"AM I BEING SILENCED BY A MACHINE?" AI-DRIVEN CONTENT MODERATION AND THE CHILLING EFFECT ON FREEDOM OF EXPRESSION

Ramalina Ranaivo Mikea Manitra
ASTA Research Center, Madagascar
Correspondent Email: manitraramalina@gmail.com

Abstract

As artificial intelligence increasingly governs online content moderation, concerns have mounted over its implications for freedom of expression and democratic participation. This paper aims to examine the legal and human rights challenges posed by AI-driven content filtering, with a focus on the emergence of chilling effects and unequal impacts across user groups. Using legal doctrinal analysis, this study interrogates how algorithmic moderation models operate and how they align—or fail to align—with international human rights norms. The findings reveal that AI systems frequently suppress lawful speech, especially from marginalised communities, due to biased training data and opaque decision-making processes. Furthermore, existing regulatory responses remain fragmented, lacking the transparency, accountability, and normative clarity required to uphold free expression. Drawing from recent UN reports and resolutions, the paper highlights growing international critiques and supports calls for human rights-based governance to ensure AI fosters an inclusive, rights-respecting digital age.

Keywords: Artificial intelligence; Chilling effects; Content moderation; Freedom of expression; International human rights law.

Abstrak

Seiring dengan meningkatnya peran kecerdasan buatan (AI) dalam moderasi konten daring, kekhawatiran berkembang mengenai dampaknya terhadap kebebasan berekspresi dan partisipasi demokratis. Artikel ini mengkaji tantangan hukum dan hak asasi manusia yang ditimbulkan oleh penyaringan konten berbasis AI, dengan fokus pada munculnya efek mengerikan (chilling effect) dan dampak tidak merata terhadap kelompok pengguna. Melalui metode hukum doktrinal, studi ini menelaah cara kerja model moderasi algoritmik serta sejauh mana praktik tersebut selaras—atau tidak selaras—dengan norma hak asasi manusia internasional. Temuan menunjukkan bahwa sistem AI sering menekan ekspresi sah, terutama dari komunitas terpinggirkan, akibat data pelatihan bias dan proses pengambilan keputusan yang tidak transparan. Selain itu, respons regulasi yang ada masih terfragmentasi dan belum memenuhi standar transparansi, akuntabilitas, serta kejelasan normatif untuk melindungi kebebasan berekspresi. Berdasarkan laporan dan resolusi PBB terbaru, artikel ini menegaskan pentingnya tata kelola AI berbasis hak asasi manusia demi era digital yang inklusif dan menghormati hak.

Kata Kunci: Kecerdasasn buatan; Efek mengerikan; Moderasi konten; Kebebasan berekspresi; Hukum hak asasi manusia internasional.

DOI: 10.24252/aldev.v7i2.57618

INTRODUCTION

Artificial Intelligence (AI) has grown remarkably in its ability to regulate digital content, particularly through the rise of automated systems designed to filter, moderate, and prioritise user-generated material. These systems, while often celebrated for their efficiency and scalability, pose significant challenges when viewed through the lens of fundamental rights (Sitabuana et al. 2024). With the explosion of online content, AI-based moderation has become a common feature across platforms, operating largely through opaque mechanisms that determine what can and cannot be seen or shared (Ashraf 2022; Babu and Darshini 2025). The rapid implementation of such technologies often occurs without adequate human oversight or user notification, raising concerns over accountability and transparency. In this context, platforms increasingly rely on algorithmic frameworks such as TensorFlow or PyTorch to manage massive volumes of content (Babu and Darshini 2025), promoting a vision of moderation that is fast and consistent but potentially devoid of nuance. While the aim may be to eliminate harmful material, the unintended consequence is a form of silent regulation—one that shapes online discourse without users' knowledge, consent, or recourse.

AI systems are frequently promoted as neutral tools serving the public interest by preventing the spread of hate speech, disinformation, or other harmful content. Yet, empirical studies have exposed their inherent susceptibility to reproducing and amplifying social biases. For instance, automated disinformation detection systems have been shown to disproportionately suppress minority or dissenting voices, often removing controversial but legally protected content (Hasimi and Poniszewska-Marańda 2024). As a concrete example, Keller noted, automated systems have unfairly burdened Arabic speakers by removing innocuous content such as a prayer on Facebook, which read "God, before the end of this holy day, forgive our sins…" for allegedly violating community standards (Alkiviadou 2022). Because these algorithms are trained on datasets that often reflect dominant cultural and ideological patterns, they may disproportionately affect expressions that differ from mainstream perspectives. As a result, marginalised groups—whether based on ethnicity, religion, gender identity, or political stance—may experience silencing not through overt censorship, but through the automated and invisible operations of algorithmic exclusion (Parmar and Murari 2025; Gentelet and Mizrahi 2024). The problem is not simply technological; it is systemic and reflective of how power operates in digital infrastructures. The legal and ethical implications of this are profound, particularly in contexts where freedom of expression forms the backbone of democratic engagement.

The risks are not abstract. Various studies have documented that AI moderation mechanisms not only suppress speech but also disproportionately affect vulnerable users. Perilo and Valença (2024) found, for example, that facial recognition technologies systematically misidentify transgender individuals, reflecting a broader issue in which AI not only mirrors but also entrenches existing social inequalities. Similarly, automated content moderation tools have been reported to wrongly flag content related to sexual and religious identity, further discouraging these communities from participating in digital spaces (Gentelet and Mizrahi 2024). These mechanisms do not operate in a vacuum. Rather, they enforce silent hierarchies where certain voices are more easily heard than others. This leads many users—especially those from marginalised communities—to experience what scholars call a "chilling effect," a situation where people choose to withhold or modify their speech due to fear of algorithmic retaliation (Khare and Raghuwanshi 2025). As AI becomes the gatekeeper of online speech, it inadvertently fosters a climate where silence becomes safer than speech.

This chilling effect represents a critical juncture in the legal and philosophical discourse on digital rights. The right to freedom of expression, enshrined in numerous international legal instruments (e.g., UDHR and ICCPR), has traditionally been understood as protection against state censorship. However, with AI assuming regulatory functions previously reserved for human actors or public institutions, questions arise concerning the extent to which algorithmically mediated communication qualifies for similar protections (Goswami 2020). When an AI system removes or deprioritises content, does this amount to censorship? And if so, who bears the responsibility—platforms, developers, or governments? These dilemmas challenge conventional legal frameworks and suggest the urgent need for legal evolution. As private platforms increasingly wield power comparable to that of state actors, the gap between legal norms and technological realities continues to widen, endangering the protection of individual liberties in digital spaces (Caruso 2025; Martínez 2023).

