CONTENT MODERATION ON SOCIAL MEDIA: A SHIELD AGAINST HARM OR A TOOL FOR CENSORSHIP? AN EMPIRICAL STUDY THROUGH USER PERCEPTION

Falak Naz, KIIT School of Law

ABSTRACT

We conduct our lives online, especially socially. But what happens when social media companies tell us what we can do and say? Are they looking out for us from harm while dictating speech, or are they working to maintain control over our speech? Social media has become a part of our everyday lives via socializing, debating, discussing, and sharing information. Social media users are left with questions regarding the fairness of speech moderation decisions, how and why the decisions are made, and whether everyone is treated equally. Social media companies moderate, eliminate, and/or filter posts they deem harmful or misleading, thereby regulating what is shared online. This paper explores continuity, or dissatisfaction, from the social media user perspective regarding online content moderation and asks, is it a trust-based safeguard against harmful content in online communities, or an outcome of unwarranted censorship due to moderating to the limits of free speech and expression? The study examines how users view content moderation, their trust in moderation decisions, and how they make sense of the rationale for content take-down or restriction using empirical methods. Finally, it explains critical topics regarding the moderation of online speech including how to best reach a balanced approach between facilitating open forms of communication and safety of members of online communities in an ever-changing digital landscape.

Keywords: Content Moderation, Censorship, Free Speech and Expression, Social media users, Harmful content, User Perception

INTRODUCTION:

In the past few years, India has seen an increase in the use of the social media platforms. The current generations highly rely on the social media platforms for their daily activity. Whether it is about posting about your daily life or commenting on someone's life this generation is highly dependent on these platforms because of which social media platforms like Facebook, YouTube, Instagram, Twitter, etc. have become critical platforms for public debates these days. Social media platforms have become an essential part of our lives. Whether we have to convey our opinions or comment on something these platforms have emerged as easily accessible platforms where we can reach out to billions of people worldwide. But with the growing use of these platforms concerns regarding content moderation are also increasing. These platforms use "norm settings"[1] to filter out harmful content such as hate speech and misinformation etc. through proactive and active moderation techniques. But the question is, are these moderations an attempt to protect society from potential harm or just a way to suppress free speech and expression? Is this governance or Dictatorship? These questions make the uniformity and transparency of these moderation rules debatable.

The "Information Technology (Guidelines for Intermediaries and Digital Media Ethics Code) Rules," which were brought in 2021, have made some significant changes with respect to India's legal framework for regulating digital platforms. The Rules prohibit users from creating, publishing, or disseminating any content that violates copyright or patents, is pornographic, contains software viruses, or jeopardizes public order or the unity of India. But the question is does it is to control internet content by maintaining a balance between the freedom of speech guaranteed by Article 19(1)(a) and the right to reasonable restrictions under Article 19(2) or not. This raises questions concerning censorship, free speech, and the opaqueness of content moderation policies.

This paper aims to examine user's perceptions of content moderation as to whether they consider it necessary to promote online safety or restrict free speech and expression. Using an empirical methodology based on questionnaires, the paper examines the effects of moderation policies on users and their opinions regarding digital censorship. The findings will support discussions related to social media governance and digital rights.

[1] Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech.*, *17*, 42.

LITERATURE REVIEW:

Langvardt, K. (2017)[2]:

Langvardt, K. (2017) discussed how social media platforms, which are governed by private companies and have their own set of rules, have developed into important forums for public debate. To decide what information is allowed or prohibited, these platforms usually employ vague and conflicting rules, which raises serious concerns about justice and freedom of speech. It makes the case that it is harmful to leave content moderation only up to private platforms since it provides a selected few companies with excessive control over free speech. The article highlights the different ways that government content moderation could be implemented, from trying to put specific legal restrictions on moderation itself to mandating platforms to be more open about their moderation procedures or giving users more control over what they view. However, it also argues that if government regulation is not done wisely, it may jeopardize free speech.

While acknowledging how harmful content might spread in the absence of content monitoring, this article also criticizes the uncontrolled power of private companies. The paper suggests some regulatory strategies such as mandatory moderation limitations, user-controlled filtering, and content policy transparency.

The paper view content moderation as a legal or regulatory matter, but fails to consider the matter from the viewpoint of the user. There hasn't been much research done to determine whether users view moderation as a kind of censorship or as a helpful safeguard against harmful content. More research may be done on the effects of user-driven moderation settings and content moderation on digital free speech.

Gerrard, Y. (2018)[3]:

This paper critically examines the efficacy of social media content moderation strategies, particularly the use of hashtag bans to restrict harmful content. ***Gerrard, Y. (2018)*** examines how users actively avoid moderation by communicating secretly through the use of visual cues, untagged posts, and coded language in pro-eating disorder (pro-ED) communities. It also

[2] Langvardt, K. (2017). Regulating online content moderation. *Geo. LJ*, *106*, 1353.

[3] Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492-4511.

highlights a serious contradiction in platform policies. Even though social media companies publicly state that they restrict harmful content, their recommendation algorithms often continue to support it.

The paper argues that social media content moderation initiatives are mainly unsuccessful because they place an undue emphasis on hashtag bans, which are simple to monitor but fail to take into consideration alternative content distribution channels. By altering terms that have been flagged, joining private groups, and utilizing platform algorithms that inadvertently magnify restricted content, users discover alternate means of disseminating information. This highlights a significant inconsistency. Although platforms openly implement moderation guidelines, their automated systems often suggest the very material they purport to control. The study highlights that in order for moderation to be effective, it must take into account user's resistance to and adaptation to these limitations in addition to hashtags.

Despite providing important insights into how users overcome content moderation, the paper ignores the broader discussion of whether these regulations serve as a safeguard or a means of censorship.

Sander, B. (2019)[4]:

By selecting what content is kept or deleted, social media platforms influence online conversation. Moderation raises questions about censorship even though its goal is to avoid harm. ***Sander, B. (2019)*** supports a human rights-based strategy that aligns moderation with the fairness and transparency requirements of Article 19 of the ICCPR. Yet, corporate interests, laws, financial incentives, and public pressure all influence governance, which results in uneven application of the law and ambiguous decision-making. Even though automated moderation is effective, it frequently lacks context and perpetuates bias, especially when it comes to politically sensitive content.

The paper emphasizes the difficulties in striking a balance between accountability and free speech. Algorithmic moderation disproportionately affects marginalized groups, appeals processes are insufficient, and policies are applied selectively on platforms. Because there is no independent oversight, platforms are susceptible to arbitrary censorship and political

[4] Sander, B. (2019). Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation. *Fordham Int'l LJ*, *43*, 939.

influence. Better governance is recommended by human rights frameworks such as the UNGPs, but due to a lack of enforcement tools, actual implementation is still lacking.

There are still a lot of unanswered questions, especially regarding how users perceive moderation and whether it censors or protects speech. Different group's experiences with moderation, trust in AI-driven decisions, and alternative, user-driven models that enable democratic content governance are all overlooked in this paper.

Brown, N. I. (2020)[5]:

As social media platforms increasingly control what content stays online, the discussion surrounding content moderation and freedom of expression has heated up. ***Brown, N. I. (2020)*** contends that the current self-regulation model is inconsistent, opaque, and unaccountable, which undermines user trust even though moderation is crucial to preventing hate speech, false information, and damaging content. Platforms put corporate interests first, applying rules arbitrarily and without adequate oversight. Although regulation by the government is frequently suggested as a remedy, it runs the risk of censorship and overreach. Therefore, a new strategy is required to maintain digital free speech while ensuring equitable and efficient content moderation.

The paper assesses three content regulation models: industry-wide governance, government intervention, and self-regulation. The current system of self-regulation gives platforms autonomy, but it has resulted in uneven enforcement and skewed decision-making. Although government regulation provides more stringent oversight, it also raises questions about political meddling.

The study promotes an industry-wide governance model in which platforms work together to create uniform, transparent regulations supported by impartial oversight organizations. This model ensures fair enforcement without undue state control by striking a balance between the protection of free speech and expression and the necessity for responsible moderation. There are still significant gaps in content moderation policies despite much discussion. First, little research has been done on how users view moderation, whether they believe it to be a

[5] Brown, N. I. (2020). Regulatory goldilocks: finding the just and right fit for content moderation on social platforms. *Tex. A&M L. Rev.*, *8*, 451.

trustworthy process or see it as censorship. Second, little is known about how moderation affects various user groups, especially when it comes to bias in automated systems.

Wilson, R. A., & Land, M. K. (2020)[6]:

Governments have historically been in charge of controlling public discourse, but social media platforms have taken over this responsibility. In their role as private speech regulators, ***Wilson and Land (2020)*** contend that these platforms enforce content moderation guidelines that frequently depend on automated systems and expansive definitions of hate speech. Although the goal of these regulations is to avoid harm, they usually result in excessive censorship and the repression of free speech. The paper examines platform regulations and draws attention to the lack of transparency and consistency in the application of moderation rules.

The significant issue with content moderation is the over-reliance on AI-driven moderation, which often ignores context and misclassifies political speech, satire, and marginalized voices. By giving actual examples where hate speech on the internet has devolved into violence, the paper demonstrates the real-world consequences of inadequate moderation. Wilson and Land argue that a more nuanced approach that considers social and political context is better than strict automated enforcement

Important research gaps persist despite in-depth discussions of hate speech online. Evidence regarding user's perceptions of content moderation specifically, whether they view it as censorship or protection is scarce. There is also a dearth of research on bias in AI moderation and how it affects various social and political groups. Furthermore, more research is required on alternative, context-sensitive moderation models that transcend automated enforcement. To create a content moderation system that is equitable, open, and efficient, these gaps must be filled.

