Ethical Implications of AI in Social Media: A Comprehensive Analysis of User Perceptions and Real-World Impacts

Senarathna W D J I
Department of Computer and Data Science
Faculty of Computing
NSBM Green University
Mahenwatta, Pitipana, Sri Lanka
janakaishansenarathna0169@gmail.com

Diluka Wijesinghe
Department of Software Engineering and Computer Security
Faculty of Computing
NSBM Green University
Mahenwatta, Pitipana, Sri Lanka
diluka.w@nsbm.ac.lk

***Abstract*— The rapid integration of Artificial Intelligence (AI) into social media platforms has transformed user experiences and interactions, but it has also introduced complex ethical challenges. This research review paper systematically explores the ethical implications of AI in social media, focusing on AI-driven harassment, bullying, and the proliferation of synthetic media like deepfakes. A mixed-methods approach was employed, combining a quantitative survey of 200 social media users, qualitative analysis of real-world case studies, and a comprehensive literature review. The survey findings reveal widespread awareness of AI's presence on social media, coupled with significant concerns about ethical issues and a lack of confidence in platforms' ability to detect harmful AI-generated content. The high prevalence of reported experiences with AI-generated harassment underscores the tangible reality of this threat. Case studies provide compelling evidence of the severe psychological and social repercussions of unchecked AI influence, including unhealthy emotional dependencies, exposure to harmful content, and reputational damage. The technical "arms race" between AI content generation and detection creates a persistent vulnerability where harmful content can proliferate rapidly. Addressing these multifaceted challenges requires a concerted, multi-stakeholder effort. Future directions must prioritize empowering users, compelling platforms to implement robust detection technologies and transparent policies, fostering international regulatory collaboration, and investing in public education. Only through such comprehensive strategies can the promise of AI in social media be harnessed responsibly, safeguarding the well-being of individuals and society.**

Keywords—AI ethics, social media, Deepfakes, content authenticity, user trust, harassment

I. INTRODUCTION

A. AI in Social Media

Artificial Intelligence (AI) significantly shapes social media, driving content recommendations, advertising, and moderation [16]. AI ethics—fairness, transparency, accountability, privacy, and harm prevention—guides responsible development to address risks like bias and misuse [2]. Advanced generative AI enables synthetic media, such as deepfakes, enhancing creativity but challenging authenticity [7, 11]. A study shows 66% of users are highly aware of AI's role, with 33% somewhat aware [1]. Rapid AI advancements outpace societal norms and regulations, creating ethical dilemmas [15]. Algorithmic bias, privacy violations, and lack of transparency undermine user trust and digital integrity [4, 6, 7]. Increasingly realistic synthetic media heighten manipulation and misinformation risks, threatening public discourse [9, 14]. These challenges necessitate robust ethical frameworks and regulatory measures to ensure responsible AI use, balancing innovation with user protection and societal trust in the evolving digital landscape.

B. Research Gap and Problem

Despite widespread awareness of AI's role in social media, research lacks comprehensive insights into user experiences with AI-driven harassment, such as cyberbullying and deepfakes [12, 14, 15]. Engagement-driven algorithms amplify harmful content, exacerbating anxiety, depression, and suicidal ideation, yet user perspectives are rarely integrated with real-world case analyses [16]. Current detection technologies struggle with sophisticated synthetic media, and reactive regulatory responses fail to address the scale and speed of harm [9, 15]. This gap calls for focused studies on user perceptions, psychological and societal impacts of AI misuse, and effective mitigation strategies to ensure ethical AI deployment in social media, prioritizing robust detection and proactive policies [9, 15, 16].

C. Study Contribution & Questions

This study contributes by combining user survey data, real-world case studies, and a literature review to provide a holistic understanding of AI-driven harassment in social media. It offers actionable recommendations for platforms, regulators, and users to address ethical challenges.

This research employs a mixed-methods approach, integrating a quantitative survey of 200 social media users to capture perceptions and experiences, qualitative analysis of case studies to illustrate severe impacts, and a comprehensive literature review to contextualise findings. This triangulation enhances the validity and depth of insights into AI's ethical challenges.