Though technological solutions such as hybrid moderation models—where AI is supplemented by human review—have been proposed as a compromise, they bring their own set of challenges. On one hand, they attempt to blend algorithmic efficiency with human discernment, aiming for a more balanced approach (Lalchhanhima, Rajendran, and Madhusudanan 2025). On the other hand, human moderators are often exposed to high-stress environments, required to review disturbing content quickly and consistently, which leads to mental health burdens and moral injury (Fahrudin, Tiwari, and Rahmi 2025). Moreover, the success of these models depends on institutional support and ethical guidelines, which are frequently lacking. Simultaneously, users grow increasingly frustrated with the opacity of moderation decisions, especially when sanctions are imposed without explanation (Subrahmanyam 2025). For many creators, this uncertainty fosters anxiety and reduces creative expression, reinforcing the chilling effect and calling into question whether AI moderation is genuinely serving the public good or merely fulfilling corporate interests.

In terms of scholarly contributions, some studies have already explored adjacent themes. For example, Vacarelu (2023) examined how AI systems were used in political discourse and warned of their potential to distort democratic participation. Similarly, Ghose, Pallav, and Ali (2025) investigated algorithmic bias in multicultural societies and concluded that AI tools tend to amplify political polarisation. Subrahmanyam (2025) also studied users' experiences with AI moderation and highlighted the growing mistrust in digital platforms. While these works offer valuable insights, they often approach the subject from either a political or a technical standpoint, without integrating legal analysis in depth. This research distinguishes itself by positioning the issue of AI filtering squarely within a rights-based legal framework, with particular emphasis on freedom of expression as protected under international human rights law. Its novelty lies in treating AI moderation not merely as a design or governance issue, but as a structural transformation of speech regulation—raising normative questions about legality, legitimacy, and accountability in the digital public sphere.

Accordingly, this research will be structured around three primary inquiries. First, it will examine how AI systems filter online speech, including the technical processes, decision-making models, and underlying data that inform content moderation practices. Second, it will analyse the impact of such AI-driven filtering on freedom of expression, with particular attention to the emergence of a chilling effect and the ways in which different user groups may be disproportionately affected. Third, it will examine international responses to the legal and ethical challenges raised by AI filtering, focusing on critiques of its impact on human rights and proposing recommendations for a human rights-based approach to AI governance. Through this structure, the study aims not only to unpack the implications of AI-based moderation on fundamental rights, but also to contribute to ongoing debates on how to regulate AI in a manner that safeguards freedom of expression in the digital age.

METHOD

This research is a legal doctrinal study that aims to examine how AI-based content moderation affect freedom of expression. It employs a combination of normative and conceptual to analyse the development and implications of AI governance within the framework of international human rights law. The study relies on primary legal materials, including the Universal Declaration of Human Rights, International Covenant on Civil and Political Rights, and relevant United Nations documents, such as UN General Assembly Resolution No. A/RES/79/1 – The Pact for the Future and UN General Assembly Report No. A/73/348. Secondary legal materials include peer-reviewed journal articles and books obtained through an in-depth literature review using the Scopus database. Tertiary legal materials, such as legal dictionaries and encyclopaedias, are used to support conceptual clarity. The data are analysed using qualitative methods in order to obtain a deductive analysis.

RESULT AND DISCUSSION

1. How AI Systems Filter Online Speech

AI systems have revolutionised the regulation of online speech by introducing automated technologies capable of detecting and filtering harmful content across digital platforms (as summarised in Figure 1). At the core of these systems lie Natural Language Processing (NLP) and Machine Learning (ML) techniques, which process immense volumes of user-generated text. These mechanisms are responsible for identifying hate speech, cyberbullying, fake news, and other unlawful expressions. Notably, advanced detection models such as Multi-Head Self-attention Bi-directional Long Short-Term Memory (MHS-BiLSTM) have been applied to improve the accuracy of detecting deceptive content within dynamic social media environments (Albraikan et al. 2023). This initial layer of AI filtering forms the primary gateway to content moderation, operating at speeds and scales that human moderators could not achieve, and ultimately determining the visibility of speech in the digital public sphere.

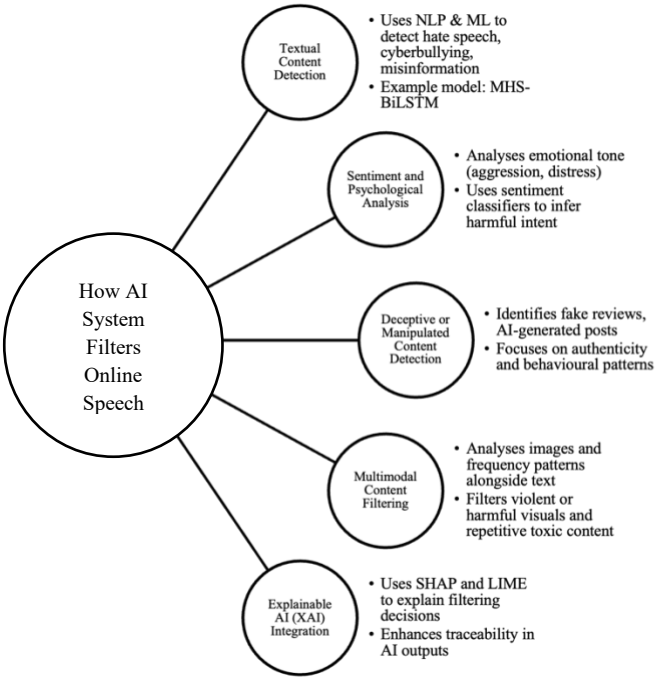

Figure 1. Technical Layers of AI-Based Content Filtering

Source: Author`s analysis

In parallel, AI filtering tools extend their function by assessing the emotional and psychological tone of digital communication. Through sentiment analysis, AI can detect aggression, distress, or behaviours indicating potential harm. These sentiment classifiers evaluate not only what is said, but also how it is emotionally charged—revealing content that may incite violence or flag early signs of mental health issues (Akhil et al. 2023; Rayhan et al. 2024). This expands the range of content flagged for moderation, including expressions that may not cross legal thresholds but are perceived as emotionally disruptive. Consequently, the filtering of online speech becomes more subjective, as AI systems infer psychological states and regulate based on behavioural cues, raising deeper human rights questions about the standards used to restrict digital expression.

