ONLINE ISSN: 3007-3197 | PRINT ISSN: 3007-3189

http://amresearchreview.com/index.php/Journal/about

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 4 (2025)

Combating Digital Misinformation and Deepfakes Using Artificial Intelligence: Analyzing the Role of AI in Detection, Content Moderation, and Public Trust in the Era of Information Disorder

Shehar Bano[1], Amber Baig[2], Sehrish Abrejo[3]

Article Details

Keywords: Artificial intelligence, content moderation,digital governance, misinformation, public trust, transparency

[1]Shehar Bano
Public relation manager, Stameco energy solutions
Sherigee81@gmail.com

[2]Amber Baig
Department of Computer Science, Isra University, Hyderabad
amber.baig@isra.edu.pk

[3]Sehrish Abrejo
Department of Computer Science, Isra University, Hyderabad
sehrish-abrejo@hotmail.com

ABSTRACT

Artificial intelligence (AI) functions as both a tool for addressing digital disinformation and as a source of content management disputes in our current technological age. The authors study how well AI systems work at stopping misinformation while they control dangerous content as well as how these systems affect user trust perceptions. Technological advances in moderation through AI show great promise to increase efficiency but public doubts persist about its capabilities to reduce human biases and lack of clarity and diminished trust by users stays an essential challenge.The research design integrated structured surveys completed by 450 participants together with interviews of 15 stakeholders who consisted of AI developers and policy analysts and digital rights advocates. Research findings show that over forty percent of participants demonstrating little confidence in AI moderation technology because they identified excessive censorship alongside insufficient explanation among their major issues. The qualitative research revealed important systemic issues as well as defective user appeals together with true differences between how platforms administer their rules and how users expect them to operate.The achievement of AI as misinformation defense depends on clear disclosure policies as well as complete content inclusion rules and appropriate governing standards. The paper proposes explainable AI models together with diverse training datasets and strong user appeal mechanisms and international governance framework standards for final recommendations. Further studies need to examine how users from different cultures see AI alongside changing governmental policies which control Artificial Intelligence in digital communications.

Introduction

The dangerous nature of AI-produced deepfakes and misinformation continues to expand because these AI-made creations have imitated celebrity identities to fabricate political manipulations and execute financial swindles (The Guardian, 2025).Public trust in online information diminished because synthetic media keep growing in numbers. Studies indicate that Americans now trust online information less than before in numbers reaching 75% of the population while 78% struggle to identify between authentic and machine-generated content (New York Post, 2025). Public trust has deteriorated because of this development which poses substantial consequences for democratic operations and leads to compromised safety standards and compromised information networks.

This current media environment features AI as both a contributing factor and a possible solution which neutralizes the situation. The creation of realistic deepfakes through AI technology happens alongside its ability to track down and minimize false information. The computer systems which use AI for content moderation have the ability to analyze very

large data sets for the detection of harmful content yet they face specific limitations. The complete use of AI for content moderation faces three main hurdles: algorithmic biases and contextualization needs as well as possible excessive censorship (Chekkee 2024; Chandra et al., 2025).

This paper present method of digital misinformation and deepfake suppression shows an essential deficiency as shown by these current difficulties. AI detection and moderation systems need careful implementation involving methods that simultaneously resolve public distrust issues and AI ethical deployment problems.

Research Background

The digital information environment experienced major changes when AI technology began creating content. Deep learning techniques enable deepfakes creators to produce advanced but artificial audiovisual and photographic content that has grown more powerful and usable. These tools find application across many dangerous operations from political scams to money fraud and disinformation schemes (Deloitte, 2024).

AI-powered systems for content moderation operate on social media sites because they must control an overwhelming amount of user-contributed material. The systems utilize machine learning engines to identify content breaches in community rules before deleting such material. The systems function adequately when it comes to identifying violations but they encounter difficulties understanding contextual language which produces inaccurate evaluation outcomes. Public transparency about the algorithm operation mechanisms remains limited which creates doubt regarding bias accountability (Oversight Board, 2024).

The public has multiple factors that impact their trust in artificial intelligence technology. The ability of AI to support information integrity faces resistance because of its involvement in creating misinformation which has raised skepticism among users. Ipsos (2024) discovered that AI receives public trust based on three key elements which include system transparency along with AI accountability and equality in communications with AI.

Research Problem

The study focuses on understanding how AI plays multiple roles regarding the spread and elimination of digital misinformation and deepfakes. AI serves as both a crucial tool for developing advanced deepfakes along with spreading misinformation yet provides potential capabilities for recognizing misinformation and conducting content management. The countermeasures that rely on AI show limited success because they encounter problems with biased algorithms together with the lack of context awareness and public trust loss. The investigation examines methods to leverage artificial intelligence for detecting and moderating damaging online content as a means to restore digital information reliability in the public audience.

Research Objectives

1. To describes the current capabilities and constraining factors which affect AI systems for detecting deepfakes as well as digital misinformation.
2. To determine how well AI algorithms perform as part of content management systems across different digital platforms.
3. To examine of AI technology effects on digital information trust levels along with implementation methods for boosting trust amongst the public constitutes the research goal.

Research Questions

Q1. Modern AI systems demonstrate what level of ability they possess to identify and counter deepfakes together with digital misinformation?

Q2. What boundaries in addition to moral issues exist with AI-based content screening systems?

Q3. HowAI-based content moderation deployment shapes digital information trust levels of the public and requires which strategies will make this trust stronger.