Bloch-Wehba, H. (2021)[7]:

Bloch-Wehba (2021) contends that law enforcement is increasingly influencing content moderation through both official and informal partnerships, despite the fact that social media

[6] Wilson, R. A., & Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, *52*, 1029.

[7] Bloch-Wehba, H. (2021). Content moderation as surveillance. *Berkeley Technology Law Journal*, *36*(3), 1297-1340.

platforms have grown to be important online speech regulators. Despite their seeming independence, platform's moderation practices typically mirror those of the government, which raises questions about oversight flaws and state censorship.

The paper emphasizes how government goals are enforced by AI-driven moderation tools, which results in excessive censorship, biased enforcement, and the repression of opposition. Platforms also assist law enforcement by supplying user data for surveillance, which turns them from impartial content adjudicators into active policing members. The concept of platform autonomy is called into question by this extensive incorporation of law enforcement into digital governance.

Important gaps still exist despite conversations about content moderation and surveillance. The ways in which government priorities shape moderation bias, how law enforcement affects platform policies, and how to set up oversight procedures to stop state-driven censorship have not all been thoroughly studied in research. To ensure independent and transparent content governance, these gaps must be filled.

Díaz, Á., & Hecht-Felella, L. (2021)[8]:

Social media companies have inconsistent content moderation policies, frequently over-policing underprivileged groups while giving influential users more leeway. According to ***Díaz and Hecht-Felella (2021)***, corporate and political interests influence moderation policies on harassment, hate speech, and terrorism, which results in unfair enforcement and double standards.

The paper draws attention to biases in moderation caused by AI, which commonly misinterpret political speech and activism, leading to excessive censorship. Additionally, platforms provide little redress for unjust takedowns and lack transparency in enforcement and appeals. Moderation is further shaped by corporate and governmental influence, which puts political and commercial interests ahead of human rights safeguards.

There are still significant gaps in the research on content moderation. Evidence on how AI biases influence online discourse, how marginalized communities experience moderation, and

[8] Díaz, Á., & Hecht-Felella, L. (2021). Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law*, 1-23.

how independent oversight could enhance fairness is lacking in research. Closing these gaps is essential to guaranteeing fair and transparent content governance.

Gillespie, T. (2022)[9]:

As a moderation technique, social media platforms are increasingly using content reduction, which limits the visibility of content rather than eliminating it. According to ***Gillespie (2022)***, this approach lacks accountability and transparency but enables platforms to suppress false information and speech that crosses the line into censorship without resorting to complete censorship. In contrast to removals, which notify users, content reduction works covertly, influencing public opinion without the user's knowledge. Algorithms effectively moderate speech while evading regulatory scrutiny by down-ranking content, restricting recommendations, and suppressing engagement.

Without explicit guidelines or user recourse, the study demonstrates how platforms such as Facebook and YouTube employ reduction techniques to handle content that is deemed problematic. Despite being meant to stop harm, this practice raises questions about fairness and bias because platforms put political and commercial interests ahead of free speech. Because there are no appeal procedures, moderation is less accountable, and users frequently do not realize that their content is being blocked. Attempts to guarantee equitable and open moderation policies are made more difficult by the dependence on opaque algorithms.

There are still a lot of gaps in spite of its increasing use. The lack of appeal mechanisms for suppressed content, user awareness of content reduction, and its effects on speech patterns and engagement are all unexplored. To provide open, equitable, and responsible content moderation that strikes a balance between harm prevention and digital rights, these gaps must be filled.

Jhaver, S., Zhang, A. Q., Chen, Q. Z., Natarajan, N., Wang, R., & Zhang, A. X. (2023)[10]:

To give users more control over their online experience, social media platforms are depending more and more on user-controlled moderation tools like toxicity sliders and word filters. According to ***Jhaver et al. (2023)***, who look at how users feel about these tools, although they

[9] Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media+ Society*, *8*(3), 20563051221117552.

[10] Jhaver, S., Zhang, A. Q., Chen, Q. Z., Natarajan, N., Wang, R., & Zhang, A. X. (2023). Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1-33.

value control, users find it difficult to deal with complicated interfaces, ambiguous categories, and the mental strain of setting things up. This raises questions about the practical effectiveness of these tools and whether moderation should be the responsibility of the platform rather than the user.

The paper draws attention to issues with transparency and usability because many users are reluctant to filter content out of concern that they will miss crucial conversations. Furthermore, cultural and contextual variations might not be taken into consideration by moderation tools, which could result in inefficient filtering or unintentional censorship. Since there is no concrete proof that these tools enhance online interactions, many users believe that moderation takes too much work and transfers accountability from platforms to individuals.

Personalized moderation is becoming more popular, but there are still significant gaps. Few studies have been conducted on how users cope with decision fatigue, whether user-driven moderation truly lowers harmful content, and the effects of prolonged use of these tools on discourse, polarization, and content visibility. To ensure equitable, efficient, and user-friendly moderation systems that strike a balance between accountability, control, and usability, these gaps must be filled.

Jain, S., Panjal, P., Sultan Ahmad, M., Ahmad Siddiqi, R., & Seth, A. (2024, December)[11]:

Community-driven moderation is an alternative to AI-based systems in voice-based participatory media platforms. Content moderation on rural Indian platforms, where local moderators enforce policies based on community norms, is examined by ***Jain et al. (2024)***. This guarantees cultural relevance, but it also brings about social bias, political pressure, and uneven enforcement, which raises questions about justice and possible censorship.

With moderators subject to outside influences, the study emphasizes how social and political contexts influence moderation decisions. Participation from users increases engagement, but it also brings up issues of bias and subjectivity. Community-driven systems might be impacted by social hierarchies, which could have an impact on free speech, in contrast to AI moderation,

[11] Jain, S., Panjal, P., Sultan Ahmad, M., Ahmad Siddiqi, R., & Seth, A. (2024, December). Filter-in or Filter-out: Complexities and Implications of Content Moderation Processes in a Voice-based Participatory Media Platform in India. In *Proceedings of the 13th International Conference on Information & Communication Technologies and Development* (pp. 84-109).

which lacks nuance.

Despite its advantages, there are still significant gaps. Research is required into how power dynamics influence content choices, whether users perceive community moderation as fair or censorship, and its long-term effects on self-censorship and free speech. To determine whether community-driven moderation promotes equity or creates new biases, it is imperative to fill in these gaps.

RESEARCH GAP:

1. **Lack of focus on user perception of content moderation:**

The existing studies mainly focus on the platform governance and legal frameworks related to content moderation but there is little or no research on how ordinary users perceive the content moderation policies and how they impact their online activity.

2. **Transparency and awareness gap in the moderation Process:**

The existing studies highlight the opacity of content moderation policies, but there is a lack of focus on how well the users understand these policies, whether they are aware of the reasons why their contents are moderated, whether there is transparency in these moderation policies, whether they perceive them as just and fair for safety from potential harm or simply an unreasonable restriction on free speech and expression.

RESEARCH OBJECTIVE:

To examine how user's perceive social media content moderation, primarily whether they consider it to be an essential safeguard against harmful content or an unjust restriction on the right to free speech and expression.

RESEARCH QUESTIONS:

1. How do social media users consider the fairness and transparency of content moderation?
2. What factors influence user's trust or mistrust of content moderation, and how much do they trust them?

RESEARCH METHODOLOGY:

The objective of this paper is to examine how user's perceive social media content moderation, primarily whether they consider it to be an essential safeguard against harmful content or an unjust restriction on the right to free speech and expression. To answer the research questions, the paper uses the quantitative method to collect data. The data was collected through an online questionnaire, which was circulated among 100 active social media users about their perception of the fairness and transparency of content moderation, as well as the factors that influence their trust or mistrust in these practices out of which 80 responses were collected. The questionnaire includes questions that were designed to understand user's perception of content moderation fairness and transparency, and trust levels, as well as their experiences and opinions. Additionally, the paper uses the doctrinal method to provide a better understanding of content moderation and Freedom of expression as safeguarded under the Indian Constitution and ICCPR, which will help provide a clear understanding of how content moderation shape user's perceptions and influence their trust in moderation practices.

CONTENT MODERATION:

Social media platforms like Facebook, Instagram, and Twitter etc. have developed into vital platforms for public debate and political participation. Because of their broad appeal, global reach, and user-friendliness, they serve as a platform for news consumption for many, while for others, they are appropriated as platforms for community organising, mobilisation, and collective action. The content shared on social media platforms often leads to community outrage which makes it more necessary to through content moderation process.

The increase in the use of Social media platforms has made people more reliant on these platforms where they share their opinions, disagreements etc. through post which often go through content moderation process based on the nature of the content. People have different opinions regarding these moderation policies and this has made it debatable in recent years. In recent years, there have been more public arguments about what should be allowed on social media platforms and what shouldn't.[12]

[12] Elmimouni, H., Skop, Y., Abokhodair, N., Rüller, S., Aal, K., Weibert, A., ... & Tolmie, P. (2024). Shielding or silencing?: An investigation into content moderation during the sheikh jarrah crisis. *Proceedings of the ACM on Human-Computer Interaction*, *8*(GROUP), 1-21.