The filtering function further extends to AI's capacity to suppress manipulated or deceptive information. Sophisticated models have been trained to assess the complexity of language and user behaviour to detect AI-generated posts or fake reviews that are not easily identifiable by humans (Gambetti and Han 2023). In such instances, speech that mimics human discourse but is intended to mislead—such as misinformation or inauthentic endorsements—is automatically flagged or removed. This proactive filtering approach reveals how AI moderates not only based on legality, but also on perceived authenticity and purpose of speech. The power of such systems lies in their ability to reshape the information landscape by quietly removing content before it spreads, often without users realising that their speech has been scrutinised or restricted.

Much of this process, however, operates within a framework of opacity that is concerning from a rule of law and human rights standpoint. Many users are unaware of how AI moderation systems function, or why their content is removed. The absence of explanation for takedowns contributes to a lack of transparency that breeds confusion and mistrust (Khare and Raghuwanshi 2025). When users do not know which standards are applied or what content may be flagged, they may self-censor to avoid potential penalties. This form of invisible regulation—where users feel surveilled by an unseen digital authority—undermines the right to freedom of expression and public participation in democratic dialogue. It also violates the legal principle of foreseeability, which requires that restrictions on speech be clear, accessible, and predictable.

One of the most controversial elements in AI filtering is the issue of bias. Moderation algorithms are often trained on datasets that already reflect societal prejudices, which the AI then reproduces and amplifies. Biased filtering can result in discriminatory treatment of users from marginalised communities, as linguistic patterns associated with specific groups are flagged more frequently (Parmar and Murari 2025). Without cultural nuance or context, these systems tend to penalise lawful but unconventional expression, disproportionately affecting non-dominant speech communities (Tiwari and Fahrudin 2025). For example, internal documents and whistleblower testimony revealed that Meta's automated moderation systems disproportionately over-enforced Arabic-language content while applying less consistent scrutiny to Hebrew posts, leading to the systematic suppression of pro-Palestinian expression during the Israel–Palestine conflict (Paul 2024). The inability of AI to understand complex human communication reinforces the concern that a one-size-fits-all approach fails to meet the requirements of fairness, inclusivity, and equality in content moderation.

Filtering systems are also tasked with combating violent extremism and hate-based content. They monitor for language, symbols, and behaviours associated with radical ideologies, such as far-right extremism or incitement to terrorism (Gunton 2022). These systems are trained to recognise known extremist narratives and remove content or users deemed to present a high risk. This process aligns with international legal obligations requiring states and platforms to prevent the spread of terror-related content and protect public safety. Additionally, content containing explicit violence, abuse, or harmful

visuals is flagged and removed automatically, particularly to shield children and vulnerable users from distressing material (Ahmed et al. 2023). Such filtering is based not only on legal prohibitions, but also on ethical imperatives to prevent psychological harm. Nevertheless, McQue (2024) reported that Meta's reliance on AI-generated child abuse reports has caused significant delays in criminal investigations, as U.S. law enforcement cannot act on these reports without human review and a search warrant—highlighting both the ethical imperative and practical limitations of AI-based filtering for violent and harmful content.

Lastly, some moderation systems have begun to incorporate explainable AI (XAI) methods, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which aim to clarify the rationale behind content filtering decisions (Gongane, Munot, and Anuse 2024). These models enable platforms and users to understand why certain posts are labelled as hate speech, misinformation, or incitement. While this level of transparency is still in early development, it represents an attempt to reconcile algorithmic governance with fundamental legal standards, particularly the principle of legality and accountability in speech regulation. Coupled with multi-modal approaches—where image and frequency analysis is used alongside textual review—AI filtering now extends beyond written words to include visual content as well, applying layered scrutiny to all forms of digital expression (Poredi, Nagothu, and Chen 2024). Through this expansive but often opaque system, AI plays a decisive role in shaping what may lawfully or unlawfully be expressed online.

2. Impact of AI Filtering on Freedom of Expression

The increasing reliance on AI filtering systems to moderate online content has led to widespread concerns about its impact on the right to freedom of expression (as illustrated in Figure 2). While these systems are developed to efficiently manage harmful or illegal content, they often operate without human supervision and within opaque decision-making processes. This creates a situation where lawful content may be flagged and removed based on internal algorithmic logic rather than legal standards (Elkin-Koren 2020). Without proper mechanisms for review or appeal, users are left without any means to challenge these decisions, thus undermining their ability to express themselves freely online. The result is not only a technical error but a structural problem where speech is suppressed without accountability, in contradiction to human rights law requiring restrictions to be lawful, necessary, and proportionate (Marsoof et al. 2023; Al-Sherman 2024).

