DOI: 10.1590/s0103-4014.202438111.020

Explainable AI to Mitigate the Lack of Transparency and Legitimacy in Internet Moderation

Thomas Palmeira Ferraz [I]
Caio Henrique Dias Duarte [II]
Maria Fernanda Ribeiro [III]
Gabriel Goes Braga Takayanagi [IV]
ALEXANDRE ALCOFORADO [V]
Roseli de Deus Lopes [VI]
Mart Susi [VII]

Introduction

Historically, societies have always established norms of coexistence that impose limits on individual rights. This was the case with the fundamental right to freedom of expression, which in modern democracies had clearly established limits and mechanisms for punishing those who did not respect it. However, with the advent of social media, public debate has shifted to the internet. If, on one hand, this makes it easy for people to participate, on the other, it makes it impossible for the same conventional means of moderating public debate to be applied. To address this issue, digital platforms have created staffs of human moderators who work to handle complaints. In recent years these groups have been replaced by the integrated use of several models based on Artificial Intelligence (AI) that automatically act in content moderation, in many cases even before there is a complaint from any party. However, the massive use of AI in this role of judge and the fact that the monopoly of this moderation is in charge of private entities raise a series of ethical and legal questions that will be explored in this work.

From a legal point of view, there is an impasse over who can moderate the content. Ideally, it would be one that could follow the principles of legitimacy, transparency, control, and enforcement capability (Sander, 2020). Although the State is endowed with democratic legitimacy, transparent public regulation, and local action power, it has limited resources and little capability to respond quickly to demands for moderation. Added to this is the fact that it does not have jurisdiction over where the content is stored (the servers and data centers), usually outside its borders, being unable to carry out this moderation at the end. Digital platforms, on the other hand, are indeed more capable of identifying users, have easy access to content, and know the structure of their portal, everything needed for enforcing decisions, but in practice, they do not have a clear regulation and their legitimacy as a public censor is questionable. This scenario places us in a dilemma, in which the State cannot act alone, and private entities, if they act alone, can be held responsible for actions that appear to be excesses or omissions.

The use of artificial intelligence models in the role of judges, in turn, also presents important ethical contradictions to be addressed (Nahmias; Perel, 2020). In addition to the philosophical issue of machines making decisions about human rights, the opaque (or "black box") nature of more complex, widely used models makes it difficult to perceive unconscious biases that can harm minority social groups compared to others, a phenomenon called algorithmic discrimination. In the context of content moderation on the internet, these unconscious biases may arise from local and cultural issues, political positioning, race, sexual orientation, among others, considering that each group may have a particular vocabulary (Oliva; Antonialli; Gomes, 2021). Tolerating that groups have less freedom of expression than others undermines the foundations of a full Democratic Rule of Law.

These questions establish, for contemporary society, a dilemma between moderating and not moderating: keeping freedom of expression inviolable, even when it is abused, or combating virtual abuse and potentially harmful content, running the risk of inadvertently suppressing a fundamental right? The evolution of social networks makes it impossible for moderation to be made entirely by humans, and even if it were, humans are also subject to bias, although they are easier to measure and mitigation methods already exist. Therefore, the use of AI in moderation was, in the last decade, the choice for the "lesser evil". Now that the use of AI is fully widespread across platforms, it is essential to debate it in order to refine this type of moderation and find a balance point. This type of discussion has been made about the use of artificial intelligence in all aspects of human life and led the European Union to introduce the right to explain algorithmic decisions in its legal framework of technology (GDPR) in 2016 (Goodman; Flaxman, 2017).

In this panorama, the field of *Explainable AI* (XAI) has emerged in Computer Science research, in which tools are researched and developed that make it possible to interpret the decision process of existing models, as well as developing models designed to be interpretable by humans (Adadi; Berrada, 2018; Felzmann et al., 2020). This is an area that has advanced a lot, including achieving performance close to opaque models for various (Arrieta et al., 2020) tasks. In the context of content moderation, this category of AI enables: (*i*) the development of models that are already built to follow a moderation pattern defined by society and be able to explain their decisions based on this pattern; and (*ii*) building models that can do parallel audit on existing black box models based on these same criteria, explaining why the black box model is making its decisions and, thus, giving greater transparency to the process.

There are several works in the literature that address the aspects discussed so far. Susi (2019) was able to formulate a mathematical instrument based on transparent criteria, towards a pattern of moderation of invasion of privacy that respects fundamental human rights. Already Reis et al. (2019) were able to define metrics to audit the XGBoost model for false news detection using *Explainable AI*. On the other hand, Mohseni et al. (2021) managed to produce evidence that transparent criteria and decisions explained to the user have the potential to reduce the recurrence of false news sharing. In this work, we approach aspects similar to those of these studies, from an interdisciplinary perspective, but with a focus on defining more clearly the aspects of the problem and the actors involved in it, and building a paradigm of transparent moderation that solves it.

Objectives and methodology

The objective of this work is to analyze the use of artificial intelligence in the context of content moderation on the internet. We bring a holistic but accessible view of the current panorama of this topic, indicating opportunities for improvement, especially with the evolution of technological maturity in the field of *Explainable AI*. With this, we propose a new moderation pattern that is consistent with current demands.

To this end, we take three points of study:

- **The current paradigm of moderation**, considering its different decision levels, the role played by digital platforms and the State in the process, as well as the AI models used, their different levels of automation, the functions they perform in the context in which are applied, and their limitations;
- **The ethical and legal view** under the current framework for the moderation of online content, exploring how law and social sciences observe the current system, and defining the state of the art of the discussion on freedom of expression in the digital context, understanding which aspects must be met by a pattern of moderation to be adopted and what role each stakeholder should play;
- **The state of the art of technology**, analyzing the panorama of the *Explainable AI* and its technological maturity at present, understanding its operation, what is necessary for its construction and its limitations, bringing conclusions about how the advances in this area can be applied to solve the difficulties that are currently found in the area of moderation.

Considering these three pillars, we seek to summarize and relate what was studied by proposing a new paradigm of moderation that is fair, ethical, and transparent, as well as defining the role of the State and digital platforms in this context. It is also discussed which technological challenges will have to be faced for its implementation, as well as their potential benefits.

This work, in general lines, seeks to diagnose the deficiencies of the use of Artificial Intelligence in the automatic moderation of content on social networks and its threats from an ethical and legal point of view to fundamental human rights such as freedom of expression, and from that identify and unravel like *Explainable AI* can be useful to mitigate these negative effects. In the *Web of Science* database, there are about 4300 studies addressing XAI. On this same basis, there are approximately 2000 studies about automatic content moderation, however, only 21 of them mention *Explainable AI*, which demonstrates the need for work that considers both concepts together. For this purpose, we use an interdisciplinary approach between sociology, computing, and law, combining technical and mathematical analysis of data, with a review of studies that address society and law.