The process of checking online user-generated content for adherence to a digital platform's rules about what can and cannot be published on the platform is known as content moderation. The process of moderating content and enforcing rules is either done manually by people or by automation, or a combination of both, depending on the magnitude and maturity of the abuse and of a platform's operations.

Social media companies claim that the purpose of their content moderation guidelines is to guarantee the security and reliability of user-generated content that is made publicly available. But this moderation process has always been on a clash with the freedom of expression guaranteed under the Indian Constitution.

Concerns over content moderation policies on various social media platforms have been raised by various organisations, Scholars, and General users, and there is a continuing public discussion concerning the practice's limitations and blind spots. Some asserts that "users of social media are subject to a regime of private censorship that was only recently unimaginable" and that platforms make "erratic and opaque" judgements[13]. Additionally, platforms may contribute to state-led censorship because they lack explicit, open policies that explain to users how they respond to requests from government officials to delete content. Some scholars contend that excessive moderation enforcement can be construed as censorship. For instance, Gillespie states that punitive measures "risk(s) limiting or chilling critical expression." He goes on to say that, "removing content or users is analogous to the most severe sort of censorship to the extent that content moderation is somewhat akin to censorship. Eliminating users from platforms not only restricts their speech but also prevents them from participating on that platform and may even prevent them from expressing in the future[14]. Although the moderation policies of various platforms vary, the majority of the moderation work on all of them is still conducted in an opaque and commercial manner[15]. Moderation techniques include blocking and deleting accounts, deleting specific content, and more covert methods of content devaluation, such as shadow banning, which are more difficult for users to detect or verify. From a legal and regulatory perspective, these practices are extremely problematic since they are opaque and so limiting to users, who are unable to challenge or protest them. People often claim that these platforms are biased, acts with political bias and favours the powerful like the

[13] Langvardt, K. (2017). Regulating online content moderation. *Geo. LJ*, *106*, 1353.

[14] Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[15] Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work.

politicians, Governments etc. while targeting the minorities, Content Creators, Activists etc. And these claims and debates makes the content moderation policies questionable as to whether they are necessary to protect against or a tool to censorship or a tool supress the freedom of speech and Expression.

FREEDOM SPEECH AND EXPRESSION:

Article 19(1)(a)[16] of the Indian Constitution1 protects the freedom of speech and expression as one of the fundamental rights. As the foundation of a democratic country, it allows people to freely express their ideas, beliefs, and thoughts without worrying about punishment but it is not an absolute right, the state can impose reasonable restrictions in the interests of the sovereignty and integrity of India, security of the State, friendly relations with foreign States, public order, decency or morality, in relation to contempt of court, defamation, or incitement to an offence and these restrictions are outlined in Article 19(2) of the Indian Constitution. The right to free expression has changed significantly over history. Be it in the form of of laws, social conventions, or technical developments. The evolution of freedom of speech was significantly influenced by technological advancement. The Internet is one of the most essential tool. The majority of people use the internet in a revolutionary way to communicate their thoughts about anything, including ideas about government programmes or anything that is growing more potent and has excellent connectivity.

People can share their thoughts online due to the freedom to free speech and expression, which can significantly help educate the public and provide them a genuine picture of any events happening on both inside and outside of India. However, there are challenges associated with the digital world. Issues like hate speech, misinformation, and online harassment have surfaced due to the ease of content dissemination and its rapid spread. Due to its global reach, the Internet has brought up particularly difficult legal and regulatory challenges as governments and Internet service providers struggle to find a balance between protecting users' right to free expression and protecting them from harmful and abusive content. The Indian Constitution's Article 19(1)(a) protects the right to free speech and expression. In the modern day, the Internet is essential to expand the citizen's freedom of speech.[17]

[16] India Const. art. 19

[17] Kumar, A. (2022). Freedom of Speech in India and Outside: Internet's Unyielding Influence. *Jus Corpus LJ*, *3*, 258.

The right to freedom of thought and expression is guaranteed under article 19 of the International Covenant on Civil and Political Rights (ICCPR), which uses the same general language as the corresponding article in the Universal Declaration of Human Rights (UDHR). All democracies require freedom of expression, which is regarded as a fundamental right on a global scale. Article 19(1) of the ICCPR protects freedom of opinion, while Article 19(2) of the ICCPR specifically mentions the right to freedom of speech. which states:

"Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice."[18]

SOCIAL MEDIA AND FREEDOM OF SPEECH:

The term "social media" denotes a group of websites and applications that are mainly used for content sharing and communication[19] and the expression, "freedom of speech" refers to the right to express one's thoughts, feelings, and opinions in public or on a public forum without worrying about repercussions.

Since the social media platforms has been growing rapidly over the past years, there is a strong correlation between social media and freedom of speech. There are no filters on any of the social networking sites, they are all free. Users are fully in charge of their content and are free to share any kind of content they like. Social media platforms also provide people a voice and the impression that if the government or authorities don't listen to their problems, they can use social media as a platform for free speech to publicly voice their concerns and expose the relevant authorities. Social media is considered as the strongest source of original expression[20]. Social media's accessibility and worldwide reach are its greatest advantages. Citizens are not granted absolute power when it comes to the right to free expression. Even while every nation in the world guarantees the right to free expression, this does not mean that it has ultimate authority because there are situations in which this right can be curtailed and is subject to reasonable restrictions. Many nations throughout the world have acknowledged social media

[18] ICCPR art. 19(2)

[19] Hanna, K. T. (2021, September). What Is Social Media? (B. Lutkevich, Ed.). Techtarget; Techtarget. https://www.techtarget.com/whatis/definition/social-media

[20] SeventhQueen. (2020, August 7). Role of Social Media and Freedom of Speech and Expression. Legal Desire. https://legaldesire.com/role-of-social-media-and-freedom-of-speech-and-expression

as a source of freedom of speech and as a fundamental human right, in light of the expanding usage and accessibility of the Internet and social media[21]. Numerous campaigns and movements, has gained international public support after beginning on social media. However, freedom of speech and expression in digital also have some challenges.

ACCOUNTABILITY AND TRANSPARENCY:

The principles of accountability and transparency is necessary for advancement and protection of human rights. State actors should be held accountable for their decisions and actions and should act transparently, according to the fundamental concepts of accountability and transparency. Democracy cannot exist without these ideas and to uphold and advance democracy, human rights, and the rule of law, the Human Rights Council (HRC) has highlighted "the importance of effective, transparent, and accountable legislative bodies, and acknowledges their fundamental role."[22]. Freedom of expression is a prerequisite for the fulfilment of these ideals. Another way to look at the right is as a fundamental component of society. They are essential components of every free and democratic society, together with the freedom of opinion. The two freedoms are interconnected, with the right to free speech offering the opportunity for opinion development and exchange[23]. It has been believed that exercising one's right to free speech is necessary in order to participate in the democratic process. Additionally, the ICCPR mandates State parties to uphold and protect the right to freedom of expression.[24] Social media platforms impact on freedom of expression is very broad-based[25]. Social media platforms bring many distinct legal challenges, and the debate is flourishing around issues like the scope of protections for speech online, and how human rights law generally can be enforced in an online environment.

[21] Social Media And Freedom of Speech And Expression. Legalserviceindia.com. https://www.legalserviceindia.com/legal/article-426-social-media-and-freedom-of-speech-and-expression.html#google_vignette

[22] Human Rights Council Nineteenth session Agenda item 3 Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development Resolution adopted by the Human Rights Council* 19/36 Human rights, democracy and the rule of law. Available at https://digitallibrary.un.org/record/725358/files/A_HRC_RES_19_36-EN.pdf

[23] Human Rights Committee. (2011). Human Rights Committee 102nd session General comment No. 34. https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf.para 2

[24] Human Rights Committee. (2011). Human Rights Committee 102nd session General comment No. 34. https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf.paras 7 and 11

[25] Koltay, A. (2021). The protection of freedom of expression from social media platforms. *Mercer L. Rev.*, *73*, 523.

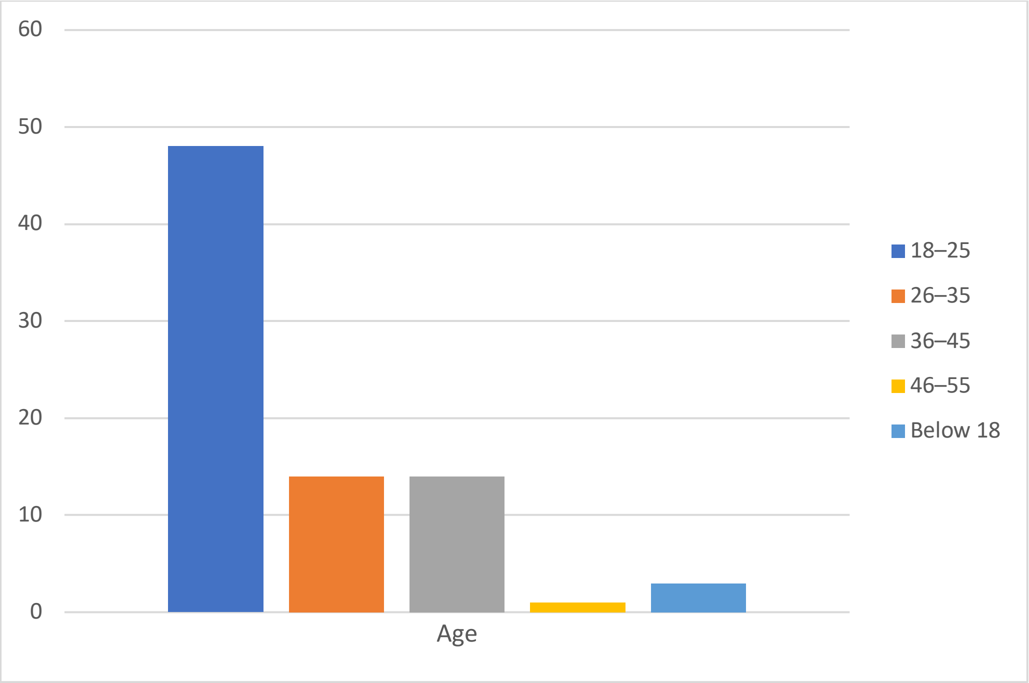
The three-part cumulative test on restrictions, outlined in Article 19(3) ICCPR[26], must be passed before restricting the right to free expression online. Restrictions are only acceptable for actors like social media sites if they pass the three-part criteria, which includes the robust suggestions made in the UNGP. Since social media platforms are necessary for the free flow of information across society, restrictions that do not conform would consequently be considered to violate Article 19.