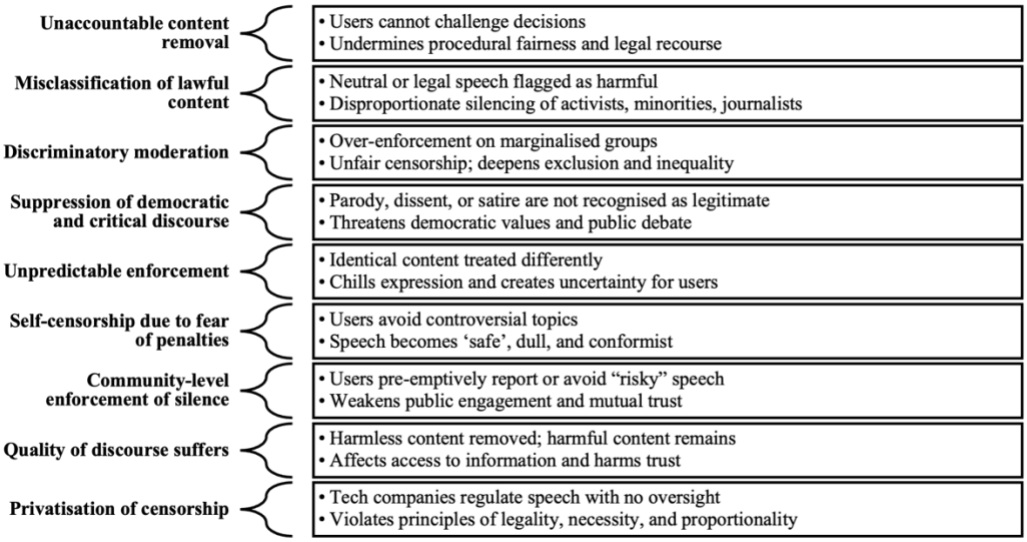

Figure 2. The Impact of AI Filtering Features on Freedom of Expression

Source: author`s analysis

A major issue lies in the misclassification of content due to AI's limited ability to understand context. Many systems incorrectly label legal or neutral content as harmful because they cannot interpret nuance or cultural meaning (Marsoof et al. 2023). This is especially dangerous for activists, journalists, or minority groups who rely on online platforms to share views that may not align with mainstream narratives. Their content, although lawful, is often disproportionately targeted, leading to their voices being silenced or hidden from public view (Parmar and Murari 2025). For example, Human Rights Watch (2023) documented over 1,050 cases on Meta platforms where peaceful pro-Palestinian content—such as Palestinian flags and nonviolent slogans—was erroneously removed or shadow-banned by moderation systems lacking cultural nuance, disproportionately silencing minority voices. Chang et al. (2025) note that the performance of AI models varies widely across different vendors, meaning that identical content might be treated differently depending on the system in use. This inconsistency in enforcement creates unpredictability and chills user participation in digital forums.

Compounding the problem is the evident bias embedded in the datasets used to train AI systems. When these systems are trained on narrow, non-representative data, they may over-detect content from certain communities while under-detecting real harm directed at them (Parmar and Murari 2025). This results in unfair and discriminatory moderation practices, where marginalised groups are penalised more frequently. The opaque nature of AI systems means that users are rarely informed why their content was removed or how the decision was reached (Frosio 2024). This lack of transparency prevents users from seeking redress, violating procedural fairness and weakening the rule of law in the digital sphere. Such unaccountable censorship reinforces existing social hierarchies and worsens exclusion.

Moreover, AI filters are typically programmed to detect specific categories like hate speech or copyright violations, without accounting for broader democratic values. This limited scope fails to protect content that might involve parody, criticism, or political dissent—forms of expression that are essential in any democratic society (Elkin-Koren 2020). Consequently, content that contributes to public debate or social awareness may be incorrectly removed simply because it does not conform to predefined filtering categories. The overreach of these systems therefore risks silencing voices that play a crucial role in shaping public opinion, especially when dissenting views are confused with misinformation (Hasimi and Poniszewska-Marańda 2024).

Another pressing concern is the emotional and behavioural impact on users. Subrahmanyam (2025) found that confusion and frustration resulting from content removals often lead users to withdraw from digital platforms. Many users are unable to understand or respond to takedown decisions, creating a sense of powerlessness. Over time, this leads to a behavioural change where users limit their expression to what they believe will not be penalised. This narrowing of discourse is not caused by any law, but by fear—fear of being flagged, suspended, or silenced. The consequence is a digital environment where users only express 'safe' views, while critical or controversial speech gradually disappears from public discussion.

This fear extends to more vulnerable groups such as activists and minority communities. Elmimouni et al. (2025) documented that these groups often self-censor due to repeated experiences of over-enforcement. Their efforts to raise awareness or challenge dominant narratives are met with frequent takedowns, pushing them to avoid using certain terms or limit the reach of their messages. This suppresses important social discourse and weakens the democratic function of online platforms. As a result, the internet risks becoming a space where only majority or state-approved narratives can thrive, while marginalised voices are effectively silenced.

In this context, the influence of AI extends beyond the technical and enters into the psychological. Penney (2020) demonstrated that users moderate their own behaviour when they feel they are being watched. The knowledge that an AI system is constantly monitoring their speech discourages users from expressing controversial or sensitive opinions. This self-censorship is a silent but serious threat to freedom of expression, as individuals start to internalise platform rules without ever being told what the exact limits are. The fear of being misunderstood by an opaque machine leads many to avoid speaking altogether, even when their content is completely lawful.

This chilling effect is not just personal but collective. Wang (2024) observed that in certain communities, users begin to engage in "participatory censorship," where they report each other to remain compliant with expected norms. In these situations, censorship is no longer imposed from above but reproduced by users themselves based on what they think the algorithm wants. This internalisation of control reflects how AI filtering can restructure human behaviour and weaken democratic engagement. Users no longer feel empowered to speak but instead become enforcers of silence in an environment ruled by algorithmic expectations.

Further complicating the matter are the frequent occurrences of false positives and false negatives in AI moderation. Systems that wrongly remove lawful content or fail to catch harmful material not only reduce the quality of discourse but also expose users to real-world harm (Ajani and Ferrante 2024). These dual failures show that AI is currently unable to strike a fair balance between content control and protection of fundamental rights. Such errors directly affect access to information, the free flow of ideas, and political participation, all of which are protected under international human rights law (Marsoof et al. 2023).

Finally, the absence of a harmonised legal framework for AI filtering aggravates the situation. Khare and Raghuwanshi (2025) point out that without clear standards, private technology companies are left to act as de facto regulators of speech, a role traditionally reserved for public authorities subject to legal oversight. The privatisation of censorship without adequate safeguards contradicts the principles of legality, necessity, and proportionality required by international law (Al-Sherman 2024). This vacuum in legal responsibility allows unchecked interference with free expression, raising serious concerns for the protection of human rights in the digital age.

3. Towards a Human Rights-Based Regulation of AI: International Community Responses

This last section explores how the international community has responded to the human rights risks associated with AI, particularly in relation to freedom of expression. It begins by outlining the main criticisms raised by international bodies concerning current regulatory gaps and the negative impacts of AI on fundamental rights, such as opaque decision-making and discriminatory outcomes. The second part then presents key recommendations for ensuring AI governance is grounded in human rights principles, including transparency, accountability, fairness, and inclusivity. At the end of the section, Figure 3 visually summarises how these proposed governance measures are designed to address the specific human rights challenges posed by AI filtering systems as raised by international critiques.

a. International Critiques of AI Regulation and Human Rights Risks

The international community has consistently raised concern about the rapid development and deployment of AI technologies, particularly their insufficient regulation and significant potential to infringe upon human rights. The Human Rights Council, through Resolution A/HRC/RES/53/29, explicitly recognises that AI systems can pose serious risks to a broad spectrum of rights—including the rights to privacy, non-discrimination, freedom of expression, and access to an effective remedy (A/HRC/RES/53/29, preamble). These concerns stem from the ways AI is currently used, including in facial recognition, behavioural scoring, and surveillance, often without meaningful safeguards. The lack of transparency and accountability in such applications underscores the urgency for a more robust global regulatory framework grounded in human rights standards.