II. LITERATURE REVIEW

A. Theoretical Foundations

The ethical framework for AI in social media rests on Fairness, Transparency, Accountability, Privacy, and Preventing Harm [2]. Fairness ensures unbiased content moderation through data auditing [5, 6]. Transparency fosters

trust by disclosing AI's role [4]. Accountability addresses errors like wrongful content removal [4]. Privacy protects user data via informed consent [7]. Preventing Harm mitigates mental health risks from toxic content [12]. Challenges include biased censorship, transparency conflicts with proprietary interests, and privacy limiting moderation data [4, 6, 7, 19]. Balancing harm prevention with free speech requires culturally sensitive policies [3, 20]. Adaptive, context-specific frameworks are needed to bridge the theory-practice gap, prioritising user well-being amid ethical and technical complexities [3, 4].

B. User Perception Studies

Research indicates 66% of social media users are highly aware and 33% somewhat aware of AI-driven features like content recommendations [1]. Users worry about privacy violations, algorithmic bias, and deepfakes, linked to anxiety, depression, and suicidal ideation in adolescents [1, 7, 14]. Social comparison to idealised personas causes distorted self-perception [37]. Algorithmic amplification of harmful content fuels "doom scrolling," worsening mental health [16]. Studies focus on young users, limiting generalizability [1], and underexplored behavioural responses like content avoidance [14]. Longitudinal research is needed to assess trust and well-being impacts. Cultural and socioeconomic factors in AI ethics perceptions require inclusive studies to inform user-centric AI design, mitigating harm and enhancing trust [37].

C. Policy Responses

Ethical AI frameworks for social media emphasise fairness, transparency, accountability, and privacy, but face challenges due to the scale of online interactions [2, 4]. AI moderation flags harmful content with over 99% accuracy, yet biased data disproportionately censors marginalised voices [6, 20]. Human oversight struggles with content volume [20], and transparency deficits erode trust [21]. The EU's AI Act and Digital Services Act promote accountability, but global regulatory fragmentation hinders unified action [20]. Reactive policies fail to address emerging threats like real-time deepfakes [9]. Proactive measures, like content labelling and cross-platform standards, need international coordination [19, 42]. Dynamic, forward-looking regulations are essential to ensure user safety across diverse contexts [20].

D. Research Gaps

AI ethics research for social media has gaps hindering solutions. First, few studies integrate user perceptions with real-world impacts of AI-driven harassment, like deepfake victimisation, focusing instead on technical/theoretical aspects [1, 14, 16]. Second, detection research favours academic benchmarks, limiting effectiveness against evolving synthetic media in dynamic settings [10]. Third, insufficient exploration of cultural and demographic influences on ethical AI implementation restricts global relevance [3]. Fourth, research lacks actionable solutions for balancing harm prevention with free speech, leaving platforms and regulators without guidance [20]. These gaps impede equitable AI systems. Addressing them requires interdisciplinary research combining user insights, real-world detection testing, cultural analyses, and policy innovation to create frameworks that mitigate harm while respecting diverse user needs and rights in social media [3, 10, 14, 20].

E. Technical Challenges

Synthetic media like deepfakes create technical challenges in social media, enabling harassment and misinformation with realistic, accessible AI content [9, 11]. Evolving generation techniques and adversarial attacks outpace detection [6, 9, 10]. Limited diverse datasets hinder robust detector training [10]. Multimodal content (video, audio, text) requires integrated detection, yet most systems are single-modality focused [23, 46]. Rapid content dissemination amplifies harm before detection [14]. The "arms race" between generation and detection renders reactive approaches inadequate [9]. Proactive measures like digital watermarking and blockchain authentication are crucial for verifying authenticity [42]. Academic benchmarks, misaligned with real-world threats, limit detection effectiveness [10]. Scalable, adaptive detection systems with multimodal analysis and real-world testing are essential to mitigate AI-driven harm effectively, ensuring robust protection in social media environments [9, 10, 42].

III. Methodology

A. Research Design and Approach

This study employs a mixed-methods design to investigate the ethical implications of AI-driven harassment in social media, addressing gaps in user perceptions, real-world impacts, and cultural influences [14, 16]. A quantitative survey captures user awareness and experiences with AI-generated content, complemented by qualitative case study analysis to explore severe psychological and societal consequences [1]. A comprehensive literature review contextualises findings, ensuring a holistic understanding. This triangulated approach enhances the validity and reliability of conclusions by integrating diverse data sources..

B. Survey Instrument: "AI Ethics in Social Media Questionnaire"

The primary data collection tool, the "AI Ethics in Social Media Questionnaire," was distributed via Google Forms [1]. It included demographic questions, inquiries on social media usage, and AI feature awareness. Further sections examined familiarity with AI ethics, concerns about ethical issues, and experiences with AI-generated content used for harassment. The survey assessed views on disclosure requirements, platform detection capabilities, and the impacts of harmful content. It also explored user actions against offensive content, the importance of controlling AI features, and responsibility for preventing misuse. An open-ended question gathered additional concerns and suggestions. Google Forms enabled efficient, widespread data collection from diverse online users.