FINDINGS:

A survey was conducted through a questionnaire which was circulated among 100 individuals of all age groups, out of which 80 responses were successfully received.

1. DEMOGRAPHY

Age

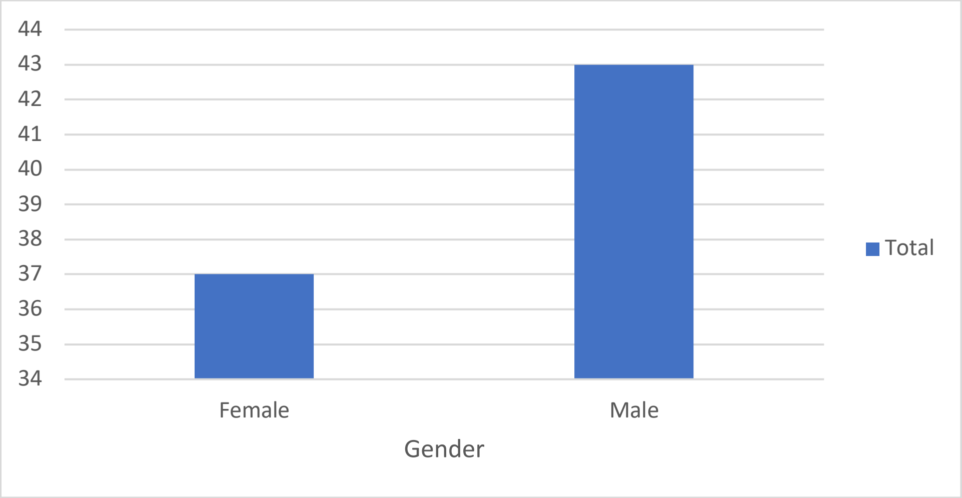


The findings show that the participants of 18-25 age group are the most prominent which suggests the engagement of younger individuals on social media platforms. There is also a

[26] Human Rights Council Seventeenth session Agenda item 3 Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. (2011). https://www.icnl.org/wp-content/uploads/Transnational_opinionexpression.pdf.para 68

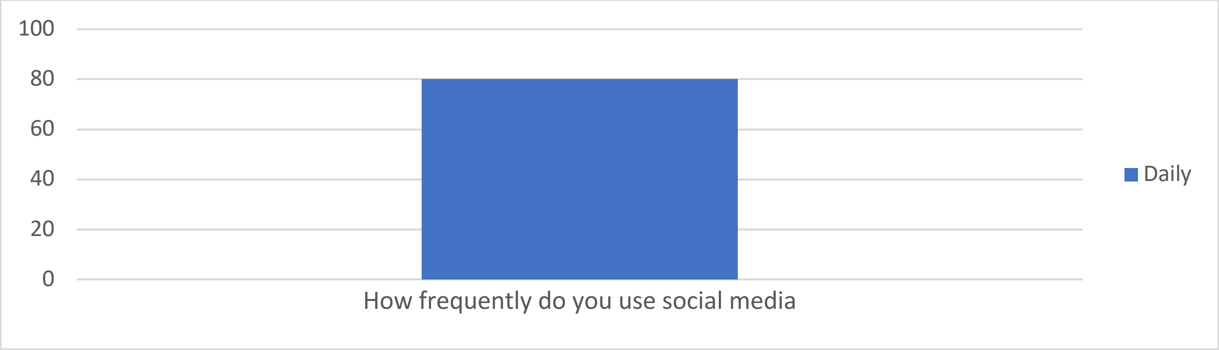
decent representation of individuals in the 26-35 age group which indicates that some people in this age range are also included in the data analysis. Fewer individuals are in the age group of 36-35, 46-55 or below 18 age group, which may suggest that the study is focused on young adults.

Gender



The findings show that out of 80 respondents, 43 are male respondents, whereas 37 are female respondents. This indicates a slightly higher representation of males in the sample, with males making up 53.75% of the respondents and females making up 46.25%.

2. SOCIAL MEDIA USAGE

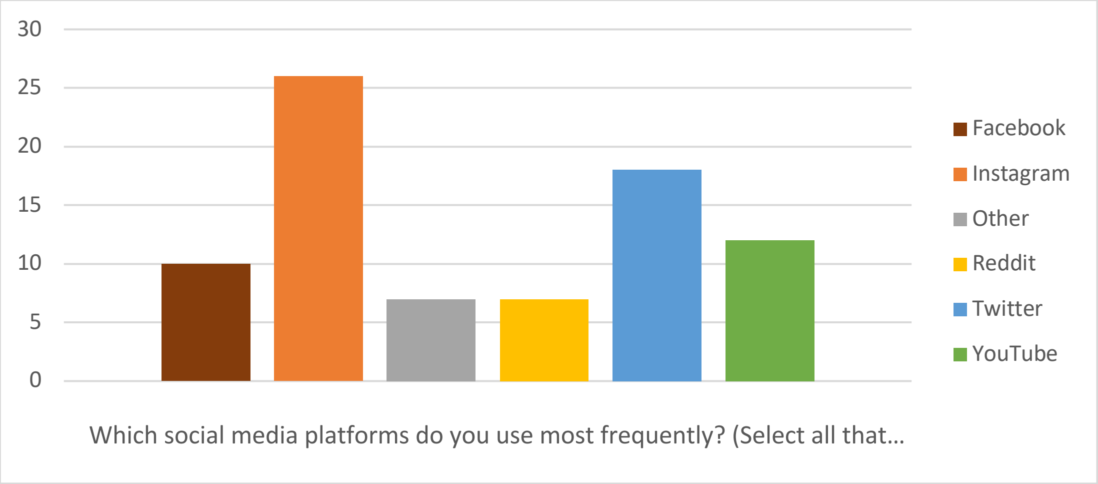


The findings show that all 80 participants use social media on a daily basis which indicates a high level of engagement with digital platforms among the population.

3. PLATFORM AND AWARENESS

The participants were asked which social media platform they use most frequently along with awareness that is whether they are aware of the content moderation practices or not.

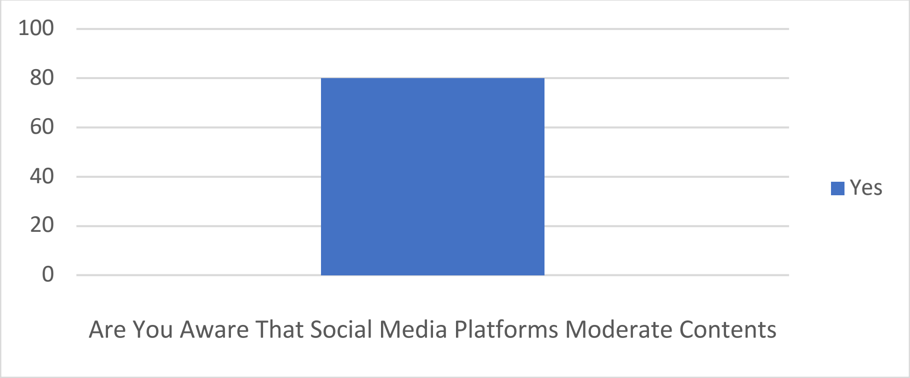
Most Frequently Used Platforms



The findings show that Instagram is most frequently used platform among the participants followed by Twitter which is the second most used platform. Facebook is used moderately other platforms like Reddit and others are less frequently used but still notable.

It shows that visual and discussion driven platforms like Instagram and Twitter are leading in terms of engagement among users.

Awareness of Content Moderation Practices



Facebook Users: Findings shows that all Facebook users are aware of content moderation which indicates that users on this platform are well-informed about how posts and online activities are monitored and controlled.

Instagram Users: Findings shows that all Instagram users are also aware about the content moderation, suggesting that moderation policies on Instagram are either sufficiently communicated to the users or the users have faced it.