First, we carry out an exploratory review of the literature and of the materials and data made available by the Meta (2021) and Google (2021) platforms themselves to characterize the way in which large platforms currently perform moderation, including seeking ways to measure their effectiveness. Then, we deepen the questions regarding the use of this technology from an ethical and legal point of view.

From an ethical point of view, we review works in Sociology, Social Computing, and Statistics that clarify which features of the "black box" model can result in inequality and injustice in the automatic decision-making process in content moderation. With that, we get the gaps and ethical challenges motivated by the use of this type of technology.

From a pragmatic point of view of Law, we carry out a comparative analysis of the current scenario regarding compliance with the principles of legitimacy, transparency, control and enforcement capability, and compliance with the legal regulations in force in the European Union (GDPR) and in Brazil (Marco Civil Internet and LGPD). This gives us an overview that allows us to diagnose the objectives set out in the regulation and legal doctrine that are not met by the moderation paradigm currently adopted, with special attention to the legitimacy to moderate the accessibility of the moderation criteria and public control over the process.

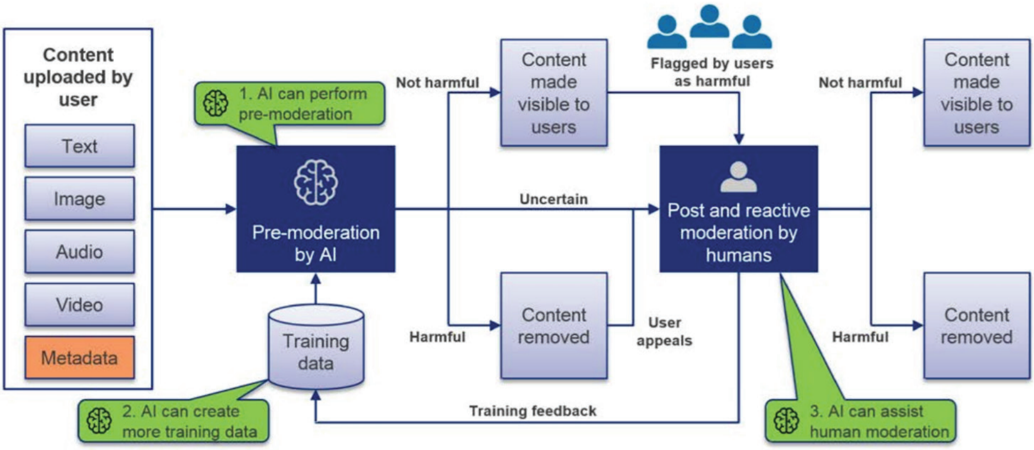
With the panorama well designed, we seek to define *Explainable AI* and unravel which of its characteristics can mitigate the harmful effects of AI use under the two aforementioned aspects. We seek to demonstrate the technological feasibility of implementing models of this type, as this is a recurrent question. Next, we outline what roles the State and digital platforms must play in order to make effective use of this new technology, creating a new paradigm of automatic content moderation in which transparency is at the center of decision-making.

Finally, we discuss advances that have already been made toward establishing this new paradigm, seeking similar practical experiences. We bring evidence that the use of these new technologies has great potential to not only cover the gaps presented in this study but also make content moderation more effective when society is introduced in the process.

Current Content Moderation Paradigm

Mainly due to societal and governmental pressure across several countries, major content providers started to develop and implement a set of standards to deal with content published on the platform that was not in accordance with local regulations or their vision about what should be the digital environment – like the *Community Standards* (for Facebook) and the *Community Guidelines & Policies* (for YouTube) (Estarque and Achergas 2021). This document usually contains the main guidelines on content and actions that are allowed or not on the social network, such as violence, harassment, hate speech, false news, nudity, and terrorism. Klonick (2017) reports, however, that the emergence of these is the result of recent efforts, and in the past, policies were based on generic guidelines. There are also recent efforts to create independent oversight committees (Klonick, 2019).

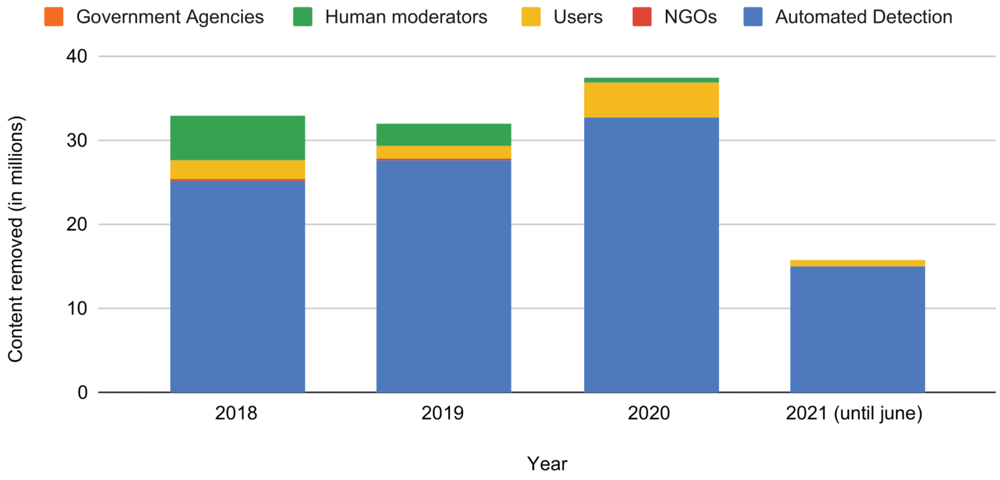
Content moderation leverages the recent revolution that has been taking place in Artificial Intelligence, with the emergence of deep learning models based on neural networks (LeCun; Bengio; Hinton, 2015) and, more recently, pre-trained language models (Devlin et al., 2019; Brown et al., 2020; Alcoforado et al., 2022), to make comply with Community Standards. Online content moderation can be implemented in several ways, but it usually adopts one of the following approaches or both (Winchcomb, 2019; Jiang; Robertson; Wilson, 2020): (i) *Pre-moderation*, when the uploaded content is moderated prior to publication, typically using AI-based systems; (ii) *Post-moderation* (or reactive-moderation), when content is moderated after its publication and it is flagged by users or AI-based systems as harmful, or when the content was previously removed but requires a second review upon appeal. Figure 1 illustrates this general case of moderation.


Content uploaded by user
Text
Image
Audio
Video
Metadata
1. AI can perform pre-moderation
Pre-moderation by AI
Training data
2. AI can create more training data
Not harmful
Uncertain
Harmful
Content made visible to users
Content removed
Flagged by users as harmful
User appeals
Post and reactive moderation by humans
Training feedback
Not harmful
Harmful
Content made visible to users
Content removed
3. AI can assist human moderation

Source: Winchcomb (2019).

Figure 1 – General Model for Content Moderation on Social Media.