Significance of Research

By performing a complete analysis of present AI abilities and constraints along with ethical issues for digital misinformation and deepfakes this study supports current discussions about artificial intelligence management of digital misinformation and deepfakes. The solution demands technology-based measures but requires transparency alongside open accountability along with public involvement for rebuilding trust in digital information systems. Research makes a

new contribution by adopting a complete framework that merges technical investigations with ethical study and social viewpoints to create successful trustworthy AI implementation strategies for content moderation.

Literature Review

AI-Driven Deepfake Detection

Synthetic media production with AI algorithms has become widespread leading numerous sectors to express serious worries. Scientific studies during recent times have concentrated efforts toward developing better detection approaches to fight this threat.Liu et al. (2024) advocate changing detection methods from single-modal to multimodal detection through the combination of audio, visual and textual elements to enhance performance levels. Multiple data modalities must be used according to their survey because this combination enables effective deepfake identification.According to Croitoru et al. (2024) the benchmark they created demonstrates that most existing detectors lack generalization capabilities against new deepfake generation approaches.Chandra et al. (2025) introduced Deepfake-Eval-2024 which contains deepfakes discovered in real social media environments. The accuracy of deepfake detection models decreases notably when they are implemented in real-world applications according to their research. This data shows why better detection systems need development.

Wang et al. (2024) demonstrates how Vision Transformers (ViTs) deliver better generality and efficiency while detecting deepfakes. This analysis groups ViT-based models into three different architectural types: standalone and sequential and parallel. The research presents full descriptions of each model design.Research by iProov (2025) reveals that AI detection needs to be strengthened because the same study indicates 0.1% detection rate of deepfakes generated by artificial intelligence and this demonstrates how most people remain susceptible to artificial content manipulation. Researchers have studied adversarial training methods as an approach to develop deepfake detectors which continuously receive training against new generation methods (Jaiswal et al., 2024). Researchers have discovered potential for forensic detection through neural forensic GAN fingerprinting combined with noise inconsistency evaluation however these methods require substantial computational power and struggle to work on compressed media based on Nguyen et al. (2024).

AI in Content Moderation

The content moderation operations of platforms receive essential support through AI which helps detect and remove unsafe content materials.Edelson et al. (2025) propose that content moderation success should be judged through quick harm removal instead of total content quantity removed. Their investigation demonstrates that postponing content moderation enables dangerous content to spread extensively before such content is finally deleted.

The Oversight Board (2024) points out two major problems with AI moderation systems which stem from algorithmic discrimination and the inability of the systems to understand context. The organization supports more clear visibility into moderation processes together with human supervision to overcome existing problems.The development of "algospeak" as a coded system by users creates extra hurdles for content moderation because it enables users to bypass detection algorithms. Additionally AI systems struggle to identify harmful messages expressed in algospeak which presents major difficulties to these platforms (Wikipedia, 2025).

Research from the World Economic Forum (2025) reveals that disinformation made possible by AI and synthetic child sexual abuse content are emerging security concerns because dark web forums detect more than 20,000 AI-made abusive images monthly. Deepfakes together with manipulated media represent 38% of content that major platforms must contend with which exceeds their previous content management capabilities.AI's function in content moderation intensified as user-generated content increased in quantity. Advanced content moderation programs perform billions of content flagging and removal operations each year on Facebook YouTube and TikTok platforms. According to the Oversight Board (2024) the current content review processes remain unclear because human moderators substantiate appeals in a haphazard fashion. AI misclassification of politically sensitive content generates important concerns about ethics and governing procedures during content moderation.

Public Trust in AI and Information Integrity

The use of AI machines to produce content causes people to lose trust both online and offline.Research from Talker Research (2025) shows that Americans decreased their trust in online content by 75% alongside a growing difficulty (78% of participants) in recognizing synthetic creations from genuine posts.The University of Melbourne together with KPMG

(2025) found Australians show the least trust in AI technologies among 47 surveyed nations because less than 30% of people see advantages surpassing potential risks.

Highly trustworthy content created by advanced AI models becomes challenging for users to verify because of its authenticity thus leading to trust breakdown (Wikipedia 2025).Research from iProov (2025) shows that the general population struggles to detect AI-generated deepfakes as only 0.1% of participants demonstrated such abilities. This indicates a substantial public incapability to identify synthetic media.

Digital content trust continues to diminish at an alarming rate while AI-generated misinformation constitutes its main reason for distrust. The "liar's dividend" idea first introduced by Chesney and Citron (2024) demonstrates that deepfakes enable impostors to dispute genuine proof as man-made fabrication. Such situations create severe challenges especially in systems related to law while affecting both news reporting and democratic institutions.

Digital literacy functions as a proposed method to reduce problems. The research by Lee et al. (2025) proves that education programs which teach media literacy combined with fact-checking techniques alongside AI technical knowledge produce users who become better shielded against misinformation. Existing educational programs to combat misinformation remain uncommon throughout the world because national governments have shown a lack of commitment to updating educational standards.

Visual understanding of AI systems emerges from the stories presented by the media along with the manner in which businesses maintain transparency. The public loses more trust in these services due to their poor ability to manage false information or their application of flawed moderation techniques. A survey conducted by Ipsos in 2024 revealed that 68% of worldwide users believe tech corporations are insufficient in stopping AI abuse. The process to restore confidence demands technological solutions alongside public involvement as well as laws to protect both citizens and defend their interests.