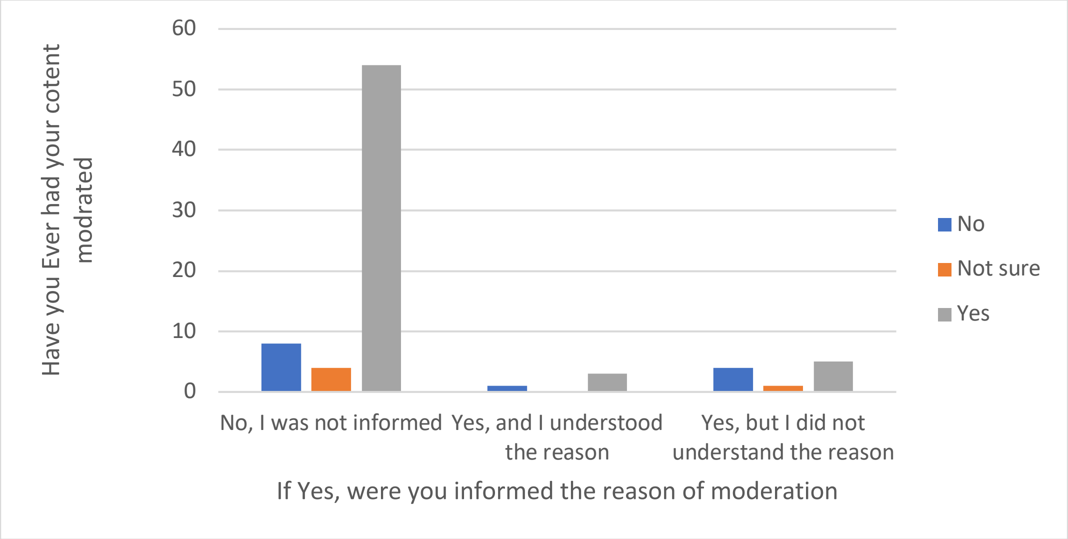
YouTube Users: Every respondent who frequently used YouTube reported awareness of content moderation practices, reflecting strong exposure to the platform's active content regulation systems.

Twitter Users: Findings shows that all Twitter users are also aware of moderation activities. This result aligns with Twitter's reputation for active discussions around moderation, free speech, and platform governance.

Reddit & Other Platforms: Similarly, users of Reddit and other platforms indicated full awareness, implying higher digital literacy and familiarity with community guidelines and moderation practices.

The data clearly shows that awareness of content moderation is universal across all major social media platforms among the participants. This widespread awareness suggests that platform communication, user experiences, and public discourse around content governance have made users consistently conscious of moderation activities, regardless of which platform they use most frequently.

4. CONTENT MODERATION:



The participants were asked whether they ever have had their content moderated and if yes then were they informed about the reason as to why such action was taken. The

findings show that out of the respondents, majority have experienced content moderation like removal, restriction, or flagging etc. by social media platforms.

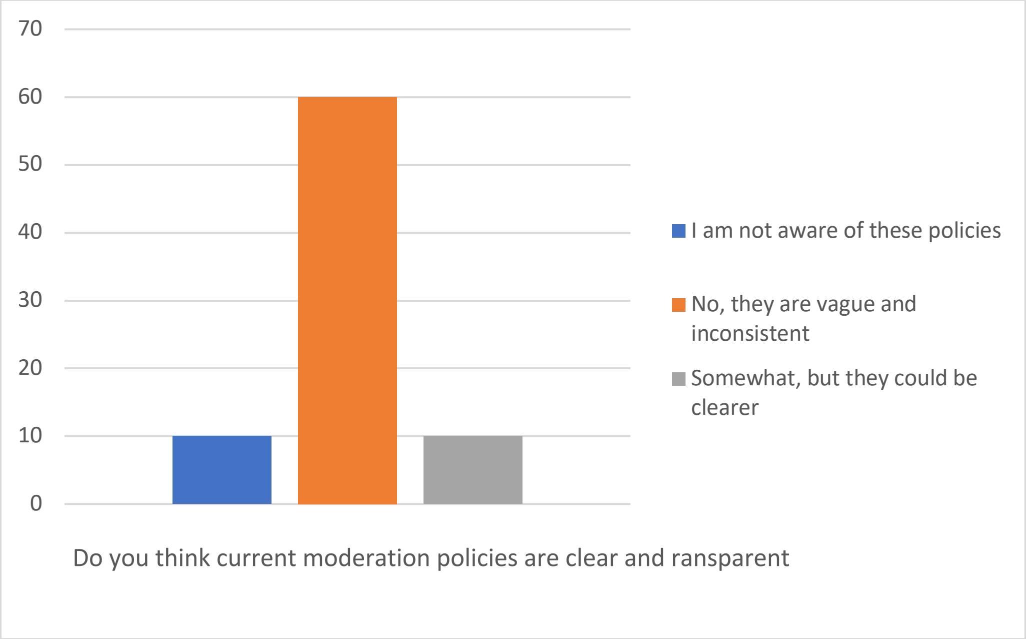
A significant number of respondents (over half) indicated that their content was removed, restricted, or flagged, but they were not informed about the reason for the moderation action.

Some respondents did report being informed of the reason for the moderation action, though the understanding of the reason was not always clear.

A smaller portion of respondents understood the reason for the moderation action when informed.

This shows that social media platforms do moderation the contents shared but the there is a lack of transparency or communication with users regarding the reasons for such actions.

5. TRANSPARENCY:

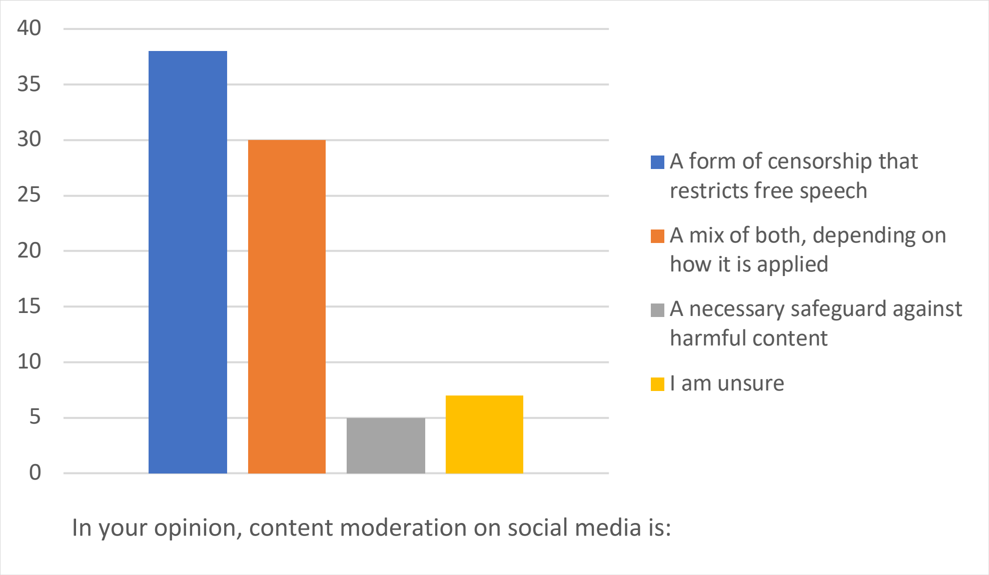


The findings show that the majority of respondents are of the opinion that the current content moderation policies of social media platforms are vague and inconsistent. A large number of respondents expressed dissatisfaction with the clarity and transparency of these policies which indicates that they find the rules and moderation practices unclear.

According to some respondents while the policies could be somewhat clearer, many are not aware of the specific content moderation policies in place. This suggests that there is a general lack of communication or understanding between social media platforms and users regarding the guidelines for content removal and moderation actions.

Over all the data shows a widespread perception that social media content moderation policies lack consistency and transparency, with a notable number of users unaware of the specific guidelines or the reasons behind moderation actions.

6. USER PERCEPTION:



The findings show that user perception on content moderation on social media are divided, with different perspectives expressed by the participants:

A Form of Censorship That Restricts Free Speech: A significant portion of respondents perceive content moderation as a form of censorship that limits free speech. These participants are concerned that moderation actions restrict their ability to express themselves freely on these platforms.

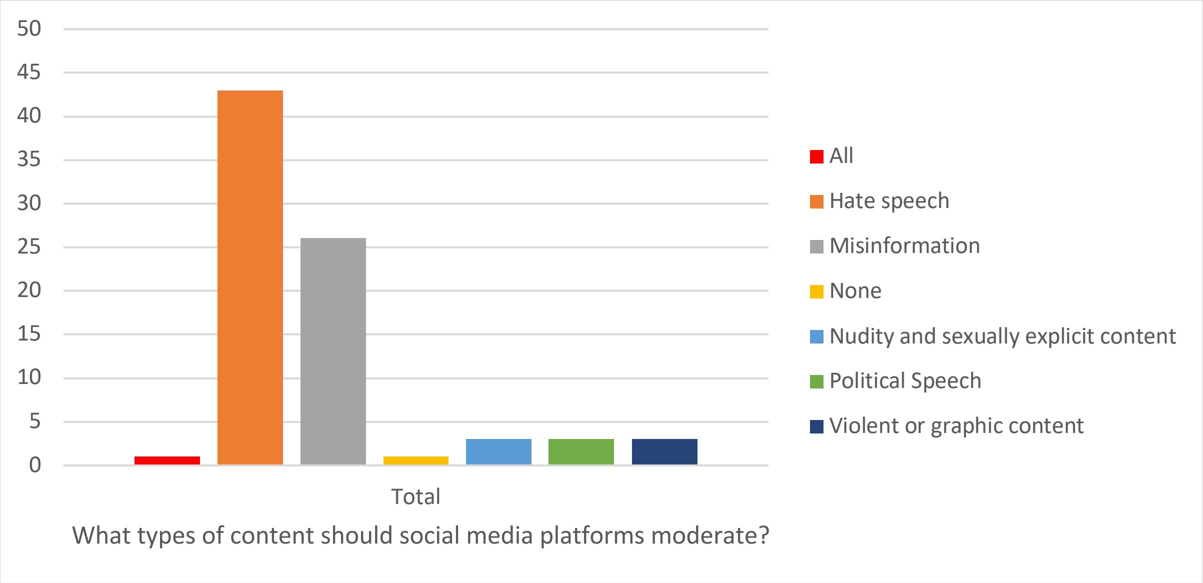
A Necessary Safeguard Against Harmful Content: Some respondents acknowledge the need for content moderation. They consider it a necessary safeguard against harmful, offensive, or inappropriate content. These participants see moderation as essential for creating a safe and respectful online environment.