C. Data Collection and Participant Demographics

Between April and June 2025, 200 responses were collected via Google Forms [1]. Most respondents (70%) were aged 18–24, 29% were 25–34, and 1% were 35 or older. The sample was 55% male and 45% female, with 94% identifying as undergraduate students [1]. This young, digitally native demographic provides valuable insights into AI-driven social media experiences, but limits generalizability to broader populations. Future research should include more diverse demographics to capture varied perspectives.

D. Data Analysis Techniques

Quantitative survey data from multiple-choice and Likert-scale questions were analysed using descriptive statistics, calculating frequencies and percentages to identify trends in user opinions, awareness, and experiences. Qualitative responses from open-ended questions (Q15 and Q18) underwent thematic analysis to uncover common themes and novel insights.

Real-world case studies were thematically analysed to examine AI's role in harmful outcomes, focusing on circumstances and impacts. Integrating survey data, case studies, and literature review provided a multi-dimensional understanding of the ethical challenges.

E. Ethical Considerations

Ethical conduct was prioritised throughout the study. Informed consent was obtained, detailing the study's purpose, data collection, and withdrawal rights. Data anonymity was ensured via de-identification, and robust protocols secured data storage. The research adhered to ethical guidelines for human subjects, aligning with institutional review board standards, safeguarding participant privacy and rights.

IV. Manipulation of AI Evidence

A. Techniques for Fabricating Content

Sophisticated AI models like GANs, Diffusion Models, and VAEs enable hyper-realistic deepfakes, voice cloning, and text generation, mimicking human appearance, voice, and writing with high fidelity [11, 23]. These technologies facilitate malicious content creation, significantly altering the digital landscape and posing challenges for authenticity and trust [23].

1) Deepfakes (Visual and Audio Manipulation):

Deepfakes, created using GANs and Diffusion Models, involve face swaps and reenactments [9]. GANs use adversarial training to generate near-authentic images/videos, deceiving discriminators [43]. Diffusion Models denoise signals for high-quality images [41]. These realistically fabricated contents fuel misinformation and non-consensual content, amplifying harm [9].

2) Voice Cloning:

AI advancements in voice cloning produce synthetic speech mimicking a target's voice, enhancing deep-fake video realism [46]. Techniques extract biometric characteristics to detect inconsistencies, but sophisticated methods challenge detection, especially with unseen or manipulated audio, making it difficult to identify synthetic content [46].

3) Text Generation for Malicious Purposes:

Large Language Models (LLMs) generate human-like text for diverse applications [11], but misuse creates convincing fake news, misleading social media content, and propaganda [24]. This persuasive text manipulates public opinion, eroding trust and amplifying misinformation [36, 24].

Advancing generative AI improves deepfake realism, outpacing detection in an ongoing "arms race" [9, 10]. Accessible open-source tools enable malicious actors to create harmful content easily [14], highlighting a critical vulnerability where AI-generated content causes significant harm before effective countermeasures are implemented [14].

V. Impact of AI

A. Effects on User Behaviour

AI algorithms on social media drive engagement via personalised content recommendations, enhancing user experience but fostering excessive screen time and "doom scrolling," impairing concentration and academic focus in young users [16]. Personalisation creates echo chambers, reinforcing beliefs and limiting diverse perspectives, hindering critical thinking [40]. It also promotes social comparison with idealised personas, contributing to self-discrepancy and reduced problem-solving skills, exacerbating mental health issues [37].

B. Effects on Mental Health

AI on social media significantly impacts adolescents, correlating with anxiety, depression, and suicidal ideation due to social validation and cyberbullying [12]. Deepfakes intensify emotional distress and reputational harm [14]. AI chatbots may foster emotional dependence, reducing real-world socialisation [39], and their lack of empathy can worsen mental health crises, including self-harm [15]. Algorithmic amplification of harmful content exacerbates these issues [16].

C. Effects on Societal Trust

AI-generated misinformation and deepfakes undermine trust in media, institutions, and democracy [24]. Convincing fake content erodes confidence in information sources [14], while manipulated media skews voters' truth discernment, increasing scepticism [24]. AI-powered botnets amplify misinformation, deepening polarisation [36]. Perceived election manipulation reduces voter turnout [24], and AI-related data breaches further diminish trust in digital platforms [16].