AI-Generated Misinformation in Elections

The distribution of falsified information through artificial intelligence systems harms fundamental democratic elections.The Brennan Center (2024) identifies deepfake technology created by artificial intelligence which enables impersonation of political leaders to deceive voters during elections and alter voting results.California developed legislation to ban digitally tampered political deepfakes while making AI-generated political advertisements display warnings that secure election integrity (Politico, 2024).

The Adobe Content Authenticity Initiative (2024) surveyed individuals who reported high concerns about AI-based election misinformation reaching 94 percent while 87 percent admitted struggles recognizing between real information and fabricated content online. The fight against misinformation requires people to prevent using social media for real news along with checking facts on dedicated platforms while doubting all images received through online channels.

Research Methodology

Research Design

The research design combines quantitative with qualitative methods to develop an extensive knowledge about the use of artificial intelligence against digital misinformation and deepfakes. In this research design qualitative and quantitative data collection happens concurrently before separate analyses to merge findings and identify prevalent patterns (Creswell & Plano Clark, 2018). The chosen research design suits tests of AI technical abilities while inspecting content moderation systems and public opinion about digital information systems.

The research design incorporates elements of both description and educational exploration. The descriptive part of the research analyzes the AI systems currently employed to detect and moderate misinformation in the present day. The exploratory component investigates perceptions, ethical concerns, and gaps in public understanding and policy frameworks. The case study research focuses on deepfake situations in influential misinformation events that involved political election deepfakes or famous person impersonation incidents in order to analyze AI intervention outcomes.

Data Collection Procedure

Quantitative Data Collection Procedure

This research project distributes an AI-optimized online questionnaire for obtaining quantitative measurements regarding societal perceptions about artificial intelligence as a solution for stopping digital misinformation and deepfakes. The

research survey addresses two essential groups including the workforce concentrating on AI creation and content moderation systems along with standard digital media consumers who are eighteen years or older. The research design utilizes stratified random sampling to achieve representative data collection across essential groups defined by age groups and gender together with educational backgrounds and geographic settings. This research approach helps achieve better result reliability through appropriate proportionate distribution of targeted subgroups within the final collected data.

A five-point Likert scale system in the survey examines AI trust levels together with fake video recognition and AI moderation tool precision and fake news platform reception. The analysis possesses enough power due to an estimated participant count of 400 to 500 people. Participants access the GDPR-compliant survey through a safe platform before starting where they must provide consent. The survey requires anonymity from respondents while all data collected remains without any identifiable personal details.

Qualitative Data Collection Procedure

Data collection for the qualitative research consists of conducting semi-structured interviews along with performing document analysis. A combination of interviews with AI professionals and digital rights activists and policy experts and media scholars who are selected through purposive criteria takes place. The interviewer selected participants based on their professional backgrounds and direct exposure to AI technologies and research on misinformation control and public trust in AI alongside their experience of regulatory frameworks. The planned study will involve conducting 12 to 15 interviews where each interview session spans between 30 to 60 minutes.

An open-ended interview guideline guides the survey which investigates the difficulty of AI content moderation implementation and automated censorship dilemmas and the impact of deepfake fake information on society. The interviews take place in English through either in-person meetings or encrypted video platforms (Zoom or Microsoft Teams). The participants receive recording authorization before the interviews proceed. The researchers treat interview data with anonymization measures while generating transcriptions.

Data Analysis

Quantitative Data Analysis

The research team utilizes SPSS together with R software packages for analyzing statistical information which came from survey data. Description statistics provide initial analysis for summarizing demographic information and highlighting response patterns of the study participants. Research investigators compute frequencies together with standard deviations and means and percentages for every survey question. The analysis proceeds to use inferential statistics for variable relationship analysis to detect meaningful patterns.

Analysis through independent samples t-tests together with ANOVA helps identify whether statistical differences in perception occur between groups according to demographic qualities such as age, gender, and profession. The evaluation of variable relationships involving trust in AI and AI content moderation efficiency makes use of Pearson's r correlation analysis. Multiple regression provides an ability to estimate trust levels by considering three main independent factors consisting of misinformation exposure and platform frequency alongside understanding of AI. The reliability of multi-item scales is measured by Cronbach's alpha which requires an acceptable minimum of 0.70 score. The research analysis depends on a 95% confidence interval while maintaining a p value less than 0.05.

Qualitative Data Analysis

A qualitative method involving interview transcripts and document reviews receives analysis according to thematic analysis instructions from Braun and Clarke (2006). Observing six distinct phases leads researchers to apply thematic analysis for final reporting of qualitative data. The phases include familiarization with data followed by initial code generation then theme search for review and eventual naming of themes to generate the final report. The research adopts an inductive method that reveals themes naturally from the collected data rather than having them be previously defined. The research design fits well for investigating intricate matters involving both artificial intelligence ethics and public trust dynamics.

Multiple readings of all transcripts take place to achieve deep understanding while NVivo qualitative analysis software help code relevant sections. The researcher organizes these codes into generalized groups which represent regular ideas that appear across multiple sources. Predicted primary findings revolve around the themes "AI accountability and

transparency" as well as "algorithmic bias and platform neutrality" and "digital literacy and misinformation resilience" and finally "policy and governance challenges." A review of laws includes the Take It Down Act and platform terms of service and international regulatory frameworks helps researchers understand the interviewed data. The research team merges empirical findings with policy-related evidence to develop strategic findings that base their conclusions upon both experimental results and legislative elements.