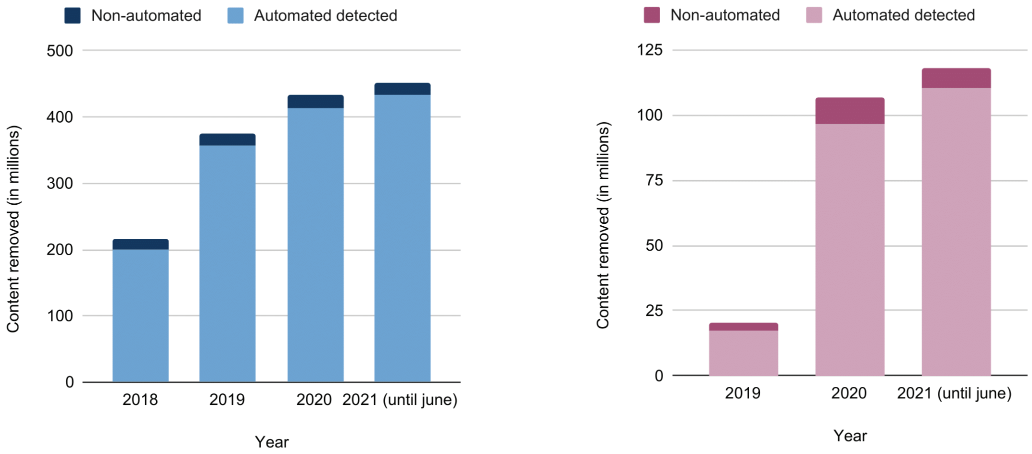
However, detecting toxic or harmful content can be quite challenging as the content can appear in many different modalities (eg. audio, image, videos, GIFs, text, and multimodal combinations), in different formats (eg. memes, deep fakes) (Winchcomb, 2019). Furthermore, some may be live streamings, which require real-time action. Still, others may depend on the context in which they were produced to be considered toxic or harmful. In addition, the language of the internet can evolve over time, and even users can learn techniques to circumvent content moderation using proprietary language encoding intentionally obfuscating certain words by misspelling, or leetspeak (informal internet language in which letters are replaced by numbers or symbols, eg. “f@ggot”, “ph*ck”, “wetback”) (Tan et al., 2020). If this difficulty were not enough, there is a broad range of content that is potentially harmful at different levels, including but not limited to: child abuse material, violent and extreme content, fake news, hate speech, false health-related claims, sexual content, cruel and insensitive material, and spam content. In practice, the success of using Artificial Intelligence for content moderation depends on the development of several expert systems for each category, working in harmony.



Source: Own Elaboration based on Data from Google Transparency Report (Google, 2021).

Figure 2 – Number of videos (in millions) removed for the first time on YouTube (not considering appeal) by year and by the author of removal. In blue, automated detection is AI removals. On the other colors, removals are made by humans: Government Agencies, Human moderators (who work for Google), User complaints, and NGOs (YouTube partners).

As shown in Figure 1, content moderation is typically implemented as an AI-human hybrid process, and may have varying degrees of automation. At a lower level of automation, pre-moderation by AI can only flag potential candidates to be removed. Moreover, AI can be implemented to synthesize training data to improve pre-moderation performance. In addition, AI can assist human moderators in post-moderation, reducing the effect of individual moderators on the final result. Analyzing the data from social networks YouTube, Facebook, and Instagram expressed in Figures 2 and 3, we can notice that there is an increase in the participation of Artificial Intelligence in the moderation process. In the case of YouTube, we can see that the percentage of content automatically removed by AI has increased from 76% in 2018 to almost 95% in 2021. The same phenomenon can be observed in the case of Facebook..



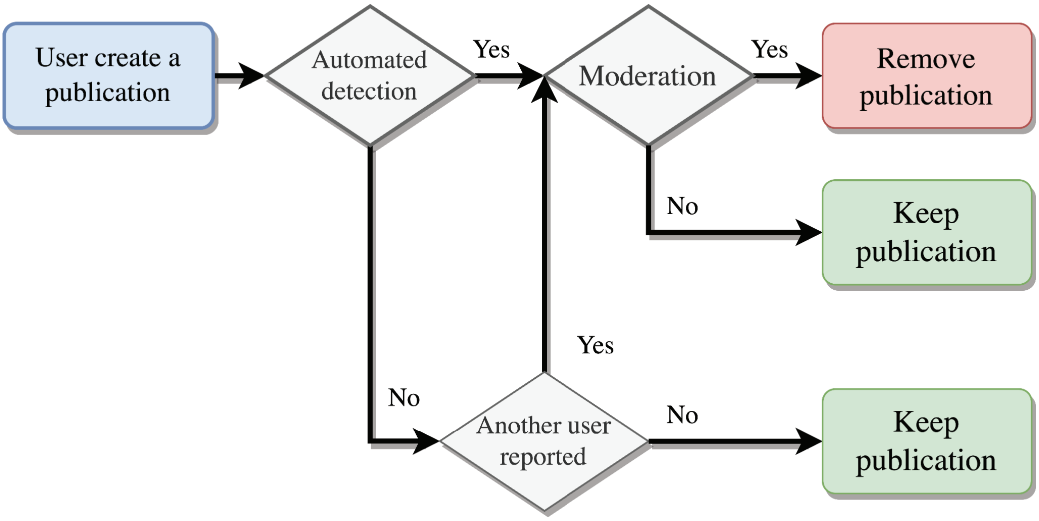
Facebook

Instagram

Source: Own Elaboration based on Data from Facebook Transparency Report (Meta, 2021).

Figure 3 – Number of content removed on Facebook and Instagram throughout the years, excluding spam remotion.

Lately, Facebook has implemented a mixed approach to moderating content,[1] presented in Figure 4. It uses human moderators, automated algorithms, and denunciations (reports) made by users to analyze content. It starts using automated algorithms to decide whether the material will be further analyzed – this verification happens using a human moderator. Facebook claims that most content that generates major concern, such as terrorism, child exploitation, or self-harm, are ranked to be first moderated by their analysts, while content such as spam are ranked last. If the content is not chosen by this moderator, then there is the option to be denounced by a user, and then again a human moderator will investigate it.



Source: Own Elaboration based on Facebook Transparency Report (Meta, 2021).

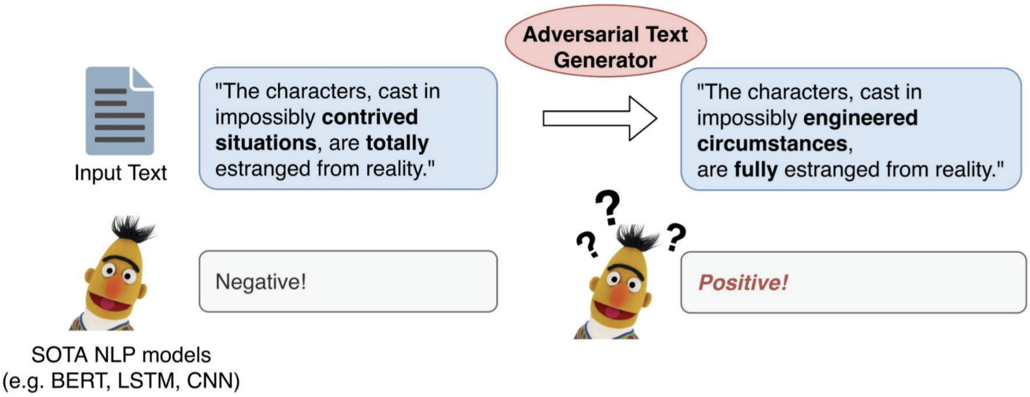
Figure 4 – New moderation procedure proposed by Meta in 2021.

