a given manifestation can be qualified as praise for “terrorism”. Therefore, the self-matrixing of its mathematical operability tends to opt for a quick result consisting of limiting or removing certain expression without considering that this results in considerably affecting the human right to receive, investigate and disseminate\[7\].

Similarly, the use of AI when uploading files on the web, in order to protect the intellectual property rights of both videos has generated doubts due to the large number of blockages that occur,
which, added to the possible filtering of content linked to terrorism or other extreme positions, can result in the opposite, i.e., instead of protecting rights, totalitarian regimes can be established, by applying an automated prior censorship.

Indeed, while the use of cryptographic matching algorithms to detect child sexual abuse images is extremely useful, their application to “extremist” content - which usually requires context assessment - is difficult without clear standards defining what “extremism” is[7].

In this sense, the UN understands that platforms should transparently explain how they use AI, revealing in detail with aggregated data illustrating examples of real cases or hypothetical cases in order to clarify how it is interpreted and the application of concrete norms[7].

Likewise, being the responsibility of companies to prevent and eventually diminish the negative effects on human rights with the use of AI, it is clear that part of their transparency policy should consist of starting by recognizing the important limitations that self-nuanced content moderation suffers from, such as those difficulties already referred to on the interpretation of context as well as the wide variation of idiomatic nuances and meaning and linguistic and cultural particularities. That is why, at the very least, current and future technology for dealing with large-scale data-related issues should be subject to rigorous auditing and, of course, have input from civil society to enrich the analysis.

To conclude this section, we would like to refer to a final aspect related to our proposal to ensure human oversight of the possible automated removal of online content.

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

To this end, the UN affirms that it is essential that adequate protection be provided to the working conditions in which they perform tasks as they must be compatible with human rights standards applicable to labor rights[7].

Such a postulation has its basis, for example, in a concrete case of “labor precariousness” of moderators working for Facebook.

Indeed, this company in 2015 had less than four thousand five hundred people as moderators of audiovisual content, but, on the occasion of COVID-19, it had to expand the workforce by hiring about fifteen thousand moderators, most of whom are under the modality of sub-con- treaties in various cities around the world (Dublin, Berlin, Manila).

The New Yorker magazine reports that moderators often work at odd hours because of the different time zones around the world, in addition to the lack of sleep and the strong psychological impact of absorbing everything they see on their screens without a standardized “protocol” to indicate which content should remain online and which should not.

As a result, in May 2020, thousands of moderators joined a class action lawsuit against Facebook alleging psychological disorders and, as a result, agreed with the company to a $52,000,000 settlement[10].

The human supervision that we postulate for the review of the automated content decision while it would not absolutely prevent online censorship, it is possible to anticipate that it would contribute to make up for the serious shortcomings of AI that does not yet manage to interpret contexts, linguistic terms, irony, satirical humor, artistic images of nudity, etc.

The following are some specific cases that, according to our position, accompany this proposal to implement human supervision in the face of disinformation and automated removal of content.
3. How Twitter and Facebook operate

3.1. Twitter rules

The social network Twitter has a set of rules entitled “General Policies and Guidelines” that must be followed in order to use the platform. One part of these guidelines relates, as far as our analysis is concerned, to online content related to topics of public interest.

Although this social network anticipates that it takes various kinds of measures regarding tweets that do not comply with its rules, at the same time it recognizes that on certain occasions - without specifying which ones, at least by way of example - it keeps certain tweets online that may be of use to society and that would otherwise be deleted. When would a tweet be considered of public interest? The platform reports that it qualifies as such when it is presented as “a direct contribution to the understanding or discussion of an issue of concern to the public at large”[11].

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 discuss 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, namely:

- 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 a local, national or supranational government or legislative branch: i) current holders of an elected or appointed leadership position in a government or legislative body; or candidates or nominees for political office.

It may happen, however, that a public official posts a tweet in violation of Twitter’s terms and conditions. In this case, as an exception, the platform informs that he/she may choose to keep the tweet, which would otherwise be deleted. To this end, it inserts a warning behind the tweet to contextualize the breach of the rules and allow people to view it, if they wish to do so.