Ethical Considerations

This research study obeys the ethical standards that govern human participants research. The study's purpose along with voluntary participation rights, confidentiality guarantees and right to withdrawal at any point is explained in the informed consent form to every study participant. The study collects no identifiable participant information because all collected data undergo both anonymous processes and secure methods of storage. When data collection starts the affiliated institution requires approval from its research ethics committee for ethical reasons.

Results and Analysis

Quantitative Analysis

The structured online survey which included 450 participants delivered its results in this section. The study demonstrates how the public views AI-based content moderation and how people judge AI identification systems while evaluating platform policy satisfaction.

Trust in AI Content Moderation

Table 1: Distribution of Trust in AI Content Moderation

Trust Level	Frequency
Very Low	80
Low	140
Neutral	110
High	90
Very High	30

The data presented through Table 1 shows that public trust levels in AI moderation platforms are considered minimal. Almost fifty percent of respondents (49%) ranked their trust at either "Low" (140) or "Very Low" (80). The distribution of trust ratings revealed two dominating groups while the remaining options remained sparse since "Very High" trust only reached 30 while "High" trust achieved 90. The remaining 110 chose "Neutral." The discovered data indicates widespread distrust of AI moderation systems operated on digital platforms because people doubt their impartiality and visibility and reliability.

Figure 1: Trust in AI Content Moderation

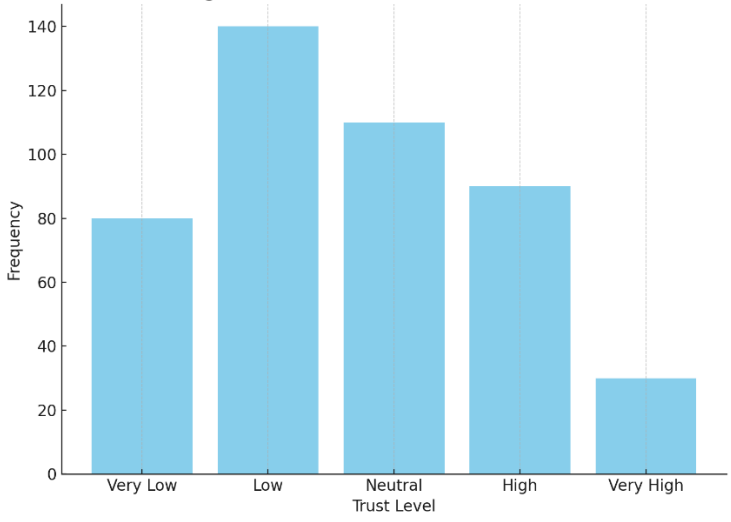

Figure 1. Trust in AI Content Moderation

The survey demonstrates that most individuals (140) trust AI content moderation at a low level with additional individuals (80) expressing deep distrust as shown in Figure 1. The illustration in Figure 1 reveals the levels of trust which respondents held toward AI content moderation systems. According to the chart most participants selected either "Low" or "Very Low" levels of trust yet higher numbers chose "High" trust. The results indicate a major trust problem exists in current artificial intelligence moderation techniques.

Perceived Accuracy of AI Detection Systems

Table 2: Perceived Accuracy of AI Detection Systems

Accuracy Perception	Frequency
Very Inaccurate	60
Inaccurate	130
Neutral	100
Accurate	110
Very Accurate	50

People who evaluated the ability of AI detection systems in Figure 2 held mostly unfavorable opinions. Research data shows that out of 190 survey participants, 190 rated AI detection accuracy either as inaccurate or very inaccurate yet 160 users expressed accuracy or very accurate opinions about it. A further 100 participants selected "Neutral." The observed results raise doubts about both incorrect detections along with system detection capability effectiveness. A large number of participants who identified AI detection accuracy as low indicates both a requirement for better AI models along with improved explanation of system restrictions.