Social and Ethical Impacts of Artificial Intelligence

Andrew Ng (2016) defines that we can probably automate using AI, either now or in the near future, any mental task that a regular human would take less than a second of thought. However, just like humans, machines are susceptible to errors. When we talk about content moderation on the internet, we are talking about an inherently imbalanced class problem, where certain categories are naturally less present in the real world. This is a classic problem in machine learning, which has been discussed by several authors, including Chawla, Japkowicz, and Kotcz (2004), Fernández et al. (2018), Ferraz et al. (2021) Krawczyk (2016), and He and Garcia (2009). In the case of social media, in general terms, there is more content to be kept than to be removed. For models to be able to learn in these scenarios, trade-off choices are made that can lead to performance loss. For instance, maximizing recall[2] can ensure that a fake news detection model classifies all fake news correctly but also may lead it to classify many true news stories as fake (type I error, or false positive). Added to this are the challenges already highlighted in dealing with different types of media (audio, video, image, text), in different scenarios. In this way, it is practically impossible to have a model with perfect performance, even in cases where they were trained to get it right (which are in their training data) (Duarte; Llanso; Loup, 2018). In many cases, models can outperform humans in speed and even standardization of judgment, but they will not always get it right. And what to do when they get it wrong? What is the cost of an error on these systems? This, of course, will always depend on the application, that is, on its purpose and how sensitive it can be.

As demonstrated by several studies, including the most recent one brought by Mehrabi et al. (2021), learning-based models are susceptible to different types of bias. The study reports that biased data generate biased models (Bias from Data to Algorithm). This bias can occur in several ways, including: Measurement Bias; Omitted Variable Bias; Representation Bias; Aggregation Bias – when false conclusions are drawn about individuals from observing the whole population; Linking bias – when network attributes obtained from user interactions misrepresent the true behavior of the users; between others. There are still biases that are added to the user by the algorithm (Bias from Algorithm to User), which are biases resulting from algorithmic outcomes and modulate user behavior as a consequence. This can occur from simple things like interface design choices, how information is presented, top-ranked results in a search, even popularity biases (fake reviews or social bots make an item popular and more exposed) (Ciampaglia et al., 2018), and biases arising from the choice of algorithm and optimization method (Danks; London, 2017). And finally, there are biases added by the user to the data (Bias from User to Data) when user behavior is affected by an algorithm, any biases present in those algorithms might introduce bias in the data generation process.

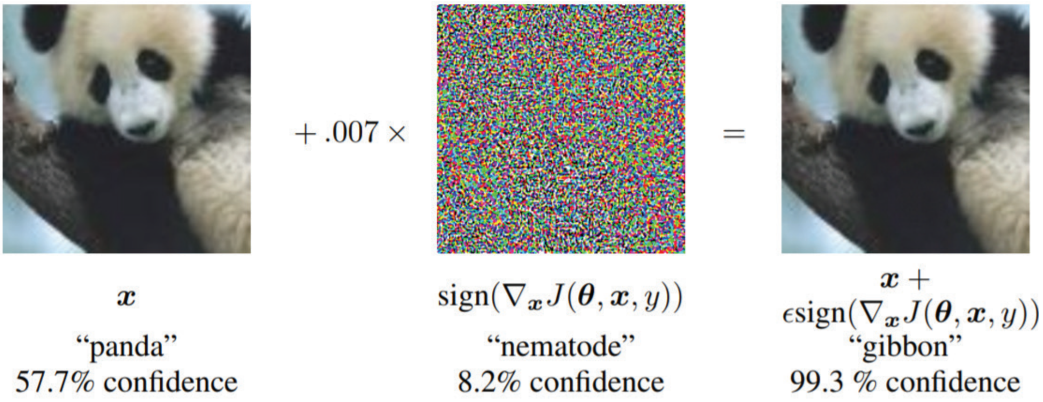
A very clear consequence of biases in content moderation models on the internet is the possibility of it being selective for under-represented groups in the data or even under-represented contexts. One example, Oliva, Antonialli, and Gomes (2021) provided evidence that AI systems may not correctly interpret the social context of discourse, failing to recognize cases in which words, that might conventionally be seen as offensive, carry different meanings in LGBTQIA+ speech. In this sense, Harrison et al. (2020) conducted a study on the human perception of fairness and unbias in bail decision systems, and concluded that realistic models are, consequently, necessarily imperfect in relation to the different definitions of fairness in AI. The truth is, the black box nature of traditional AI models makes it difficult to understand why an AI fails in certain cases, which cases are they, as well as making it difficult to predict its behavior.


Adversarial Text Generator
Input Text
"The characters, cast in impossibly **contrived situations**, are **totally** estranged from reality."
"The characters, cast in impossibly **engineered circumstances**, are **fully** estranged from reality."
Negative!
Positive!
SOTA NLP models
(e.g. BERT, LSTM, CNN)

Source: D. Jin et al. (2020).

Figure 5 – Demonstration of susceptibility to adversarial attacks of a sentiment analysis system in movie reviews. It is noticed that when modifying words by lesser-used synonyms leads the classifier to a wrong decision.

And how do the AI models behave in cases he has not been trained to handle (cases not on the training data)? The answer is that learning-based models can produce unexpected responses to unexpected inputs. Their vulnerability, above all, can be assessed through *Adversarial Attacks* (Goodfellow; Shlens; Szegedy, 2015), where the data are modified so as not to present themselves in the conventional way, confusing the model. Adversary attack methods seek to discover which are the cases in which the model will fail and generate an entry that causes this effect. A clear example is when a word is misspelled, as reported in the previous section. In this case, the model has never seen this word and cannot recognize it. However, this can happen in a more subtle way, as in the example of Figure 5 swapping words for their synonyms led to a different and incorrect prediction of the classifier. Or even in images, as shown in the Figure 6, where introducing noise in the image misleads the classifier to the wrong class prediction. The notion of *Adversarial Robustness* of AI systems is, then, the ability of the model to consider equal two things that are equal to human eyes, such as the examples in Figures 5 and 6.



+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Source: Goodfellow, Shlens and Szegedy (2015).

Figure 6 – A demonstration of fast adversarial example generation (Goodfellow; Shlens; Szegedy, 2015) applied to the GoogLeNet network (Szegedy et al., 2015) on the ImageNet dataset (Krizhevsky; Sutskever; Hinton, 2012). By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, GoogLeNet's classification of the image can be changed.