One of the central criticisms raised by international bodies is the persistent assumption of AI's objectivity. The Special Rapporteur on contemporary forms of racism highlights this misconception in A/HRC/56/68, stating that technology is not neutral but reflects the interests and values of its creators (A/HRC/56/68, para 7). This flawed belief allows biased and discriminatory outcomes to be embedded in AI systems without sufficient scrutiny. The so-called "black box" problem—where algorithms evolve beyond human understanding—further complicates accountability and increases the risk of opaque decision-making (A/HRC/56/68, para 21). Without transparency, victims of discrimination caused by AI systems may find it impossible to challenge decisions or seek remedies.

The discriminatory impact of AI is particularly alarming in high-stakes contexts. The Special Rapporteur cites examples in law enforcement, education, and healthcare, where algorithmic bias has already resulted in unequal treatment, particularly along racial lines (A/HRC/56/68, para 25). In some instances, facial recognition technology has been linked to the wrongful arrests of individuals of African descent (A/HRC/56/68, para 27). These cases illustrate the real-world consequences of unregulated AI and underscore the need for regulatory approaches that account for historical and structural discrimination. In this regard, international law's commitment to non-discrimination, enshrined in instruments such as the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR), must be meaningfully applied to emerging technologies.

Another pressing concern relates to AI's role in amplifying disinformation, hate speech, and incitement to violence. As outlined in the Human Rights Council document A/HRC/59/28, AI-powered content moderation tools on social media platforms have been used both to mitigate and to unintentionally escalate such harms. The Special Adviser on the Prevention of Genocide warned that new technologies are enabling the proliferation of hate speech at unprecedented speed, contributing to the risk of atrocity crimes (A/HRC/59/28, para 11). The global spread of such content, often targeting

vulnerable communities, reflects a failure by both states and technology companies to implement effective, rights-based governance mechanisms.

The *Rabat Plan of Action*, referenced in A/HRC/59/28, offers a framework for identifying speech that constitutes incitement to violence or discrimination. However, its implementation remains inconsistent, and many AI content moderation systems fail to meet its threshold tests (A/HRC/59/28, para 15). Furthermore, automated moderation tools frequently lack linguistic and contextual sensitivity, resulting in the underenforcement or overenforcement of policies depending on the region or language (A/HRC/59/28, paras 18, 60). This uneven application not only undermines trust in AI systems but also raises questions about the equality of access to protections under international human rights law.

The critique extends beyond governments to business enterprises, particularly large technology companies that design and operate AI systems. The Human Rights Council emphasises that these entities must respect human rights under the UN Guiding Principles on Business and Human Rights, including by conducting human rights due diligence and participating in accountability processes (A/HRC/RES/53/29, preamble; para 7). However, as noted in multiple UN reports, current industry practices often fall short. Transparency reports are delayed and superficial, and oversight mechanisms such as independent review boards remain rare or ineffective (A/HRC/59/28, paras 16, 62). The lack of binding international regulation allows corporations to operate with considerable discretion, often prioritising commercial interests over rights protection.

A recurring theme in UN documentation is the mismatch between the speed of technological development and the pace of regulatory responses. The Special Rapporteur warns that regulatory measures have failed to keep up with AI's rapid expansion, allowing systems with harmful biases to be deployed at scale (A/HRC/56/68, para 6). This regulatory lag creates a permissive environment for rights violations, especially in jurisdictions lacking strong domestic safeguards. Furthermore, this imbalance disproportionately affects marginalised populations, reinforcing existing inequalities and obstructing efforts to close global human rights gaps.

Despite repeated calls for stronger action, including the rejection of a "colour-blind" approach to regulation (A/HRC/56/68, para 66), many stakeholders continue to view AI as a technical issue rather than a human rights concern. This perception limits the scope of regulation to technical fixes rather than legal and ethical reforms grounded in human rights law. As long as this narrow view prevails, international critiques will likely persist. Effective regulation must address not only algorithmic design but also the socio-political structures in which AI is developed and deployed.

In conclusion, the international community's critiques reveal a wide range of human rights risks stemming from the unregulated or poorly regulated use of AI. These include systemic discrimination, opaque decision-making, suppression of expression, and the spread of hate and disinformation. Through UN resolutions, reports, and expert discussions, a consensus is forming around the urgent need for a rights-based approach to AI governance. Yet, the realisation of this vision remains hindered by regulatory inertia, inadequate corporate accountability, and persistent myths about technological neutrality. Moving forward, international law must serve not only as a reference point but also as a binding framework to ensure that the development and use of AI uphold human dignity and equality.

b. Recommendations for Human Rights-Based AI Governance

The increasing deployment of AI in moderating digital content and shaping online interactions has profound implications for human rights. As highlighted in United Nations reports, effective governance of AI must be grounded in international human rights law and

focused on ensuring transparency, accountability, and fairness (A/RES/79/1, para 52). The starting point for a human rights-based approach is the reaffirmation of States' obligations under the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, particularly regarding the rights to freedom of expression (Article 19), privacy (Article 17), and non-discrimination (Articles 2 and 26). These rights are not suspended in digital spaces and must be protected in AI applications.

A key recommendation emerging from the UN documents is the need for robust human oversight over AI systems, especially those used for content moderation (A/RES/79/1, para 55(d)). While AI can be effective in detecting harmful content, it lacks the capacity to fully assess context, irony, or cultural nuances (A/73/348, para 29). Therefore, a hybrid model of automated moderation complemented by trained human moderators is necessary. Human oversight ensures that rights-infringing decisions made by algorithms can be reviewed and reversed, offering a critical safeguard against arbitrary or discriminatory content takedowns.

Transparency is another core principle for AI governance. *The Pact for the Future* stresses the urgency of enhancing transparency in algorithmic systems, particularly those managing content moderation and personal data (A/RES/79/1, para 36(a)). A rights-based approach requires that individuals be informed of when and how AI systems are making decisions that affect them. This includes transparency in how content is flagged, how user data is handled, and how AI-generated material is identified. Techniques such as labelling, watermarking, and authenticity certification (A/RES/79/1, para 36(c)) can empower users to distinguish between genuine and synthetic content, thus strengthening user agency and autonomy.