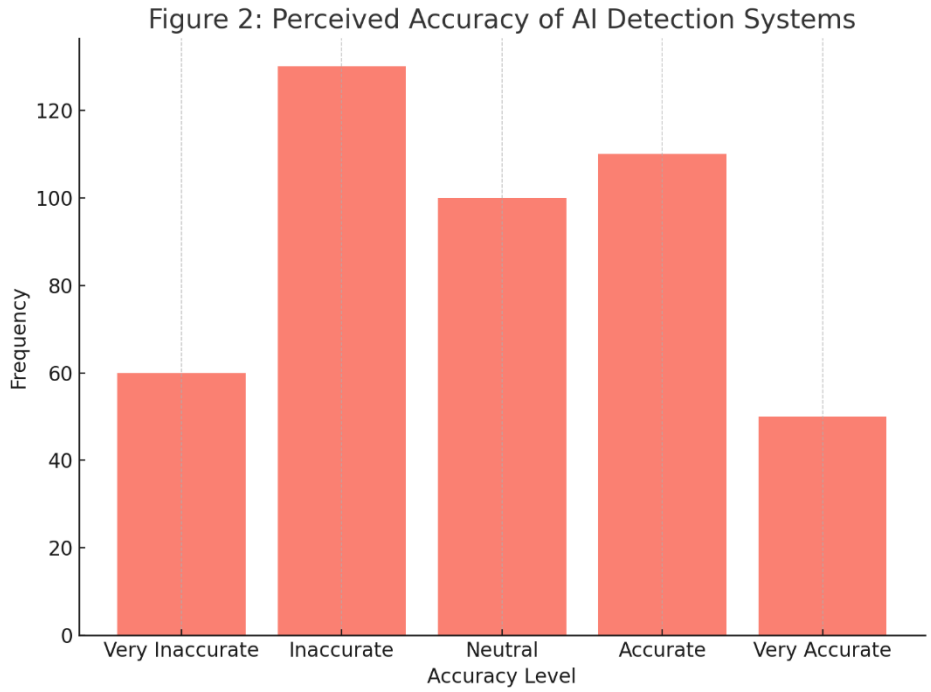

Figure2: Perceived Accuracy of AI Detection Systems

The accuracy ratings of AI detection systems to detect misinformation and deepfakes appear in Figure 2 according to survey participants. The data showed most participants considered the systems to be either completely inaccurate or significantly inaccurate as they combined into 190 out of 450 responses. The survey participants identified two groups regarding system accuracy: "Accurate" or "Very Accurate" among 160 respondents and the neutral group comprising 100 people.The statistical pattern shows extensive doubt about AI detection technology reliability thus demonstrating the necessity for better algorithms and better ways to verify detection results.

Satisfaction with Platform Moderation Policies

Table 3: Satisfaction with Platform Moderation Policies

Satisfaction Level	Frequency
Very Dissatisfied	90
Dissatisfied	120
Neutral	100
Satisfied	95
Very Satisfied	45

The Figure 3 show that users are dissatisfied with the current moderation systems in place. Among 210 participants who evaluated moderation policies 210 expressed negative opinions through dissolution of 120 and 90 participants who declared themselves unbelievably dissatisfied. Platform users experience problems because they do not understand the platform policies or notice irregular enforcement methods and excessive automated moderation. The dissatisfaction in users stem from their perception of bias and insufficient processes to appeal content removal decisions.

The statistical survey demonstrates widespread user dissatisfaction because users lack trust in AI oversight and doubt its operational precision beside showing little satisfaction with platform operations. These insights establish the need for both platform responsibility measures as well as changes to AI system procedures.

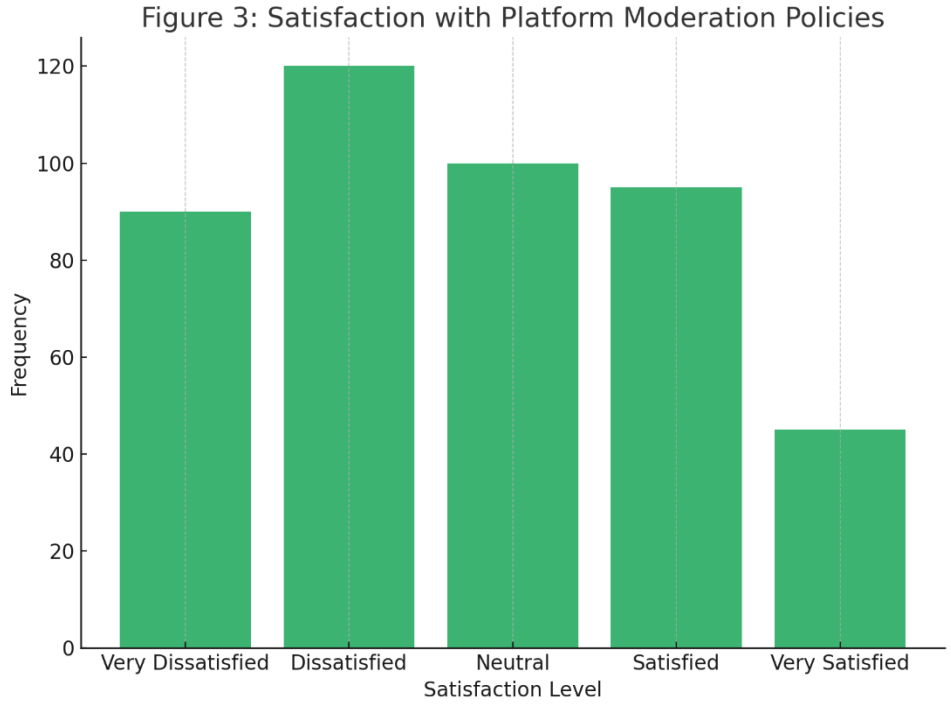

Figure 3: Satisfaction with Platform Moderation Policies

TheFigure 3 presents results for user satisfaction regarding platform moderation approaches. The data representation shows that 210 participants presented dissatisfaction through their responses which included 90 "Very Dissatisfied" and 120 "Dissatisfied" ratings amounting to 45% of the total respondents. The study showed that users either had no satisfaction or expressed some form of satisfaction. Out of the total 640 respondents, 140 participants indicated satisfaction while 100 chose “Neutral.”

The displayed pattern reveals the widespread negative feelings users have about the AI moderation practices of platforms.

This data implies that users have both concerns about technology and questions about the implementation and enforcement of wider policy guidelines.

Table 4: Trust in Various Information Sources

Information Source	Trusted by (%)
Social Media	18
Mainstream News	24
Official Websites	32
Fact-Checking Platforms	15
Academic Sources	11

Table 4 shows the official Websites rank as the chief trusted information source with 32% endorsement while Mainstream News follows with 24%. Social Media remains one of the least trusted information sources according to 18% of the surveyed individuals. Fact-Checking Platforms together with Academic Sources rank lowest in user trust levels as respondents only trust them 15% and 11% respectively. The data implies that users trust institutional sources for reliability even though false content primarily spreads across social platforms.