Although there is strong research on how to mitigate these effects in Artificial Intelligence, the potential flaws presented have been raising several concerns, questions, and discussions about the evolution of AI (Sichman, 2021), especially regarding the applicability of the enabled tools. The human-computer interaction of learning-based agents is an important and much-discussed aspect, but the role of these agents as moderators, effectively judging the content of users, is a new feature that adds complexity to the topic. In addition, the lack of transparency in how the moderation process takes place, as well as access to data, makes it difficult for researchers to better understand the process and seek to improve it. The opaque nature of this process can seed conspiracy theories such as the one investigated and refuted by Jiang, Robertson, and Wilson (2020), that social media moderation has a left-wing bias. Furthermore, the claims brought forward are difficult to validate, as neither researchers nor critics can access data referring to content removed by moderation decisions. It would, therefore, be of great importance for moderated/removed content to be preserved and protected by large platforms, enabling research that seeks to improve the decision process regarding moderation.

In this context, it is important to note that approaches that do not use AI can mitigate the effect of harmful content on platforms. Techniques already used in social networks range from automatic algorithms, based on sets of rules specified by human beings, to the active reinforcement of authentication policies, aiming to reduce the risk of fake accounts and forcing users to leave anonymity, exposing them to legal risks of sharing harmful content. Also, monetization policies that observe the nature of the content posted can also discourage the sharing of that type of content. Winchcomb (2019) details these techniques, also citing censorship of offending users, which seeks to impose social restrictions (such as limiting the number of interactions the user can have), and, finally, content curation performed by algorithms. This content curation, which involves determining what content the recommendation system suggests to users, can inadvertently prioritize harmful content or even unintentionally encourage the production of such content, depending on the curation approach. For example, giving preference to content with high numbers of views or interactions carries risks, as harmful content could potentially attract more engagement and sharing. Modern curation algorithms tackle this concern, and they are not exclusively reliant on AI techniques to manage it. For instance, platforms like Instagram and YouTube minimize engagement and even demonetize content that contains terms listed as harmful.

Legal dilemmas of the current moderation paradigm

Under this section, we intend to provide a brief panorama on how content moderation is debated around the world, so that we may better understand how explainable AI models may influence – helping or hindering – the protection of rights on the Internet. Online content moderation has sparked important discussions that related theoretical aspects and propositions with practical limitations of both States and digital platforms to implement regulations in the virtual sphere.

One can consider the starting point of the discussion to be the issue of whether human rights in the digital domain have been sufficiently conceptualized and discussed so that with a robust theoretical framework, we may be able to assess concrete problems and plan policy accordingly. A policy designed to protect rights related to new technologies may be seen as either in need of more factual evidence (Waldron, 2003) or as equivalent to offline rights, which is followed as a normative premise in international legal instruments (Tuori, 2019).

Either way, the current paradigm of balancing human rights in order to protect them is what guides the interactions and policy planning as of today. Paradigms may change if new rights arise with the development of technology, but if that is not the case, the current paradigms will still remain (Jóri, 2016). Theories such as Robert Alexy's view on proportionality (Alexy, 2014) are used to balance rights in studies to gather more factual evidence or applied to online rights the same way they would be applied to offline rights. Instead of shifting the paradigm, it is worth noting the standards through which a proportionality analysis would guide itself to balance rights so that we can better assess how explainable AI can improve content moderation.

Ideally, apart from legality, content moderation online should follow four main ideas in order to protect a truly democratic environment of debate and the sharing of ideas: legitimacy, transparency, control, and enforcement capability (Sander, 2020).

In the case of legitimacy, the restriction of freedom of expression of users is to be provided for by law (UN Human Rights Committee, 2011), as in accordance with the General comment n.34 of the UN Rights Committee. With this in mind, we see that the terms of service and community standards of companies such as social media platforms have adapted to comply with several different regulations at the same time, creating a dialogue between the legislation of different countries that regulate the same object, e.g. user interaction and comments moderation. This is not to say that States have forfeited their sovereignty, but rather that in complying with several human rights regulations, the approach needed by companies is an extensive one.

This leads us to the aspects of control and capability to execute (*enforcement*), which are intrinsically connected. Although States have the legitimacy to create the regulation on comment moderation, they do not necessarily possess the capability to act and monitor comments, even if we take into account specialized agencies.

The issue of time is a central one in this case, as the rights at stake can be forfeited by the capability of the virtual environment to disseminate and perpetuate information that might violate an individual's privacy. The general tendency to ensure control is the creation of standards and legal guidelines by States on how companies – the entities that have the most control over said situations, might act on these situations.

Such paths may vary but still fall under the same umbrella of creating standards for companies to exercise their ability to execute moderation. The Delphi case showed that the European Court of Human Rights decided to apply online the same patterns used for traditional media:

> Defamatory and other types of clearly unlawful speech, including hate speech and speech inciting violence, can be disseminated like never before, worldwide, in a matter of seconds, and sometimes remain persistently available online. These two conflicting realities lie at the heart of this case. Bearing in mind the need to protect the values underlying the Convention, and considering that the rights under Article 10 and 8 of the Convention deserve equal respect, a balance must be struck that retains the essence of both rights. (ECtHR, 10 October 2013, para. 110)

This shows us that in order to achieve transparency, we require clear legal standards to regulate moderation and the technology associated with it. This has been the rationale behind both the General Data Protection Regulation of the European Union (GDPR) and the Brazilian Internet Bill of Rights (*Marco Civil da Internet*). Clear limitations and even clearer principles to be applied in standards of content moderation are used to provide control, and allow for the capability to execute and regulate comments, the former including an obligation for companies to balance conflicting rights online.[3]

As the paradigm of balancing rights asserts itself, creating transparent AI that can have its standards in compliance with such regulations can only be achieved through explainable AI technology.

Explainable AI and Opportunities in Content Moderation

One of the biggest challenges in AI is dealing with its complexity and opacity. Understanding how technology works is a critical step for the realization of other principles, such as accountability, human control of technology, safety and security, and fairness and non-discrimination (Kiritchenko; Nejadgholi; Fraser, 2021). Transparent AI intends to shed light on the process of creating an automatic system and make it understandable to different stakeholders (Arbix, 2020). Transparency can refer to various practices depending on who is the audience and the beneficiaries of the explanations (Weller, 2019). For users, it alludes to *explainability* (or interpretability) associated with good documentation that allows them to use it. In this sense, explainability can be seen as the ability of the models to provide explanations for their decisions.

When it comes to explainability, more complex models (such as deep neural networks) tend to have their decisions more difficult to be interpreted. An Explainable AI system can consist of a model that has less complexity and therefore is intrinsically capable of providing explanations for its decisions (Model-Specific XAI), or it can rely on another model that is responsible for auditing its decisions and providing a set of post-hoc added explanations (Model-Agnostic XAI). These explanations can be generated for each decision (Local Explanation), identifying reasons for this specific decision, or can be globally brought (Global Explanation), when what matters is the understanding of the whole logic of a model, and following the entire reasoning leading to all the different possible outcomes (Adadi; Berrada, 2018). These explanations are generally provided in a more technical way, making relationships with the variables, and leaving the system to generate views for the user. However, the demand for developing systems capable of producing Natural Language Explanations (Natural-XAI) (Camburu et al., 2018) is also growing.