Accountability frameworks are essential to ensure that AI developers and digital platforms do not evade responsibility for rights violations. The UN General Assembly calls on companies to co-develop such frameworks in collaboration with States and civil society (A/RES/79/1, para 32(b)). These frameworks should include clearly defined obligations, reporting mechanisms, and avenues for redress. Importantly, auditability must be a core component. Platforms should publish periodic reports on AI system performance, including error rates and the impact of moderation decisions on vulnerable communities, as these groups are most at risk of being unfairly targeted (A/73/348, para 15).

In this regard, the implementation of redress mechanisms is critical. Victims of unjust AI-driven decisions must have access to effective remedies. This aligns with the UN's broader commitment to establishing risk mitigation and redress measures that also respect privacy and freedom of expression (A/RES/79/1, para 30). Companies should provide users with timely explanations of AI decisions and accessible procedures to challenge them. This could include appeals against content removals, account suspensions, or algorithmic rankings that suppress certain voices.

Another important solution is the promotion of inclusive and participatory governance models. Paragraph 31(c) of the Pact calls for regular collaboration among national institutions to share best practices (A/RES/79/1, para 31(c)). Extending this principle to the international

level, multi-stakeholder platforms should include underrepresented regions, especially from the Global South, in global AI standard-setting. This ensures that the norms guiding AI governance are not solely dictated by powerful private actors or developed States, but reflect diverse values and experiences.

There is also an urgent need to address the discriminatory risks embedded in training datasets and AI design. The UN reports underscore how biased data can lead to harmful outputs, as in the case of algorithms associating the word "Mexican" with negative connotations (A/73/348, para 15). Developers must implement fairness-by-design principles, such as diverse data curation, bias audits, and context-sensitive moderation tools. International guidance, like UNESCO's Recommendation on the Ethics of Artificial Intelligence offers a valuable blueprint for integrating ethical norms into AI system development (A/RES/79/1, para 52).

Finally, public digital literacy must be strengthened. As the UN documents suggest, users must be equipped to make informed choices and to provide or withdraw consent meaningfully (A/RES/79/1, para 36(a)). This calls for global initiatives to enhance digital education, especially in communities with limited access to information technology. Governments and platforms should support user awareness campaigns explaining how AI functions, what risks it poses, and how to navigate online environments safely and critically.

In sum, a human rights-based governance framework for AI must balance innovation with the protection of individual rights. This requires binding commitments to transparency, oversight, redress, inclusion, and ethical development. While technological efficiency is important, it cannot be pursued at the cost of fairness or justice. The UN resolutions provide a timely and coherent set of recommendations that, if operationalised effectively, can help States and corporations govern AI in a manner that genuinely serves humanity (see Figure 3).

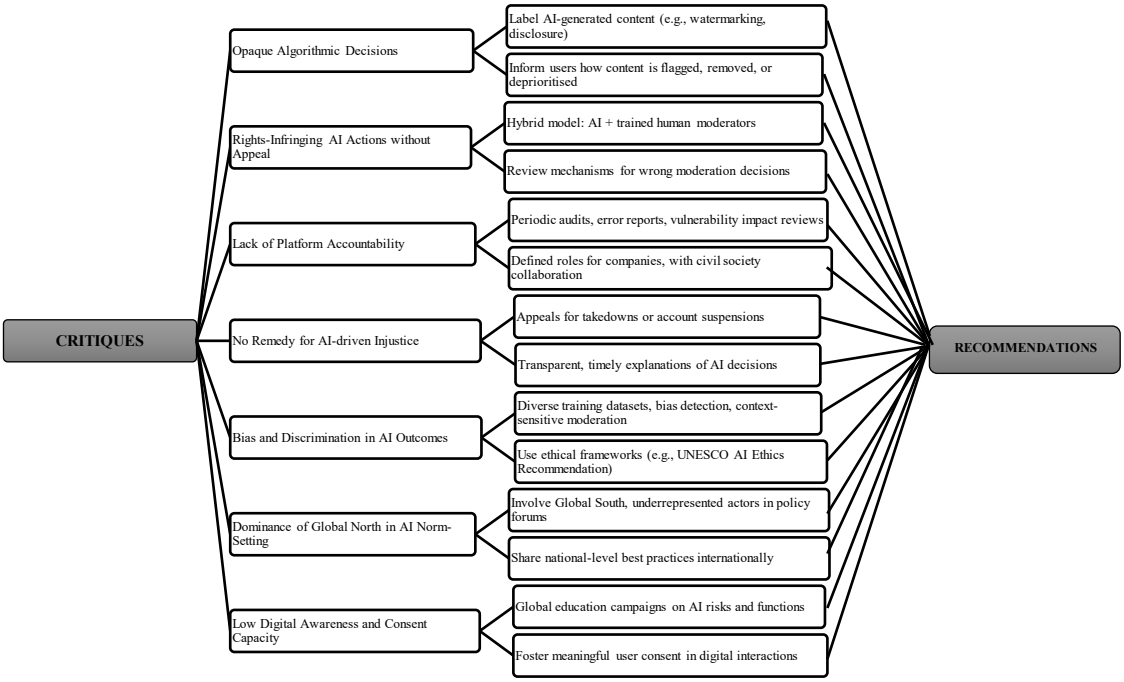

Figure 3. International Human Rights-Based Critiques and Recommendations for AI Governance on Freedom of Expression

Source: Processed UN Documents by the Author

CONCLUSION

As AI technologies increasingly mediate access to information and shape public discourse, governments, tech companies, and civil society should urgently reimagine digital governance frameworks through the lens of human rights. The unchecked expansion of AI filtering systems risks entrenching opaque, unaccountable, and exclusionary practices that distort the democratic function of online platforms. To prevent a future where silence is engineered and dissent invisibly erased, regulatory efforts should move beyond reactive fixes and commit to systemic safeguards rooted in legality, transparency, and inclusion. This calls not only for robust legal interventions, but for deeper interdisciplinary engagement that bridges technical design, ethical reasoning, and participatory policymaking. Future research should continue to scrutinise how AI is deployed, whose voices are being silenced, and what institutional architectures can ensure that freedom of expression survives—indeed thrives—in algorithmically governed spaces.