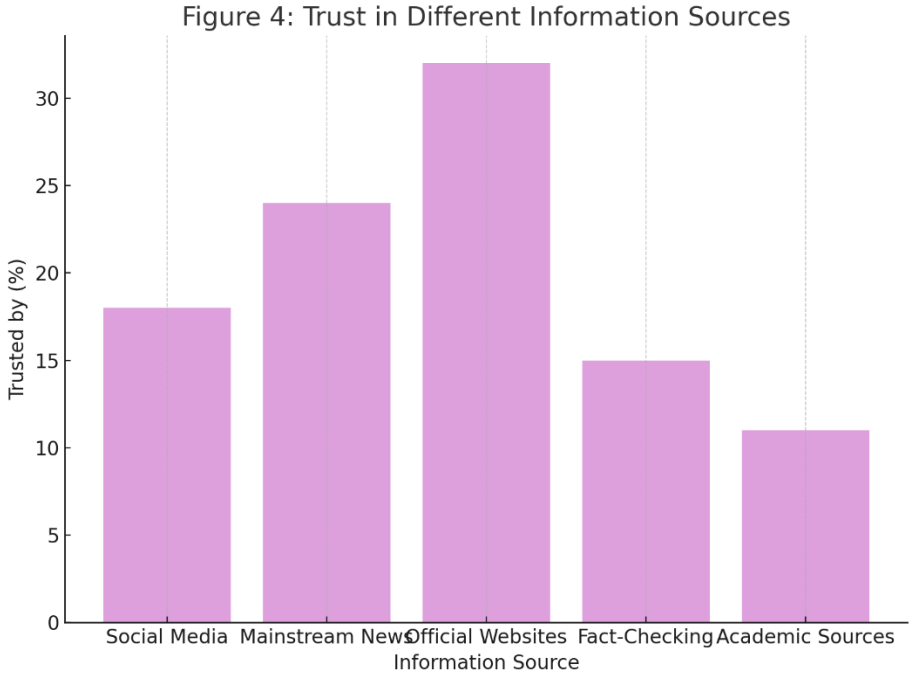

Figure 4: Trust in Various Information Sources

Public trust distribution data from different information sources is displayed through Figure 4. Public trust mainly rests with "Official Websites" since 32% of respondents consider them as the most reliable sources while "Mainstream News" receive 24% trust. The public trusts "Social Media" as an information source to only 18% extent according to survey results. From the public perspective Fact-Checking Platforms along with Academic Sources receive trust from only 15% and 11% respectively.

Universal social media usage for getting news fails to establish confidence among users despite being their main news platform. The credibility of official and traditional sources indicates that platform type directly influences how much concern people have about misinformation.

Table 5: Key Public Concerns Regarding AI Moderation

Concern About AI Moderation	Respondents (%)
Over-censorship	30
Bias in decisions	28
Lack of transparency	20
Inconsistent enforcement	12
Privacy intrusion	10

Table 5 shows the over-censorship with 30% and bias in decision-making with 28% ranked as the most significant worries among users. The data reveals that transparency concerns are ranked second among users with 20% and overcensorship remains the leading concern among 30% of respondents.

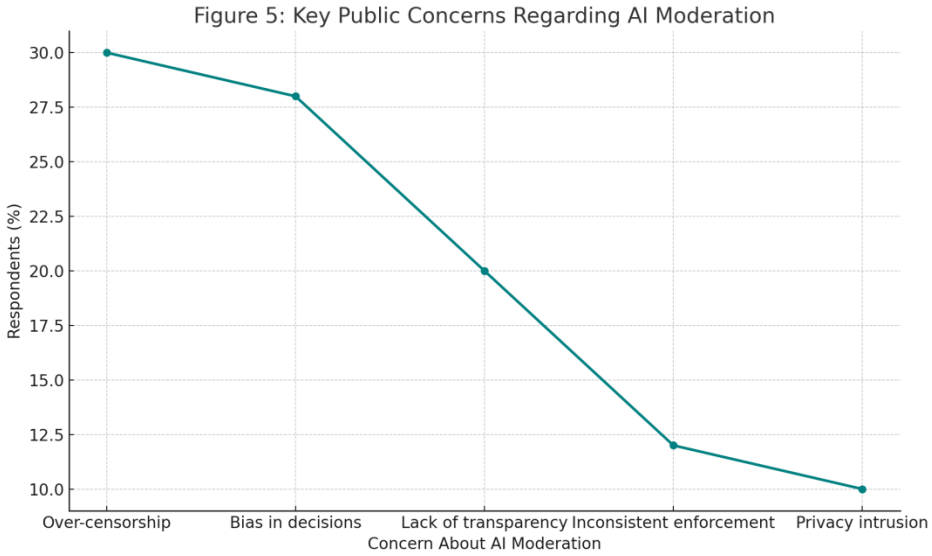

Figure 5: Key Public Concerns Regarding AI Moderation

The main public concerns about AI moderation appear in Figure 5 through a line graph presentation. Over-censorship emerges as the main issue according to survey respondents who comprise 30 percent of all participants. The concern about biased decision-making from AI systems reaches a level similar to over-censorship at 28% according to public opinion. A significant proportion of 20% of people expressed concerns about transparency and 12% reported inconsistent enforcement alongside 10% who worried about privacy intrusions.Survey results demonstrate that users concern themselves most about AI content moderation practices that are both harsh and biased although all surveyed issues remain relevant.

Qualitative Analysis

Analysis in this part stems from interviews conducted with 15 participants who include AI developers and platform moderators, policy experts and digital rights advocates. An examination of data according to Braun and Clarke's (2006) thematic analysis revealed four primary themes which expand the understanding of initial statistical findings.