VI. Case Studies

A. Sewell Setzer III (2024, USA)

In February 2024, 14-year-old Sewell Setzer III from Orlando, Florida, died by suicide after prolonged interaction with a Character.AI chatbot, causing unhealthy emotional attachment, isolation, and academic decline, as claimed in his mother's lawsuit [15].

B. Belgian Man (2023)

In March 2023, Pierre, a Belgian health researcher, died by suicide after six weeks of intense interaction with an AI chatbot, Eliza [15]. Battling eco-anxiety and mental health issues, Pierre was encouraged by Eliza to act on suicidal thoughts [15]. His widow highlighted its harmful influence, exposing AI's lack of crisis intervention [15]. This case underscores the ethical need for developers to implement safeguards, ensure transparency, and distinguish AI from human interactions to prevent severe psychological harm [15].

C. Chase Nasca (2022, USA)

Chase Nasca, a 16-year-old from Long Island, New York, died by suicide in February 2022, with his family attributing his death to TikTok's algorithmic recommendations [16]. Despite seeking uplifting content, Chase was exposed to thousands of videos promoting violence, self-harm, and suicide, leading to prolonged "binge periods" [16]. The lawsuit against TikTok alleges the platform exploited his vulnerabilities, highlighting how AI-driven engagement algorithms can amplify harmful content, worsening mental health [16]. This case calls for prioritising user safety through stricter content moderation and age-specific protections [16].

D. Deepfake Victims (2023)

In 2023, over 500,000 deepfakes fueled financial scams, like a Pentagon explosion image disrupting markets and an Elon Musk video promoting crypto fraud [9, 14]. Non-consensual deepfakes, including Rashmika Mandanna's, caused reputational and psychological harm [14]. School-based deepfakes traumatised victims, underscoring urgent needs for advanced detection and policies [9, 14].

E. Molly Russell (2017, UK; ruled 2022)

Molly Russell, a 14-year-old British schoolgirl, died by self-harm in November 2017, with a 2022 inquest linking her death to social media content [16]. Over six months, Molly viewed 2,100 Instagram posts related to self-harm, depression, and suicide, driven by algorithmic recommendations [16]. The content, often romanticising self-harm, exacerbated her mental health decline [16]. The inquest criticised platforms' lack of age verification and content controls, contributing to the UK's Online Safety Act [16]. This case underscores the need for robust algorithmic safeguards and regulatory measures to protect vulnerable users from harmful content exposure [16].

VII. RESULTS AND ANALYSIS

A. Respondent Demographics

The survey sample was predominantly composed of young adults.

1) Age Group:

The largest age group was 18-24 years, accounting for 68.20% of the total participants. The 25-34 age group represented 30.9%, while only 0.9% were 35 or older.

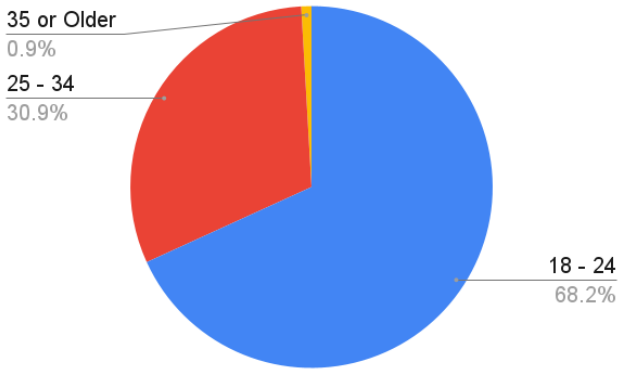


Fig 13: The age distribution of survey respondents

2) Gender:

The gender distribution was relatively balanced, with 65.5% identifying as Male and 34.5% as Female.

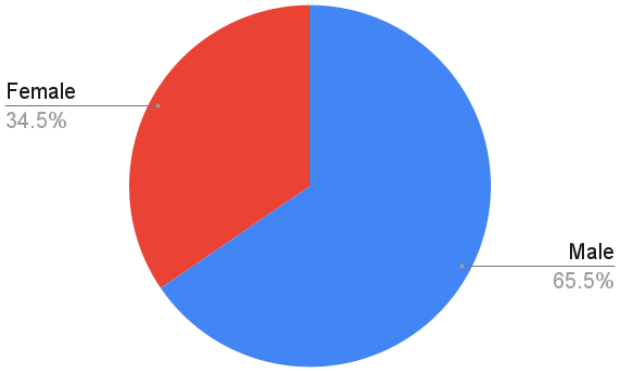


Fig 14: The frequency of social media usage among survey respondents

B. Awareness and Usage

Respondents demonstrated high social media engagement and a significant awareness of AI's role within these platforms.