Using AI, it states that, by placing this notice, the possibility of interacting with the tweet by “liking”, “Retweeting” or sharing it on the same social network is also being reduced, so that the Twitter algorithm avoids recommending it. It is thus noticed that through these actions it would be attempted to restrict the reach of the toity, at the same time, guaranteeing the public the possibility of visualizing it and debating on the topic in question.

As a first observation, we would like to highlight the limited and restrictive framework implemented by Twitter when it requires an account to have 100,000 followers in order to be included within the conditions of the public interest standard. Quantitative measurement based only on the number of followers which could well be made up mostly of bot accounts. We believe that it would go against a qualitative analysis of the discourse in question, as to define whether or not a discourse is of public interest. This is because human supervision should be used, following jurisprudential standards such as that of the Inter-American Court of Human Rights, which defines public interest as those opinions or information on matters in which society has a legitimate interest in being kept informed about the functioning of the State or general rights and interests[12]. Returning then to the analysis of the measures implemented by Twitter on this point, an example of this can be seen in one of the many
tweets issued on August 23, 2020 by President Donald Trump on the presidential election contest.

As can be seen, in that tweet President Trump alluded to the possible electoral fraud that could be committed through the citizen mail in ballot system. In that case, Twitter inserted a warning in the tweet that warned the public about the non-compliance with the rules related to the integrity of the civic electoral processes, although it was also decided that the tweet would remain accessible. For more information, a link was attached to refer the user to the reading of the general policies and 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 professional experts in various fields, will implement a second analysis, in order to analyze the tweet and give an opinion on whether or not to retain the visibility of the tweet based on the 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 their opinion, the Trust & Safety leaders will finally decide whether to apply the warning or delete the tweet.

However, it seems that this personalized review modality would not be applied uniformly by Twitter for situations of public interest. An example of this can be seen when in October 2020 Twitter prevented users from sharing a New York Post article linked to presidential candidate Joe Biden and his possible contacts with a Ukrainian businessman. Why did it prevent it? The notice stated the following rationale: “Your Tweet could not be sent because Twitter or our partners identified this link as potentially harmful”[13]. No additional information was provided as to whether a team of Trust & Safety professionals may have been involved in such a decision, as it appears, they did by referring to their “public interest” policies in the notice inserted in President Trump’s tweet.

Moving on to the analysis of this platform’s policies in general, it is worth referring to the case of the dissemination of multimedia content. Twitter thus announces that it will focus its attention on content that is altered or falsified with the deliberate intent to mislead. However, it does not explain how it would arrive at such a conclusion, i.e., how it would determine that a certain audiovisual content has been altered or falsified. To this end, Twitter warns that it has the power to apply its own technology - it does not specify or report it - or to collect complaints 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 has been modified or is a copy, it may be - it does not assure - that it does not take any measures to restrict or reference it[14].

Likewise, and always in relation to the dissemination of multimedia content about which it omits to provide details about how it concludes that it could give rise to confusion or suggest a malicious intent to deceive, it reports that it analyzes the context of the tweet to determine whether it warns that the content is modified or falsified, although it does not specify whether professionals intervene for that purpose, as it expressly indicates for public interest content. Thus, the lack of precision leads us to infer that Twitter would use AI for review purposes:

- The text of the tweet attached to or included in the multimedia element.
- The metadata associated with the media item.
- The profile information of the account broadcasting the media item.
- The websites linked in the tweet in the profile of the account that disseminates the multimedia element.

In this sense, we note that the automated
measures that Twitter adopts when faced with content that the same platform qualifies as false or altered, as it prevents it from being shared on Twitter and, consequently, it could be deleted at the same time that the account from which the referred content emanates could be permanently suspended.

3.2. Facebook

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

- Remove accounts and content that violate its community standards or advertising policies.
- Reduce the distribution of fake news and inauthentic content such as “bait titles”.
- Inform people by providing more context to the posts they view.

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

It also notes that 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 sign up[^15].

It reports that it is taking “measures” - although it does not explain what they consist of - against entire pages and websites that repeatedly share fake news, which would reduce its overall distribution of news. They stress that since Facebook has no intention of making money from misinformation or helping those who generate it to make a profit, these publishers are prevented from running ads and using its monetization features such as Instant Articles.