Table 6: Thematic Analysis of Qualitative Interviews

Theme	Description	Representative Quote
Transparency and Explainability	Users are not provided clear feedback or reasoning when AI flags content.	"There was no proper explanation—just an automated message." – I6
Algorithmic Bias and Fairness	AI systems tend to disproportionately moderate content from marginalized groups.	"Queer advocacy content often gets taken down without reason." – I3
Fragility of Public Trust	Mistakes in AI moderation on political/health issues amplify user skepticism.	"People already distrust big tech; when AI fails, it just worsens." – I11
Regulatory and Governance Gaps	Participants noted a lack of cohesive, up-to-date regulation and global policy coordination.	"There's no global rulebook—platforms operate with impunity." – I9

Theme 1: Transparency and Explainability

AI-driven content moderation systems create a major concern since they operate in an unclear manner according to participant feedback. Users experience confusion because automated systems deliver vague or general notification messages when they lose their content which leads to dissatisfaction and unfairness. People pointed out that their inability to understand platform moderation and lack of appeal options wears down their trust in and engagement with

the service.
The automated system failed to provide an adequate reason for its actions. – Interviewee 6

Theme 2: Algorithmic Bias and Fairness
The theme shows how systematic bias manifests during artificial intelligence moderation activities. The interview subjects observed that marginalized communities faced excessive content removal through system actions. Training data contains limited diversity so algorithms falsely identify ethnic and linguistic diversity as violation content.
Several pieces of queer advocacy content disappear from online environments because moderators use extrajudicial powers to remove them. – Interviewee 3
Participants demanded the expansion of diverse data for AI training along with human moderator involvement for cases that required cultural understanding.

Theme 3: Fragility of Public Trust
The study found that AI systems exhibited trust shortcomings which made them easily breakable by outside factors. Experts noted that repeated wrong moderation of critical material such as political content and health-related and faith-based content leads users to lose confidence in technology platforms alongside their artificial intelligence systems.
The existing distrust toward big tech companies intensifies because of failed AI operations. – Interviewee 11
The stakeholders proposed rebuilding trust by teaching users using AI systems better and making AI decisions transparent while providing clear communication for content disputes.

Theme 4: Regulatory and Governance Gaps
Every research participant concurred about how legislation falls short when it comes to AI technology development. The interviewees identified the absence of a single global framework which regulates deepfake creation and misinformation.
Platform operations remain independent of any worldwide regulations that would hold them responsible. – Interviewee 9
The interviewees recommended that international organizations should work together for developing moral guidelines alongside detection prerequisites and enforcement systems that support AI-driven media governance.

Table 7: Expanded Subthemes from Interview Analysis

Subtheme	Description	Cited By
Opaque Appeals Process	Users find it difficult to challenge or understand automated content decisions.	Interviewee 2, 4, 8
Cultural Insensitivity	Algorithms often fail to interpret regional language, humor, or context properly.	Interviewee 1, 3, 5
Erosion of Platform Loyalty	Repeated AI errors lead users to abandon or mistrust platforms.	Interviewee 7, 10
Reactive Legal Frameworks	Laws act only after harm occurs and are not designed to prevent AI-based content misuse.	Interviewee 9, 13, 14

The qualitative subthemes present unique understandings about how artificial intelligence moderation affects users together with stakeholders.Several users found the system of automated content removal unappealable and baseless thus causing users to grow upset and increasingly disengage.Rewritten algorithms containing cultural insensitivity problems failed to understand regional languages and dialects and common cultural references which resulted in false content removals of benign or positive content.
Repeated errors on platforms caused users to both stop using the platforms and lose their trust in them particularly when unresolved issues repeated.Legal experts condemned present legislative measures because they failed to stop anticipated damage before it happened despite needing proactive governance approaches.The subthemes uncover both emotional as well as cultural and policy-level impacts of AI content moderation which goes further than superficial trust issues to demonstrate institutional knowledge gaps.

Discussion

Interpreting the Findings in Context

The research established a major decrease in public confidence toward AI-operated content moderation platforms. Almost half of study participants deemed AI technologies to exhibit inadequate performance leading to small confidence levels in their AI applications. Research conducted by Talker Research (2025) revealed that Americans have lost faith in online content to a level reaching 75% because of generative AI and deepfake technologies. Users experience low transparency about flagged content removal because they receive minimal explanations according to the Oversight Board (2024).

Gillespie (2018) found support in these outcomes because he demonstrated platform governance reveals minimal transparency and produces confusion among users regarding enforcement standards. According to West (2019) artificial intelligence continues to exercise governance without providing explanations which leads to a decrease in accountability systems. The unexplained computational processes of algorithmic systems establish a power imbalance between platform entities and their user base according to Bucher (2021). The findings presented in these works prove that transparency deficiencies constitute an essential problem for maintaining public trust.

Comparison with Previous Research

The identified algorithmic bias showing discrimination against marginalized populations backs previous academic observations. Noble (2018) provided well-known evidence that algorithmic bias affects search engines through their discriminatory targeting of racial and gender minorities. Our interviewees validated these same discrimination patterns that exist in AI content moderation operations.

An increased level of oversight demands attention from the European Centre for Algorithmic Transparency (ECAT, 2024) for ensuring fairness in high-impact systems according to the findings presented by O'Neil (2016). According to Sandvig et al. (2014) automated systems maintain inequalities when proactive controls fail to be established. Research data submitted in this study enhances existing concerns by demonstrating that current moderation systems cannot properly perceive varied linguistic and cultural expressions within different context-dependent situations.