Promoting explainability in decisions made by algorithms has a number of advantages. In the following subsection, we present who can benefit from it and discuss concrete cases of application.

Concrete application of Explainable AI

Shin (2021) brings to the discussion the concept of causability, which precedes explainability. Causability provides justification for what must be explained, and why. This is done by determining the relative importance of the explainability features, a subsequent step to promote transparency to the algorithms. In the case study presented, it is evident that giving users explanations about why certain news items are recommended generates trust, while offering causal information regarding the generated explanation promotes emotional security for users. The results highlight the opportunities we discussed in this paper, showing the importance of including causability and explainability in AI systems.

This importance has multiple aspects. On the one hand, it is common to associate XAI with the possibility of technically auditing model decisions, helping engineers to interpret what they have produced. However, XAI can and should be used to generate explanations for users. Despite this, Bhatt et al. (2020) show how there is a gap between desired transparency and XAI in practice, as the adoption of explainability lately has served more internal stakeholders than end users yet.

Another important aspect is the lack of unity regarding the objectives/opportunities of XAI: groups from different areas, when studying the adoption of XAI, seek different objectives. Therefore, Mohseni, Zarei, and Ragan (2021) study this phenomenon in an attempt to systematize the evaluation methods and objectives of XAI design. Four main objectives/opportunities are defined, relating to lay users in AI – AI novices – to be pursued when offering explanations:

- **Algorithmic Transparency:** explain how the intelligent system works.
- **User Trust and Reliance:** improve end-users trust in the intelligent system.
- **Bias Mitigation:** help human users to inspect if the intelligent systems are biased.
- **Privacy Awareness:** provide a means for end-users to assess their data privacy.

The work carried out by Mohseni, Zarei, and Ragan (2021) also analyzes the opportunities generated for other types of users, defining data experts and AI experts. Thus, it also defines the objectives:

- **Model Visualization** and Inspection: similar to AI novices, *expert users* can also benefit from machine learning interpretability. It allows them to inspect model uncertainty and trustworthiness.
- **Model Tuning and Selection:** visual analytics approaches, for instance, can help *experts* to tune machine learning parameters for their specific domains. This makes it easier for them to compare multiple models and select the right model for their data.
- **Model Interpretability:** it allows getting new insights into the learning patterns of deep models.
- **Model Debugging:** increases the ability of *researchers* to use interpretability techniques aiming to improve model architecture and training process.

In this sense, XAI creates benefits for end users and also allows for accountability by developers and maintainers. Promoting explainability for lay users can generate relevant social impacts since the overwhelming majority of users are included in this category. However, in the context of engineering algorithms and models, society is represented by the other two categories (data experts and AI experts). The effects of promoting explainability in order to improve the visualization and inspection of models or their selection and training, for example, can improve **society's trust** in the decisions of these models. Furthermore, promoting interpretability and enabling debugging allows for possible errors to which all AI is subject to be more easily abstracted away by society, *reducing distrust and increasing confidence in the process.* Below, we discuss some cases where Explainable AI was applied in the context of content moderation.

Mohseni et al. (2021) were able to produce evidence that transparent criteria and explained decisions to the user have the potential to reduce the recurrence of false news sharing. Furthermore, they showed that the transparency in content moderation brought about by XAI helps to build proper trust in AI models, who understand their limitations but also their potential. This result raises the pedagogical potential that the use of Explainable AI has, which can help the user to understand the risks of sharing a certain type of content.

On the other hand, Kou and Gui (2020) in analyzing a large number of comments and interactions regarding the AI system that punishes players for breaking rules in the game *League of Legends*, noticed a need for an explanation for the decisions of AI criteria, especially in terms of its criteria in terms of its values and social norms, of clarifications regarding the specifics of its operation and of ways to avoid penalties in the future. The study concludes that in the ideal case the Explainable AI could not be a satisfactory universal answer for any and all cases as there is an important role for the community in helping to resolve these doubts. However, the XAI could play an essential role in bringing the community closer to the developers, contributing mainly on three points: making explanations of more complex technical level accessible; enabling users themselves to have more tools for a better dialogue and understanding on how moderation works; and contextualizing AI's decisions by explaining which rules it took into account in each decision.

A project that can be inspiring and relevant in this regard is Twitter Birdwatch (Coleman, 2021), which allows people to identify information in Tweets they believe is misleading and write notes that provide informative context. This enables quick response when misleading information spreads, adding context that people trust and find valuable. The notes are made visible directly on Tweets for the global Twitter audience when there is consensus from a broad and diverse set of contributors. Instead of using an AI to generate the explanations, this system uses the people in the community to generate explanations that can be used in the decision-making of the moderation algorithm and fulfill a pedagogical role among users.

Towards a new content moderation paradigm

Despite being private entities, digital platforms have become the public space for discussion. They are so big that they cannot be alien to society. This raises important questions about the legitimacy and transparency with which they exercise moderation and define their internal laws (Community standards), as their decisions start to affect a huge portion of the population. Here we argue that Explainable AI has the potential to alleviate some of these shortcomings noted in content moderation.

Regarding the legitimacy of moderation, it is possible that society, through the State and democratic channels, creates standards of universal moderation that guarantee respect for fundamental rights, including freedom of expression, but also guarantee that this freedom does not exceed the limits of the law. The use of accountability tools presented in the framework of the Explainable AI can allow these rules defined in society to be implemented by platforms, and in this way to be monitored, audited, and improved.

Allowing society to audit AI models being implemented and provide input on moderation policies makes the process more legitimate. There are studies, such as the one carried out by Vaccaro et al. (2021) that calls for representative moderation, which guarantee that people can have representatives within the moderators bodies of the digital platforms. In this sense, it is important that as much of the data produced in the moderation process as possible is in the public domain: removed content, moderation policies, decision criteria, model prediction and confidence scores, among others. This is essential to enable researchers to investigate and collaborate in these processes.

With regard to transparency for the end user, the production of local explanations in natural language of why a particular content of theirs was removed may play an even more important role than the removal itself for the safety of the public debate. Instead of excluding this citizen from the debate, it includes and pedagogically gives users the opportunity to learn to be more critical of the content they share on social media. There is a potential, to be explored in future research on Human-computer interaction in Explainable AI, that the use of XAI will help to reduce the recidivism of harmful content sharing. There is also an opportunity to use other approaches instead of directly removing the content, such as flagging, limiting the number of interactions that the post can have, and demonetizing, depending on the scope that the publication has had.