ACKNOWLEDGEMENT

The author would like to express his gratitude to ASTA Research Center for its invaluable support.

REFERENCES

Ahmed, Syed Hammad, Muhammad Junaid Khan, Hafiz Muhammad Umer Qaisar, and Gita Sukthankar. 2023. "Malicious or Benign? Towards Effective Content Moderation for Children's Videos." Edited by Franklin M. and Chun S.A. *The International FLAIRS Conference Proceedings* 36 (May). https://doi.org/10.32473/flairs.36.133315.

Ajani, Taiwo, and Tammy Ferrante. 2024. "Cyber-Analytics: An Examination of Machine Learning

Algorithms for Spam Filtering." *Issues in Information Systems* 25 (2): 203–13. https://doi.org/10.48009/2_iis_2024_116.

Akhil, Karnam, Jangili Sireesha, Gundu Venkata Sai, Katanguri Sai Shashidhar Reddy, and Jonnalagadda Haripriya. 2023. "Harnessing Artificial Intelligence for Preventing and Detecting Addiction in Digital Healthcare and Social Media among Students of Age Group 12 to 18." In *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 936–42. IEEE. https://doi.org/10.1109/ICAC3N60023.2023.10541714.

Al-Sherman, Naser. 2024. "Legal Protection from Artificial Intelligence Technology Used to Filter Visual Contents via the Internet." *Pakistan Journal of Criminology* 16 (1): 505–17. https://doi.org/10.62271/pjc.16.1.505.517.

Albraikan, Amani Abdulrahman, Mohammed Maray, Faiz Abdullah Alotaibi, Mrim M. Alnfiai, Arun Kumar, and Ahmed Sayed. 2023. "Bio-Inspired Artificial Intelligence with Natural Language Processing Based on Deceptive Content Detection in Social Networking." *Biomimetics* 8 (6). https://doi.org/10.3390/biomimetics8060449.

Alkiviadou, Natalie. 2022. "Artificial Intelligence and Online Hate Speech Moderation." *SUR: International Journal on Human Rights* 19 (32): 101–12. https://sur.conectas.org/en/artificial-intelligence-and-online-hate-speech-moderation/.

Ashraf, Cameran. 2022. "Exploring the Impacts of Artificial Intelligence on Freedom of Religion or Belief Online." *International Journal of Human Rights* 26 (5): 757–91. https://doi.org/10.1080/13642987.2021.1968376.

Babu, C V Suresh, and S Deva Darshini. 2025. *Detecting Hate Speech in the Digital Age: AI Solutions for Real-Time Moderation*. Edited by Swati Chakraborty. *Ethical AI Solutions for Addressing Social Media Influence and Hate Speech*. IGI Global. https://doi.org/10.4018/979-8-3693-9904-0.ch012.

Caruso, Corrado. 2025. "Towards the Institutions of Freedom: The European Public Discourse in the Digital Era." *German Law Journal* 26 (1): 114–37. https://doi.org/10.1017/glj.2024.68.

Chang, Bingtao, Weiping Wen, Xiaojie Wu, Siyang Cheng, Jianchun Jiang, and Rui Mei. 2025. "TCLens: Towards Toxicity Tags Aggregation of Massive Labels Generated by Content Moderation for AIGC." In *CSAI 2024 - Proceedings of 2024 8th International Conference on Computer Science and Artificial Intelligence*, 466 – 473. Association for Computing Machinery, Inc. https://doi.org/10.1145/3709026.3709114.

Elkin-Koren, Niva. 2020. "Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence." *Big Data & Society* 7 (2): 1–13. https://doi.org/10.1177/2053951720932296.

Elmimouni, Houda, Sarah Rüller, Konstantin Aal, Yarden Skop, Norah Abokhodair, Volker Wulf, and Peter Tolmie. 2025. "Exploring Algorithmic Resistance: Responses to Social Media Censorship in Activism." *Proceedings of the ACM on Human-Computer Interaction* 9 (2): 1–24. https://doi.org/10.1145/3710970.

Fahrudin, Adi, Siddhartha Paul Tiwari, and Kus Hanna Rahmi. 2025. "Psychological Well-Being of Human Content Moderators and Wellness Strategies in AI-Driven Content Moderation Organizations." In *Content Moderation in the Age of AI*, edited by Siddhartha Paul Tiwari, 221 – 249. IGI Global. https://doi.org/10.4018/979-8-3373-0335-2.ch009.

Frosio, Giancarlo. 2024. "Algorithmic Enforcement Tools: Governing Opacity with Due Process." In *Driv. Forensic Innovation in the 21st Century: Crossing the Valley of Death*, edited by Simona Francese and Roberto S. P. King, 195 – 218. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-56556-4_9.

Gambetti, Alessandro, and Qiwei Han. 2023. "Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media." In *International Conference on Information Systems, ICIS 2023: "Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies."* Association for Information Systems.

Gentelet, Karine, and Sarit K. Mizrahi. 2024. "A Human-Centered Approach to AI Governance: Operationalizing Human Rights through Citizen Participation." *Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users*, 215–30. https://doi.org/10.1201/9781003320791-24.

Ghose, Anuttama, Pallav Pallav, and S M Aamir Ali. 2025. "AI and Political Polarization: Analyzing the Impact of Algorithmic Bias on Governance in Diverse Political Systems." In *Economic and Political Consequences of AI: Managing Creative Destruction*, edited by Siddhartha Paul Tiwari, 135 – 159. IGI Global. https://doi.org/10.4018/979-8-3693-7036-0.ch007.

Gongane, Vaishali U, Mousami V Munot, and Alwin D Anuse. 2024. "A Survey of Explainable AI Techniques for Detection of Fake News and Hate Speech on Social Media Platforms." *Journal of Computational Social Science* 7 (1): 587 – 623. https://doi.org/10.1007/s42001-024-00248-9.

Goswami, Manasvin Veenu. 2020. "Algorithms and Freedom of Expression." In *The Cambridge Handbook of the Law of Algorithms*, edited by Woodrow Barfield, 558–78. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108680844.027.

Gunton, Kate. 2022. "The Use of Artificial Intelligence in Content Moderation in Countering Updates Violent Extremism on Social Media Platforms." In *Artificial Intelligence and National Security*, edited by Reza Montasari, 69 – 79. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-06709-9_4.