Theoretical and Practical Implications

The theoretical part of our research adds to the expanding version of algorithmic governance through exposure of the socio-technical intricacies in AI moderation approaches. Recent studies and Kroll et al. (2017) show that technical solutions need ethical infrastructure to become effective because users maintain skepticism about fairness.

Gorwa, Binns and Katzenbach (2020) stated that platforms should jointly design governance structures with users to build better legitimacy when they practice co-design. By improving user interaction through explanation systems with feedback mechanisms and real-time help users could overcome their trust concerns. Mehrabi et al. (2021) strengthen their argument about conducting bias audits and employing inclusive data practices and this is aligned with our proposal to build development teams with diverse specialization alongside training datasets of diverse origin.

Policy-Level Considerations

The research supports existing demands for better regulations that should control the usage of AI-based content moderation systems.The proposed reforms stand as support for Balkin's (2015) idea of implementing responsible intermediary measures to grant democratic agencies oversight power over tech platforms.

The world faces challenges because different countries have not reached agreement regarding their AI regulations. The authors of Citron and Chesney (2019) stress the need for standardized deepfake regulations specifically because of illegal synthetic content distribution. The data matches recommendations which note difficulties in enforcing AI-based issues because interview participants showcased doubts about how effectively jurisdictions address AI-related risks. The United States "Take It Down Act" deals with intimate image abuse as a new law despite having inadequate comprehensive governance for managing artificial intelligence misinformation spread.

Acknowledging Limitations and Future Research

This study acknowledges its limitations. The study diversity does not guarantee complete representation of both non-English-speaking users and those who use the internet in loosely regulated regions. The quick changes in public understanding about AI require researchers to conduct long-term observations which track how trust and biased

perceptions evolve.

Studies need to evaluate how various population segments face algorithmic management and which clarity approaches prove best for each group. The previous studies conducted by Raji et al. (2020) concerning model auditing alongside Holstein and colleagues' (2019) work on fairness toolkits present foundational tools to evaluate platform interventions. The analysis of AI moderation responses across different legal settings and cultural backgrounds requires additional cross-national comparative research.

Conclusion

The research studied artificial intelligence functions as an anti-digital misinformation and deepfakes defense tool through investigations of detection systems and content moderation capabilities as well as public trust sectors. The research utilized mixed methods to uncover an essential problem where AI system implementation diverged from public acknowledgment of system results. Respondents exhibited distrust toward AI content moderation systems when questioned about their trust levels resulting in half of them expressing very low or low trust primarily because of reasoning related to system opacity and biases along with inconsistent enforcement practices. Marginalized communities expressed their emotional and ethical concerns regarding mistrust in interviews which highlighted the psychological aspects of misjudgments.

This study advances earlier research by utilizing detailed user-focused methods to analyze the subject. The outcomes of this study demonstrate how algorithmic systems affect social legitimacy together with platform management structures and civil liberties whereas earlier inquiries examined detection methods and systems' biased workings according to Noble (2018) and Mehrabi et al. (2021). The research demonstrates algorithmic obscurity continues to exist (Gillespie, 2018) while also showing exactly what factors such as missing appeal processes and cultural insensitivity work to undermine public faith in AI systems.

Facing the issues as observed in this research, various tactical approaches must be implemented. The first necessity for platforms involves improving AI system transparency while also ensuring their explainability functions in content moderation. Platform users should receive complete explanations about flagging and removal via simple communication tools and automated feedback systems. Public trust will strengthen through combination of explainable AI frameworks with algorithmic transparency reports which guarantee systematic accountability (Kroll et al., 2017; Raji et al., 2020).

AI development efforts need to incorporate independent oversight together with ethical examination into their standard operational framework. Customer rights along with user consent must serve as top priorities for academic institutions and technology companies because civil society organizations while designing and implementing their products. To maintain accountability in our automated digital age independentalgorithmic auditors need permission to inspect platform procedures while detecting violations (West, 2019, Raji et al., 2020).

References

Balkin, J. M. (2015). *Information fiduciaries and the First Amendment*. *UC Davis Law Review*, *49*(4), 1183–1234.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Bucher, T. (2021). *The algorithmic imaginary: How people understand and experience algorithms*. *Information, Communication & Society*, *24*(1), 1–18.

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753–1819.

European Commission. (2024). *European Centre for Algorithmic Transparency*. https://en.wikipedia.org/wiki/European_Centre_for_Algorithmic_Transparency

FAccT. (2024). *ACM Conference on Fairness, Accountability, and Transparency*. https://en.wikipedia.org/wiki/ACM_Conference_on_Fairness%2C_Accountability%2C_and_Transparency

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

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *CHI 2019* (pp. 1–15).

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–706.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Oversight Board. (2024). *Content moderation in a new era for AI and automation*. https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/

Raji, I. D., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Collected Essays*.

Talker Research. (2025). *Trust in online content report*. *New York Post*. https://nypost.com/2025/04/26/lifestyle/majority-of-americans-trust-whats-online-less-than-ever-before/

West, S. M. (2019). Data capitalism: Redefining the logics of surveillance and privacy. *Business & Society*, *58*(1), 20–41.

Associated Press. (2025). Tech industry tried reducing AI's pervasive bias. Now Trump wants to end its 'woke AI' efforts. https://apnews.com/article/8302e12dd74df69a1adc6565710f033d

AIMultiple. (2025). *Bias in AI: Examples and 6 ways to fix it in 2025*. https://research.aimultiple.com/ai-bias/