A Mix of Both, Depending on How It Is Applied: A considerable number of respondents believe that content moderation is both a necessary safeguard and a form of censorship, depending on how it is applied. This finding suggests that while moderation can be helpful in controlling harmful content, its implementation can sometimes be arbitrary or inconsistent, leading to restrictions on free speech.

Uncertainty: A number of respondents were unsure about their stance on the matter which indicates a lack of clarity or understanding of the complexities surrounding content moderation.

Overall all the data shows that user perception of content moderation is mixed, with concerns about censorship and free speech being a central concern for many. However, there is also recognition of its necessity in managing harmful content, with the overall view depending on how the policies are implemented.

7. WHAT SHOULD BE MODERATED:



The findings show that respondents believe social media platforms should moderate a variety of content types, with particular emphasis given on:

Hate Speech: A majority of respondents are of the opinion that hate speech should be actively moderated by social media platforms. It was the most frequently selected option, highlighting the widespread concern about the harmful impact of such content.

Misinformation: A significant number of respondents also identified misinformation as a type of content that should be moderated. This suggests a growing awareness of the harm of false or misleading information, which often influences public opinion and behavior.

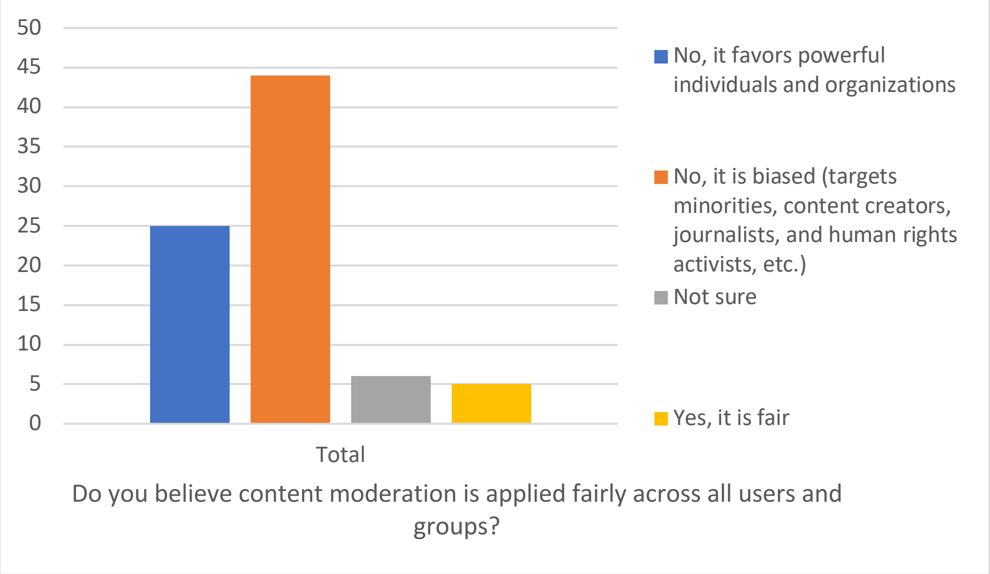
Nudity and Sexually Explicit Content: Some respondents believe that nudity and sexually explicit content should be moderated which can help maintain community standards and protect vulnerable audiences.

Violent or Graphic Content: Few participants pointed out the need to moderate violent or graphic content, which shows the concerns about the mental and emotional impact such material may have on viewers.

Political Speech: A smaller group of respondents suggested that political speech should be moderated, although this remains a controversial area because of its connection to the right to free speech and expression

None / All: A few respondents expressed extreme positions, either advocating for the moderation of all types of content or none at all, indicating polarized views on the scope of content regulation.

8. APPLICATION OF CONTENT MODERATION:



The data shows a strong perception among respondents that content moderation is not applied fairly across users and groups on social media platforms.

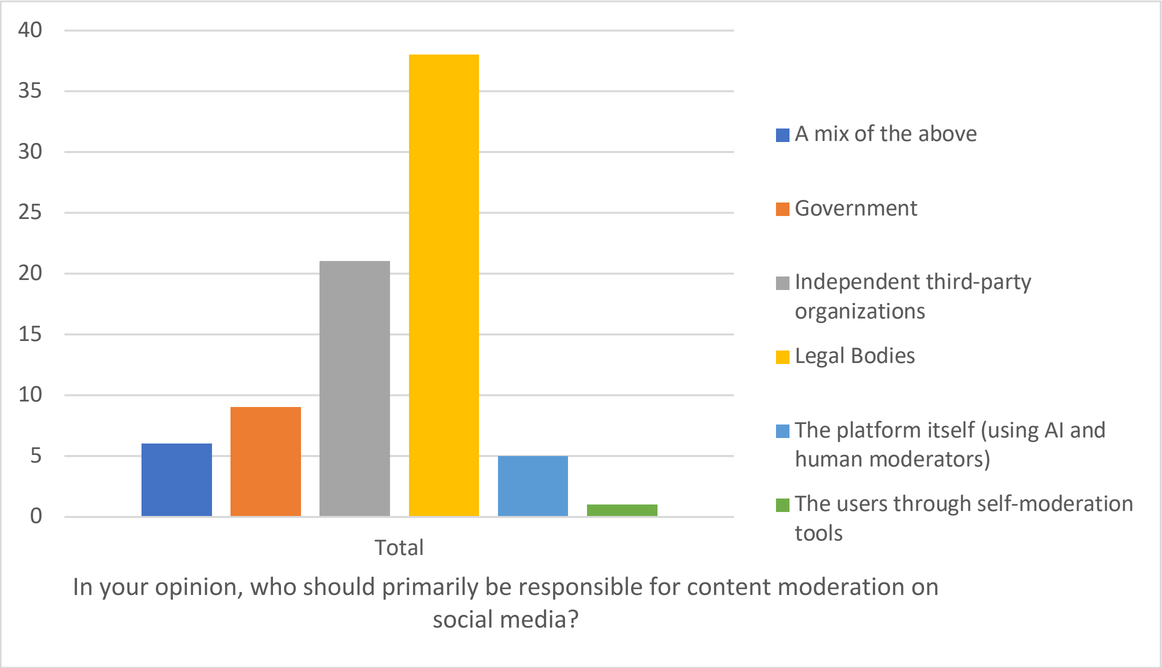
The majority of respondents believe that moderation practices are biased and target minorities, content creators, journalists, and human rights activists. This shows a widespread concern about discrimination and suppression of marginalized voices.

Many respondents also feel that content moderation favors powerful persons and organizations, which shows that influential groups may receive preferential treatment, while ordinary users are more likely to face moderation.

Only a few respondents believe that moderation is fair and applied equally across all users and groups. This shows a low level of trust in the neutrality and objectivity of content moderation systems.

A small group of participants are not sure whether content moderation is fair or biased, which shows a lack of transparency that leaves users confused or sceptical.

9. PRIMARY RESPONSIBILITY:



The findings show a diverse range of opinions among the respondents regarding who should who should be primarily responsible for content moderation on social media.

The majority of respondents believe that legal bodies such as courts, regulatory agencies, or independent third-party organizations should primarily be responsible for content moderation. This shows a strong demand for neutrality, accountability, and fairness, with the belief that external regulation would prevent bias and undue influence.

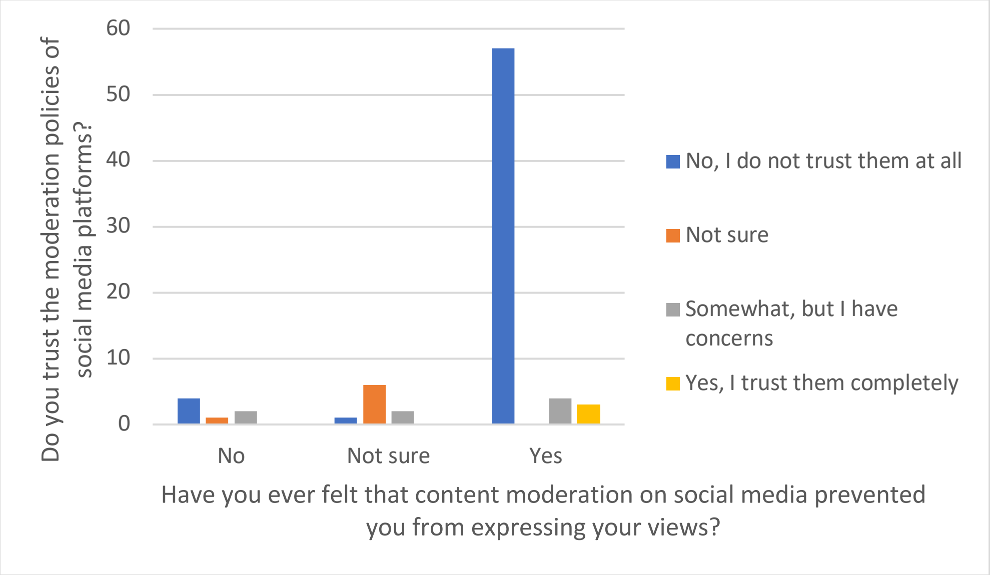
Many participants believe a mix of all, such as platforms, governments, and independent bodies, should be responsible for creating a balanced and transparent system. This shows a recognition that content moderation is complex and cannot be left solely to one entity.