More interdisciplinary research, involving specialists from all areas (law, sociology, computer science, ethics, politics), is needed so that we can build a more concrete model for content moderation that encompasses the concepts presented here. This study was limited to interdisciplinary exploratory research that brings a preliminary proposal of how content moderation on the internet should be. However, this work could demonstrate that the introduction of the Explainable AI in this debate has a great potential to collaborate so that the internet becomes a fairer and healthier debate space.

Conclusion

In this work, we provided a comprehensive overview of the application of Artificial Intelligence in Internet Content Moderation. We adopted an interdisciplinary approach, with the aim of offering a comprehensive perspective of the moderation process employed on large platforms. Among the results identified, the exponential growth in the percentage of content automatically removed through AI stands out. Furthermore, we explored the ethical and social implications associated with the use of AI systems, outlining their limitations and the legal challenges that arise as a result of their implementation. We also introduced Explainable AI, highlighting the opportunities that the concept provides, as well as illustrating practical cases where this approach has been successfully applied in content moderation, potentially serving as a source of inspiration.

The main contribution of this study was to present an alternative to make the use of artificial intelligence in content moderation more transparent for the end-user and legitimate to society, and this involves the adoption of Explainable AI associated with moderation criteria defined together by society. This configuration could constitute a new paradigm of content moderation, fairer and more ethical, in which the State, Digital Platforms, and the citizens themselves have their relevant and well defined roles.

Adopting a more legitimate process guarantees respect for individual freedoms without neglecting legal limits. In turn, adopting a transparent moderation standard does not solely offer benefits tied to safeguarding minority rights. Transparency at the heart of the moderation process holds the capacity to enhance its overall efficacy, including reducing the recurrence of committing virtual abuse, sharing false news, etc., contributing to a healthier virtual environment.

Notes

[1] Available at: <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai--and-hate-speech-detection/>.

[2] The recall metric in statistics is the ratio of positive predictions that are correctly performed to all predictions that are actually positive. It is a relevant but limited metric, as it does not measure how many false positives the system produced. Related concepts that may be helpful for understanding are: precision, accuracy, and confusion matrix.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, IEEE, v.6, p.52138-60, 2018.

ALCOFORADO, A. et al. Zeroberto: Leveraging zero-shot text classification by topic modeling. In: PINHEIRO, V. et al. (Ed.) *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2022. p.125-36. ISBN 978-3-030-98305-5.

ALEXY, R. Constitutional rights and proportionality. *Revus. Journal for Constitutional Theory and Philosophy of Law/Revija za ustavno teorijo in filozofijo prava*, Klub Revus–Center za raziskovanje evropske ustavnosti in demokracije, n.22, p.51-65, 2014.

ARBIX, G. A transparência no centro da construção de uma IA ética. *Novos estudos Cebrap*, SciELO Brasil, v.39, p.395-413, 2020.

ARRIETA, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, Elsevier, v.58, p.82-115, 2020.

BHATT, U. et al. Explainable machine learning in deployment. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. [S.l.: s.n.], 2020. p.648-57.

BROWN, T. B. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, v.33, p.1877-901, 2020.

CAMBURU, O.-M. et al. E-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, v.31, p.9539-49, 2018.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v.6, n.1, p.1-6, 2004.

CIAMPAGLIA, G. L. et al. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, Nature Publishing Group, v.8, n.1, p.1-7, 2018.

COLEMAN, K. *Introducing Birdwatch, a Community-Based Approach to Misinformation*. [S.l.]: Twitter, 2021.

DANKS, D.; LONDON, A. J. Algorithmic bias in autonomous systems. In: PROCEEDINGS OF THE 26TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. [S.l.: s.n.], 2017. p.4691-7.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: PROCEEDINGS OF THE 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ACL. [S.l.: s.n.], 2019. p.4171-86.

DUARTE, N.; LLANSO, E.; LOUP, A. Mixed Messages? The Limits of Automated Social Media Content Analysis. In: PMLR. *Conference on Fairness, Accountability and Transparency*. [S.l.], 2018. p.106.

ECTHR. Delfi v. Estonia 64569/09. 2013. Delfi (n 9) para 110, Ibid para 59 (10 october 2013).

ESTARQUE, M.; ACHERGAS, J. V. *Redes Sociais e Moderação de Conteúdo*: criando regras para o debate público a partir da esfera privada. [S.l.], 2021. Disponível em: <https://itsrio.org/wp-content/uploads/2021/04/Relatorio_RedesSociaisModeracaoDeConteudo.pdf>.

FELZMANN, H. et al. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, Springer, v.26, n.6, p.3333-61, 2020.

FERNÁNDEZ, A. et al. *Learning from imbalanced data sets.* [S.l.]: Springer, 2018. v.10.

FERRAZ, T. P. et al. DEBACER: a method for slicing moderated debates. In: SBC. *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional.* [S.l.], 2021. p.667-8.

GOODFELLOW, I.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS. [S.l.: s.n.], 2015.

GOODMAN, B.; FLAXMAN, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, v.38, n.3, p.50-7, 2017.

GOOGLE. *Google Transparency Report.* 2021. URL: <https://transparencyreport.google.com≥ [Online;accessed 15-Out-2021]. Disponível em: <https://transparencyreport.google.com≥.

HARRISON, G. et al. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. [S.l.: s.n.], 2020. p.392-402.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, IEEE, v.21, n.9, p.1263-84, 2009.

JIANG, S.; ROBERTSON, R. E.; WILSON, C. Reasoning about political bias in content moderation. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. [s.l.: s.n.], v.34, n.9, p.13669-72, 2020.

JIN, D. et al. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. [S.l.: s.n.], v.34, n.5, p.8018-25, 2020.

JÓRI, A. Protection of fundamental rights and the internet: a comparative appraisal of german and central european constitutional case law. *The Internet and Constitutional Law: The protection of fundamental rights and constitutional adjudication in Europe.* London; New York: Routledge Taylor and Francis Group, 2016.

KIRITCHENKO, S.; NEJADGHOLI, I.; FRASER, K. C. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, v.71, p.431-78, 2021.

KLONICK, K. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, v.131, p.1598, 2017.

_______. The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, HeinOnline, v.129, p.2418, 2019.

KOU, Y.; GUI, X. Mediating community-ai interaction through situated explanation: The case of ai-led moderation. In: PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION, ACM New York, NY, USA, v.4, n. CSCW2, p.1-27, 2020.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, Springer, v.5, n.4, p.221-32, 2016.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v.25, p.1097-105, 2012.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v.521, n.7553, p.436-44, 2015.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v.54, n.6, p.1-35, 2021.

META. *Facebook Transparency Report.* 2021. Disponível em: <https://transparency.fb.com≥. Acesso em: 10 nov.2021

MOHSENI, S. et al. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In: PROCEEDINGS OF THE INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA. [S.l.: s.n.], v.15, p.421-31, 2021.

MOHSENI, S.; ZAREI, N.; RAGAN, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM New York, NY, v.11, n.3-4, p.1-45, 2021.

NAHMIAS, Y.; PEREL, M. The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harvard Journal on Legislation, Forthcoming*, 2020.