It also notes that part of its strategy to combat disinformation is to partner in several countries with third-party fact-checkers to review and rate the accuracy of articles and posts on Facebook. These fact-checkers would be independent since, he notes, they are certified through the nonpartisan International Fact-Checking Network. Thus, when these organizations rate a story as false, Facebook significantly downgrades that story in the News Feed. Thus, they claim that this reduces future views by more than 80%[^16]. 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 reports of content that violated its rules, to a system of activation and proactive ‘enforcement’ of its terms and conditions of service. With regard to the so-called endemics, this change implied that this social network...
automates the moderation of content that is “possibly” false, and then directly forwards the same content to other users or to the so-called “verifiers”, even before someone has filed an internal complaint about such content\(^{[17]}\).

In closing, it is illustrative the following graph prepared by Facebook as a global sample of content removal from 2013 to 2019 in the last six years\(^{[18]}\) in Figure 1.

4. Conclusions: The necessary human supervision as a rule and not as an exception in the final decision of online content removal

Throughout this paper we have analyzed and succinctly described the international legal framework related to the protection of freedom of thought and expression as a human right that is exercised no matter through which medium or platform it is done. With the exponential growth of the various online platforms and the volume of data that grows second by second thanks to the interaction of users, we have given an account of how international organizations highlight the use and employment of AI in the distribution and in the automated restriction of online content.

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

With the sole 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 natural language reading capabilities, simply decides what information we receive through the social networks with which we interact daily. This situation openly contradicts the standard provided by Article 13 of the American Convention on Human Rights, Article 19 of the ICCPR, Article 10 of the European Human Rights Treaty, among others, inasmuch as in general “restrictions on free expression are only admissible through the enactment of a necessary law, which pursues a legitimate aim and is proportionate to the right it seeks to protect”.

In order to ensure 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 essential for the necessary review of all automated content removal decisions. The illustrative examples referenced throughout this paper allow us to infer that, although human supervision of the decision taken by the AI would not be able to absolutely prevent online censorship, it is possible to anticipate that it would contribute to make up for the serious shortcomings of the latter, which still fails to interpret contexts, linguistic terms, irony, satiric 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:

Compatibilized their domestic legislation while respecting international standards of freedom of expression. While each country is sovereign and has the power to regulate speech on Internet platforms more directly, there are specific cases such as what is happening in Germany where the NetzDG Law has been in force since 2018. This law requires social networks to promptly remove illegal speech, with a specific focus on hate speech and hate crimes, otherwise they should pay fines of thousands of euros. The laudable purpose that may have been served by the enactment of such regulation confronts an undeniable fact: that an allegedly illegal online content fast removal ignores relevant constitutional guarantees such as due process and right of defense by delegating to private platforms the decision to confirm which
content deserves or not to remain online when, in its case, such resolution should be adopted by a natural judge, at least as far as those democratic States are concerned.

It is also necessary that they develop public policies consisting in implementing legislation protecting the working conditions (psychophysical aspects in particular) of the dependent personnel who perform supervisory tasks of any automated decision of removal of online content under the orders of platforms.

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

On the other hand, it would also be appropriate to require 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 received by any person interacting with their platform. All of this, of course, together with the necessary collaboration that these companies should provide to improve the current internal appeal mechanisms in the event of an automated and supervised decision to block an account or remove online content[7].

And finally, complementing the above, these same companies should conduct due diligence through human rights impact assessments, i.e., what are their rules, how they are applied and what measures are taken to prevent them from being violated. While 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[19].

Conflict of interest

The authors declare no conflict of interest.

References

12. Inter-American Court of Human Rights. Fontevecchia y D’amico Case Vs. Argentina Series C (Spanish) [Case of Fontevecchia and D’Amico v. Argentina Series C] [Internet]. 2011. Available from: http://corteidh.or.cr/docs/casos/articulos/seri


17. Del Campo A. ¿La desinformación en democracia o la democracia de la desinformación? (Spanish) [Disinformation in democracy or the democracy of disinformation?]. Buenos Aires: Universidad de Palermo; 2020.