Some believe that platforms (using AI and human moderators) should handle content moderation, which shows a degree of trust in technological solutions and platform-specific policies.

A noticeable number of respondents favored direct government regulation which shows a belief that official government intervention is necessary to ensure fairness and public accountability.

While very few participants also believe that users alone, through self-moderation tools, can effectively manage harmful or inappropriate content.

10. CREDIBILITY AND FREE EXPRESSION:



The findings show that the majority of respondents do not trust the moderation policies of social media platforms at all. This shows a significant credibility gap between platforms and their users, which is driven by perceptions of bias, inconsistency, or opaque decision-making processes.

A few of the participants expressed that they "somewhat" trust social media platforms but still harbour concerns.

This shows that while some users recognize efforts made by platforms still issues like perceived unfairness, lack of transparency, and questionable enforcement of policies continue to erode full confidence.

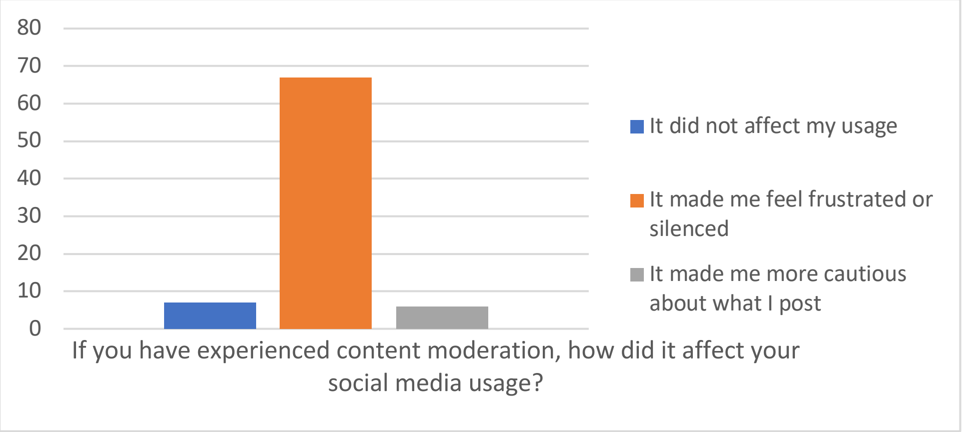
A small group of participants completely trust the moderation policies which shows either a strong belief in platform integrity or perhaps limited negative experiences with content regulation.

A substantial number of respondents reported that content moderation has prevented them from expressing their views. This shows that many users feel their freedom of expression is being curtailed by current moderation practices, potentially contributing to the broader distrust.

Some users chose "Not sure" for both trust and censorship impact, suggesting that the complexity and opaqueness of moderation processes make it difficult for users to fully understand or judge them.

Overall all the responses reveal a critical lack of trust in social media content moderation systems, coupled with a widespread sentiment of being censored or silenced. There is a clear demand for more transparent, fair, and accountable moderation practices to rebuild user trust and ensure that platforms truly support free and open dialogue without unjust suppression.

11. IMPACT OF MODERATION:



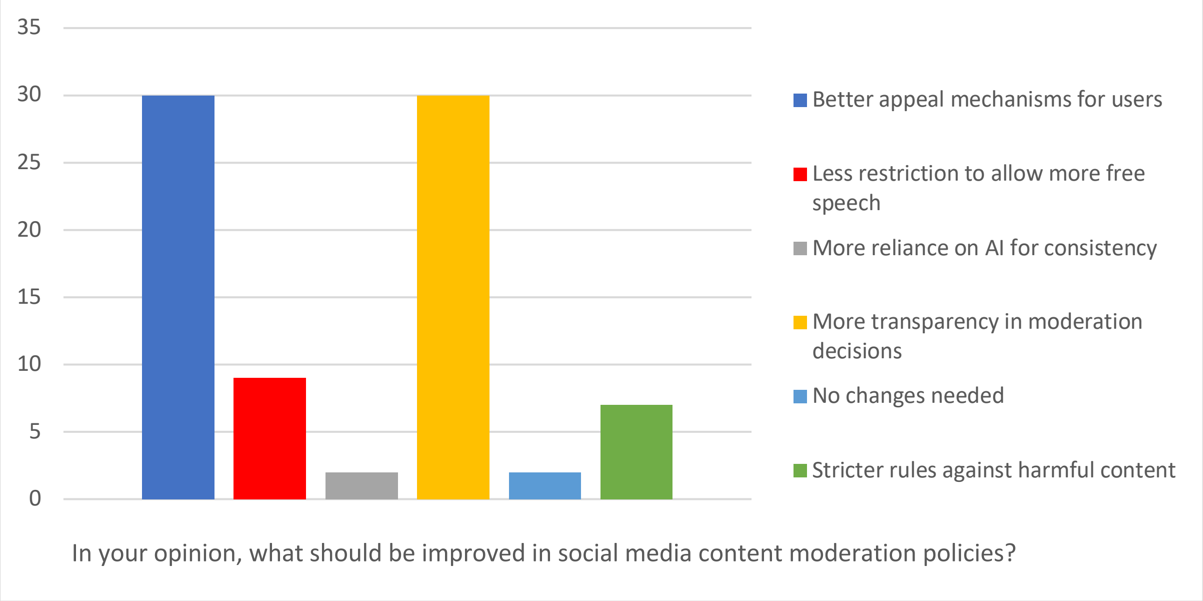
The findings show that majority of respondents believe that content moderation made them feel frustrated or silenced. This shows a deep emotional impact where users feel that moderation policies often suppress legitimate expression and create a sense of being unfairly targeted.

A significant number of participants indicated that they became more cautious about what they post following moderation experiences. This shows how the users are engaging in self-censorship, possibly limiting their genuine opinions or creativity to avoid potential penalties.

A smaller group of users stated that moderation did not affect their social media behavior. This group appears to have either accepted the moderation environment or simply adjusted their usage habits without feeling personally restricted.

Overall, the results reveal that content moderation has had a notable negative psychological and behavioural impact on most users. Feelings of frustration, silencing, and self-censorship dominate the responses, with only a minority continuing to use social media unaffected. These findings indicate that current moderation practices may be damaging user experience and trust, calling for reforms that better balance platform safety with freedom of expression.

12. WHAT SHOULD BE IMPROVED?



The data shows that an overwhelming majority of respondents emphasized the need for more transparency in moderation decisions. This shows that users want to know and understand why their content was flagged, removed, or restricted, and seek clearer guidelines to avoid future violations.

Many participants emphasized the need for stronger, fairer appeal systems. This shows a perception that current appeal processes are either inadequate or ineffective, and users want the ability to challenge moderation decisions more meaningfully.

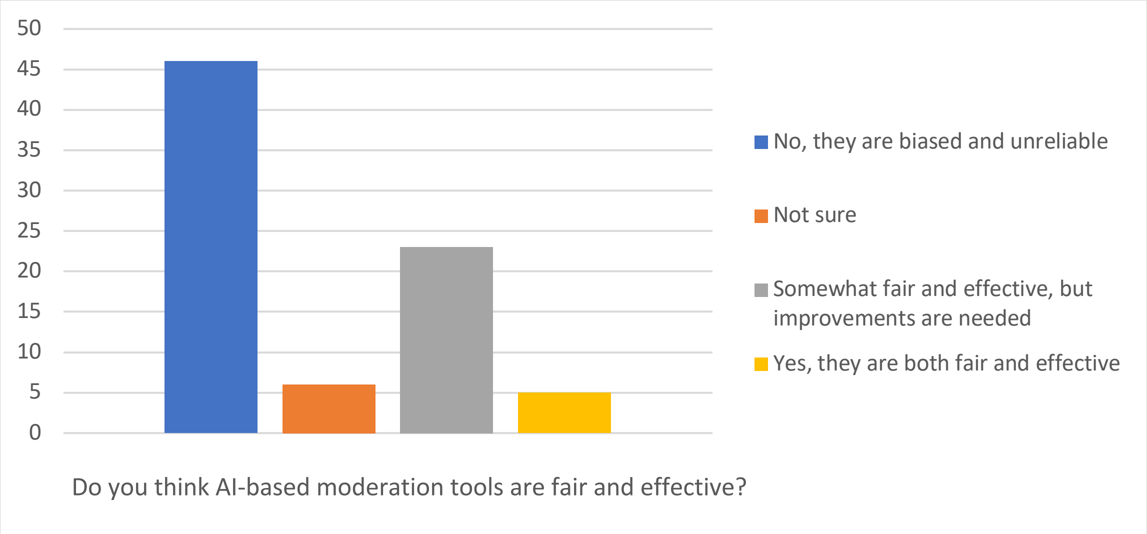
Several participants called for stricter rules against harmful content. This shows that users are not only concerned with free speech but also expect platforms to better protect them from hate speech, violence, harassment, and misinformation.