NG, A. What artificial intelligence can and can't do right now. *Harvard Business Review*, v.9, n.11, 2016.

OLIVA, T. D.; ANTONIALLI, D. M.; GOMES, A. Fighting hate speech, silencing drag queens? Artificial Intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, Springer, v.25, n.2, p.700-32, 2021.

REIS, J. C. et al. Explainable Machine Learning for Fake News detection. In: PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE. [S.l.: s.n.], p.17-26, 2019.

SANDER, B. Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation. *Fordham International Law Journal*, v.43, n.4, 2020.

SHIN, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, Elsevier, v.146, p.102551, 2021.

SICHMAN, J. S. Inteligência artificial e sociedade: avanços e riscos. *Estudos Avançados*, v.35, p.37-50, 2021.

SUSI, M. The Internet Balancing Formula. *European Law Journal*, v.25, n.2, p.198-212, 2019.

SZEGEDY, C. et al. Going deeper with convolutions. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. [S.l.: s.n.], p.1-9, 2015.

TAN, F. et al. TNT: Text Normalization based Pre-training of Transformers for Content Moderation. In: PROCEEDINGS OF THE 2020 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP). [S.l.: s.n.], p.4735-41, 2020.

TUORI, K. Principles and policies: once more. In:_____. *The Quest for Rights*. [S.l.]: Edward Elgar Publishing, 2019.

UN Human Rights Committee. General comment no. 34: Article 19: Freedom of opinion and expression, u.n. doc. ccpr/c/gc/34. 2011. 12 Sept. 2011 [hereinafter General Comment No. 34], para. 25.

VACCARO, K. et al. Contestability for content moderation. *Proceedings of the ACM on Human-Computer Interaction*, ACM New York, NY, USA, v.5, n.CSCW2, p.1-28, 2021.

WALDRON, J. Security and liberty: The image of balance. *The Journal of Political Philosophy*, v.11, n.2, p.191-210, 2003.

WELLER, A. Transparency: motivations and challenges. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. [S.l.]: Springer, 2019. p.23-40.

WINCHCOMB, T. *Use of AI in Online Content Moderation*. [S.l.], 2019. Disponível em: <https://www.cambridgeconsultants.com/insights/whitepaper/ofcom-use-ai-online-content-moderation>.

ABSTRACT – The massive use of Artificial Intelligence in Content Moderation on the internet is a reality of our times. However, this raises a number of questions, such as whether the use of opaque automatic systems is pertinent, or even whether platforms alone can make decisions that used to be made by the State. In this context, the use of *black box* AI comes to be considered a threat to freedom of expression. On the other hand, keeping content that promotes virtual abuse is equally harmful to this fundamental right. In this scenario, this study summarizes the main problems pointed out by the literature regarding the current paradigm, evaluates the responses that new technologies bring, and proposes a path for a new moderation paradigm that is fair and ethical in which the State and social media platforms play a relevant role. That involves the adoption of *Explainable AI* associated with transparent and legitimate criteria defined by society.

KEYWORDS: Digital humanities, Automatic content moderation, Explainable AI, Freedom of expression on the internet, Ethics in Artificial Intelligence.

RESUMO – O uso massivo de Inteligência Artificial na moderação de conteúdo na internet é uma realidade dos tempos atuais. No entanto, isso levanta uma série de questionamentos, seja sobre a pertinência do uso de sistemas automáticos opacos, seja se as plataformas podem sozinhas tomar decisões que antes cabiam ao Estado. Nesse contexto, o uso de IA "caixa-preta" passa a ser considerado uma ameaça à liberdade de expressão. Por outro lado, manter conteúdos que promovam abuso virtual é igualmente danoso a este direito fundamental. Nesse cenário, este estudo sumariza os principais problemas apontados pela literatura quanto ao paradigma atual, avalia as respostas que as novas tecnologias trazem, e propõe um caminho para um novo paradigma de moderação que seja justo e ético, no qual Estado e plataformas de mídias sociais têm papel relevante. Esse passa pela adoção de IA explicável associada a critérios transparentes e legítimos definidos pela sociedade.

PALAVRAS-CHAVE: Humanidades digitais, Moderação automática de conteúdo, IA explicável, Liberdade de expressão na internet, Ética na Inteligência Artificial.

Thomas Palmeira Ferraz is a PhD candidate in Computer Science at Télécom Paris and at École Polytechnique, Institut Polytechnique de Paris. He has graduated in Engineering from the University of São Paulo (USP) and holds a master's degree in Applied Mathematics and Artificial Intelligence from the École Normale Supérieure Paris-Saclay (ENS). @ – thomas.palmeira@telecom-paris.fr / http://orcid.org/0000-0002-5385-9164.

Caio Henrique Dias Duarte is a master's student at the Department of International and Comparative Law at the Law School of the University of São Paulo (USP). He holds a bachelor's degree in Law from the same university. @ – caio.henrique.duarte@usp.br / https://orcid.org/0000-0002-1720-7249.

Maria Fernanda Ribeiro is a collaborating researcher at the University of São Paulo (USP). She has graduated in Engineering from the same institution.
@ – maria.fernanda.ribeiro@alumni.usp.br / http://orcid.org/0000-0002-4340-9901.

Gabriel Goes Braga Takayanagi is an undergraduate at the Department of Computer Engineering and Digital Systems at the Polytechnic School of the University of São Paulo (USP).
@ – gabriel.takayanagi@usp.br / https://orcid.org/0000-0002-6498-6716.

Alexandre Alcoforado is a master's student in Computer Engineering at the Polytechnic School of the University of São Paulo (USP). He has graduated in Electrical Engineering from the same institution.
@ – alexandre.alcoforado@usp.br / http://orcid.org/0000-0003-3184-1534.

Roseli de Deus Lopes is a full professor at the Polytechnic School of the University of São Paulo (USP), vice-coordinator of the Interdisciplinary Center for Interactive Technologies (Citi-USP), and a researcher at the Integrated Systems Laboratory (LSI-USP), where she coordinates research projects in interactive electronic media with an emphasis on applications in education, inclusion, and health. She is currently the director of the Institute of Advanced Studies at USP. @ – roseli.lopes@usp.br /
https://orcid.org/0000-0001-8556-6473.

Mart Susi is a professor of Human Rights Law and Head of Legal Studies at Tallinn University, Estonia. @ – martsusi@tlu.ee /
https://orcid.org/0000-0002-2624-4797.

Received on 27.2.2022 and accepted on 1.22.2024.

[I] Institut Polytechnique de Paris, École Polytechnique, Paris, France.

[II] University of São Paulo, Law School, São Paulo, Brazil.

[III] University of São Paulo, Engineering School, São Paulo, Brazil.

[I,IV,V] University of São Paulo, Polytechnic School, São Paulo, Brazil.

[VI] University of São Paulo, Institute of Advanced Studies, São Paulo, Brazil.

[VII] Law School, Tallinn University, Tallinn, Estonia.