Hasimi, Lumbardha, and Aneta Poniszewska-Marańda. 2024. "Detection of Disinformation and Content Filtering Using Machine Learning: Implications to Human Rights and Freedom of Speech." In *CEUR Workshop Proceedings*, 3677:68–77. CEUR-WS. https://ceur-ws.org/Vol-3677/paper6.pdf.

Human Rights Watch. 2023. "Meta's Broken Promises Systemic Censorship of Palestine Content on Instagram and Facebook." *Human Rights Watch*, December 20, 2023. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and.

Khare, Pranjal, and Vishambhar Raghuwanshi. 2025. "Legal Frameworks Surrounding the Use of AI in Online Content Moderation." In *Ethical AI Solutions for Addressing Social Media Influence and Hate Speech*, edited by Swati Chakraborty, 235 – 260. IGI Global. https://doi.org/10.4018/979-8-3693-9904-0.ch011.

Lalchhanhima, J K, Vijayalaxmi Rajendran, and Aradhana Madhusudanan. 2025. "Reducing the Workload: AI as a Support Tool for Human Content Moderators." In *Content Moderation in the Age of AI*, edited by Siddhartha Paul Tiwari, 251 – 284. IGI Global. https://doi.org/10.4018/979-8-3373-0335-2.ch010.

Marsoof, Althaf, Andrés Luco, Harry Tan, and Shafiq Joty. 2023. "Content-Filtering AI Systems–Limitations, Challenges and Regulatory Approaches." *Information and Communications Technology Law* 32 (1): 64–101. https://doi.org/10.1080/13600834.2022.2078395.

Martínez, María Barral. 2023. "Platform Regulation, Content Moderation, and AI-Based Filtering Tools: Some Reflections from the European Union." *Journal of Intellectual Property, Information Technology and E-Commerce Law* 14 (1): 211–21.

McQue, Katia. 2024. "Revealed: US Police Prevented from Viewing Many Online Child Sexual Abuse Reports, Lawyers Say." *The Guardian*, January 17, 2024.

https://www.theguardian.com/technology/2024/jan/17/child-sexual-abuse-ai-moderator-police-meta-alphabet.

Parmar, Hemlata, and Utsav Krishan Murari. 2025. "Human-AI Synergy in Ethical Content Moderation: Navigating Fairness, Accountability, and Transparency Challenges." In *Ethical AI Solutions for Addressing Social Media Influence and Hate Speech*, edited by Swati Chakraborty, 191 – 212. IGI Global. https://doi.org/10.4018/979-8-3693-9904-0.ch009.

Paul, Kari. 2024. "Meta Struggles with Moderation in Hebrew, According to Ex-Employee and Internal Documents." *The Guardian*, August 15, 2024. https://www.theguardian.com/technology/article/2024/aug/15/meta-content-moderation-hebrew.

Penney, Jonathon W. 2020. *Measuring Surveillance Chill and Other Regulatory Impacts at Scale*. Edited by Ryan Whalen. *Computational Legal Studies: The Promise and Challenge of Data-Driven Research*. Edward Elgar Publishing Ltd. https://doi.org/10.4337/9781788977456.00012.

Perilo, Michel, and George Valença. 2024. "How Facial Recognition Technologies Affect the Transgender Community? A Systematic Mapping Study." In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 1153–60. New York, NY, USA: ACM. https://doi.org/10.1145/3605098.3635935.

Poredi, Nihal, Deeraj Nagothu, and Yu Chen. 2024. "Authenticating AI-Generated Social Media Images Using Frequency Domain Analysis." In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, 534–39. IEEE. https://doi.org/10.1109/CCNC51664.2024.10454640.

Rayhan, Tapu, Ayesha Siddika, Mehedi Hasan, and Nafisa Sultana Elme. 2024. "Social Media Emotion Detection and Analysis System Using Cutting-Edge Artificial Intelligence Techniques." In *Lecture Notes in Networks and Systems*, edited by Xin-She Yang, R. Simon Sherratt, Nilanjan Dey, and Amit Joshi, 1000 LNNS:501 – 514. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-97-3289-0_40.

Sitabuana, Tundjung Herning, Ramalina Ranaivo Mikea Manitra, Dixon Sanjaya, Ibra Fulenzi Amri, and Nethan. 2024. "The Urgency of Artificial Intelligence Code of Ethics." *Indonesia Law Review* 14 (3): 25. https://doi.org/10.15742/ilrev.v14n3.6.

Subrahmanyam, Satya. 2025. "Psychological Impact of AI Moderation." In *Content Moderation in the Age of AI*, edited by Siddhartha Paul Tiwari, 191 – 219. IGI Global. https://doi.org/10.4018/979-8-3373-0335-2.ch008.

Tiwari, Siddhartha Paul, and Adi Fahrudin. 2025. *Content Moderation in the Age of AI*. *Content Moderation in the Age of AI*. IGI Global. https://doi.org/10.4018/979-8-3373-0335-2.

United Nations Human Rights Council. (2023, July 14). *Resolution adopted by the Human Rights Council on 14 July 2023: New and emerging digital technologies and human rights (A/HRC/RES/53/29)*. United Nations. https://digitallibrary.un.org/record/4020206/files/A_HRC_RES_53_29-EN.pdf.

United Nations Human Rights Council. (2024, June 3). *Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance (A/HRC/56/68)*. United Nations. https://docs.un.org/en/A/HRC/56/68.

United Nations Human Rights Council. (2025, April). *Annual report of the United Nations High Commissioner for Human Rights (A/HRC/59/28)*. United Nations. https://docs.un.org/en/A/HRC/59/28.

United Nations General Assembly. (2024, September 22). *Resolution adopted by the General Assembly on 22 September 2024: Pact for the Future (A/RES/79/1)*. United Nations. https://docs.un.org/en/A/RES/79/1.

United Nations General Assembly. (2018, August 29). *Promotion and protection of the right to freedom of opinion and expression: Note by the Secretary-General (A/73/348)*. United Nations. https://docs.un.org/en/A/73/348.

Vacarelu, Marius. 2023. "Malicious Use of Artificial Intelligence in Political Campaigns: Challenges for International Psychological Security for the Next Decades." In *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, edited by Evgeny Pashentsev, 203 – 230. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9_8.

Wang, Erika Ningxin. 2024. "Participatory Censorship With Illusory Empowerment: Algorithmic Folklore and Interpretive Labor beyond Fandom." *Social Media and Society* 10 (4). https://doi.org/10.1177/20563051241295800.