At the same time, a significant group emphasized the need for less restriction, emphasizing the protection of free speech. This shows a delicate balance that platforms must maintain ensuring safety without unnecessarily stifling legitimate expression.

A few respondents also suggested that greater use of AI could help bring more consistency and fairness to moderation decisions, reducing human bias and error. While very few participants believe that no improvements are required.

Overall the results show a strong user demand for greater transparency, better user rights through appeals, and a more balanced approach that safeguards both protection from harm and freedom of speech. There is a clear call for platforms to rebuild trust by making their moderation policies and actions more open, fair, and user-friendly.

13. AI BASED MODERATION:



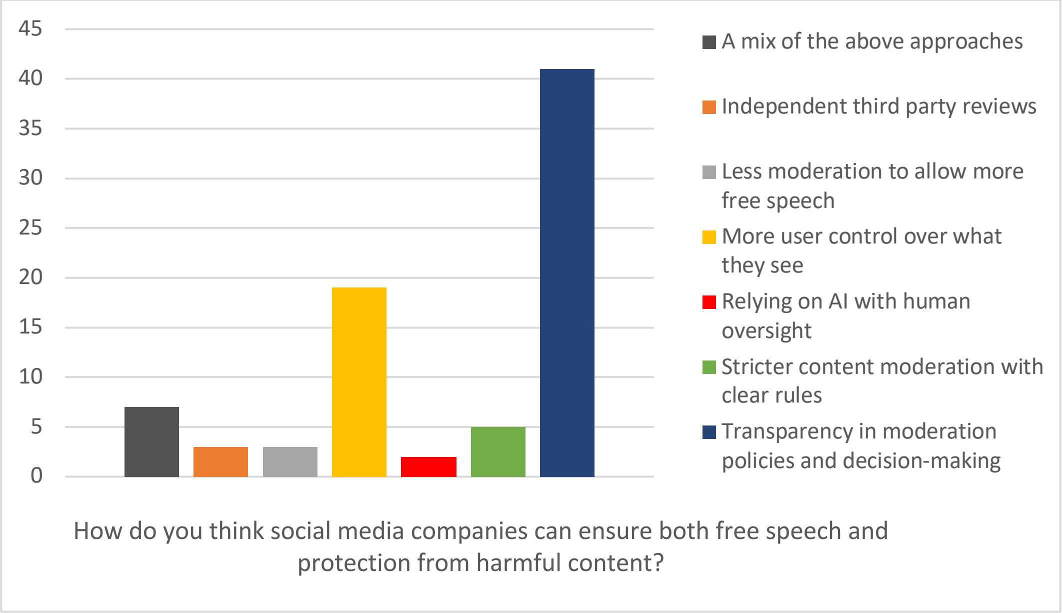
The data shows that a large number of participants believe that AI tools are biased and unreliable. This shows a serious concern that algorithmic moderation may unfairly target certain views or groups, and that errors in judgment by AI are still too frequent.

A significant respondents of respondents feel that AI-based moderation tools are somewhat fair and effective, but improvements are clearly needed. This shows a recognition of AI's potential, but also a widespread awareness of its current limitations, especially in dealing with complex or nuanced content.

Only a relatively small group expressed that AI-based moderation is both fair and effective. This suggests that while there is some optimism about AI's role in content moderation, it is not yet the dominant view among users.A noticeable portion of respondents selected "Not sure", highlighting the fact that many users either do not fully understand how AI moderation works or feel disconnected from its processes.

Over all the responses reveal a critical but open-minded attitude toward AI-based moderation tools. While many users acknowledge the usefulness of AI, there is still a strong call for improvements, especially to address issues of bias, reliability, and fairness. Platforms need to focus on enhancing AI transparency, minimizing biases, and building public trust through clearer communication about how these systems operate and are evaluated.

14. BALANCING FREE SPEECH AND PROTECTION FROM HARMFUL CONTENT:



The data shows that an overwhelming number of respondents believe that transparency in moderation policies and decision-making is the most important step for social media companies to ensure free speech and protection from harmful contents. This shows that users want to clearly understand how content is moderated, why decisions are made, and what guidelines are being applied to ensure fairness and accountability.

A large group of participants emphasized the need for more user control over what they see. Rather than relying solely on centralized moderation they want personalized tools to filter or curate the content they interact with which will empower them to shape their own experiences.

Several respondents preferred a mix of strategies that is combining transparency, user control, AI with human oversight, and clear rules. This shows a nuanced understanding that no single method is sufficient; a comprehensive and flexible system is needed to balance competing concerns.

Some users also highlighted the need for stricter content moderation as long as the rules are clearly communicated. The emphasis is on consistency and fairness rather than arbitrary or hidden decisions.

A smaller group of participants called for less moderation to better protect free speech, and a few suggested involving independent third-party reviews to ensure that moderation practices remain impartial and unbiased.

Overall, the responses show a clear call that transparency must be the foundation of all content moderation efforts at the same time, empowering users with more choices and control is equally critical. Balancing methods, including AI, human review, and independent oversight, will be crucial for creating fairer and more trusted platforms. Social media platforms must build trust by making their systems more transparent, user-driven, and accountable.

ANALYSIS:

The empirical study demonstrates that user’s perceptions of content moderation on social media platforms are complex, multi-layered, and sometimes contradictory. The survey's results indicate that although users clearly understand the value of moderation in preventing harmful content like hate speech, false information, and graphic violence, there is also a considerable and pervasive lack of trust in the way this moderation is implemented.

The survey found that most participants believe that the present methods of content moderation are opaque, ambiguous, and inconsistent. A vast majority of users believe that decisions about moderation are not explained well, and that frequently, when content is tagged, removed, or restricted, users are either not told why it was done or they do not comprehend the explanations when they are given. The measures intended to create safer online environments are also making users feel more alienated and doubtful, as this lack of clarity creates feelings of dissatisfaction, silence, and mistrust towards social media platforms.

The data clearly showed the problem of perceived bias in moderation decisions. According to several respondents, social media platforms favour powerful people, businesses, and political organisations while disproportionately targeting minority voices, human rights activists, independent journalists, and small content creators. This apparent disparity raises the possibility that moderation is not always applied impartially and is instead impacted by political or economic factors. Such a view seriously damages platforms' legitimacy as impartial arbiters of online debate.

The results show that users are significantly impacted emotionally and behaviourally by content control. After going through moderation actions, a significant portion of participants said they felt silenced or frustrated, and many acknowledged that they started being more selective about what they posted. The idea that moderation policies, if not properly regulated and transparently implemented, may discourage legitimate expression, suppress dissenting opinions, and inhibit democratic dialogue all of which are protected under the fundamental right to freedom of speech and expression makes this effect on speech especially worrisome.

According to the study, users are generally unhappy with the way AI-based moderation tools are now performing, even though they are aware of its potential. Although AI is seen as a useful technological answer, it is prone to biases, mistakes, and a lack of contextual knowledge, especially when it comes to satire, political speech, or content from underrepresented groups. Many responders emphasised that in order to guarantee accountability, contextual sensitivity, and justice, AI-based moderation must always be combined with human oversight.

The overwhelming need for improved appeal procedures and greater openness is another important finding of the study. Respondents underlined that social media businesses need to set up fair, easily available, and unbiased mechanisms for users to challenge and appeal decisions, give users more precise instructions, and explain to users the rationale behind moderation measures. Moderation will probably continue to be viewed as arbitrary and oppressive rather than just and protecting in the absence of these advancements.

The data also shows that users have a complex perspective of moderation. Many respondents agree that it is a "necessary evil" that must be used justly and carefully to protect online places from harm. A mixed approach to content management, in which platforms, governments, independent third parties, and users themselves all have a place to play, is widely supported. It is evident that users are afraid of political overreach and do not fully trust governments or platforms to control themselves. In order to guarantee impartiality, openness, and the defence of both security and free expression, they suggest a cooperative framework.

Lastly, the results also show that people strongly support the right to free expression even while they want harmful content to be restricted, especially hate speech and false information. Therefore, the difficulty lies in finding a fine balance such that moderation does not equate to censorship. "More transparency," "user empowerment," and "better appeals" are all desired, which emphasises that users are not opposed to moderation in general but rather want a democratic, participative, and accountable system.

Essentially, the study's user perceptions make it abundantly evident that social media companies need to radically reconsider the way they design and conduct content moderation if they want to safeguard user trust and their significance as forums for public discourse. In order to truly protect online communities without compromising basic human rights, moderation must be changed from a covert act of control to an open, equitable, and participatory process.

CONCLUSION:

The paper concludes that according to users ' perceptions, content moderation on social media platforms is complicated and frequently inconsistent. Although it is commonly acknowledged that content moderation is essential to safeguarding online communities against hate speech and false information, users continue to have serious concerns over the transparency, fairness, and consistency of these procedures. The results show that a many users believe that moderation rules are unclear and applied inconsistently, especially when automated systems are used. This can give the impression that some groups are being censored or biased. However, there is widespread agreement that platforms have a moral obligation to refrain from promoting harmful information, and that moderation is necessary to preserve a secure and welcoming online environment. The report further emphasises. that more accountability, more transparent moderation decisions, and more sophisticated, context-sensitive strategies that strike a compromise between the necessity of user protection and the fundamental right to free speech are all necessary. Providing open, equitable, and participatory moderation mechanisms that promote trust and protect user safety and digital rights in a constantly changing online environment is ultimately the challenges being faced by these platforms and governments.