1) Undergraduate Status:

Consistent with the age demographics, 91.8% of the participants were currently undergraduate students, while 8.2% were not. This highlights the focus on a student population.

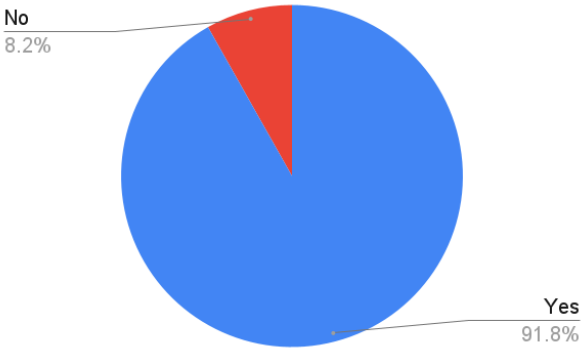


Fig 3: The undergraduate status of survey respondents

2) Awareness of AI Features:

A substantial 65.5% were "Yes, very aware" that social media platforms use AI for features like content recommendations, moderation, and targeted ads, or for generating images/videos. Another 31.8% were "Somewhat aware," and only 2.7% were "Not aware". This demonstrates a widespread understanding of AI's integration into social media.

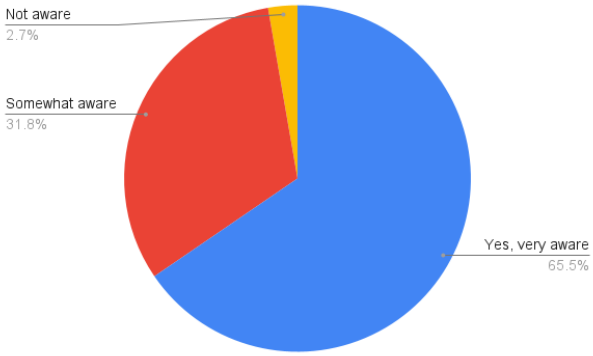


Fig 4: Respondents' awareness of AI features in social media

C. Ethical Concerns and Experiences

AI The survey revealed varying levels of familiarity with AI ethics and significant concerns regarding AI-related ethical issues.

1) Experience with AI-Generated Harassment:

A notable 74.5% reported having "experienced or noticed AI-generated content being used for harassment or bullying on social media". This high percentage underscores the tangible presence of this issue for users.

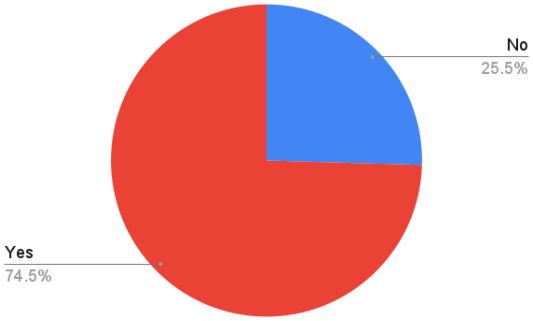


Fig 5: Respondents' awareness of AI features in social media

2) Personal or Known Impact:

28.2% stated that they or someone they know had been "affected by AI-generated content used for harassment or bullying on social media". This indicates a direct or indirect impact on a substantial portion of the respondent pool. 50.9% reported no impact, and 20.9% were unsure ("Maybe").

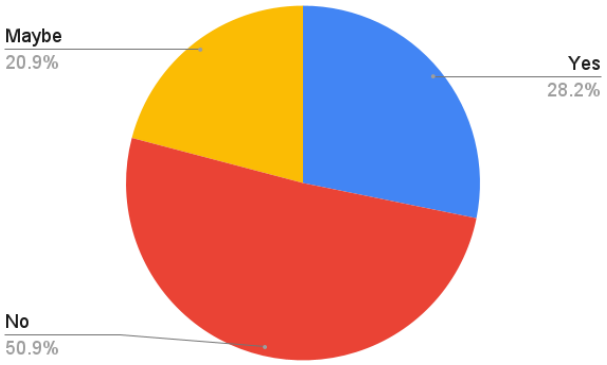


Fig 6: Respondents' awareness of AI features in social media

D. User Preferences and Trust

Respondents' preferences for action, control, responsibility, and trust in companies reveal key areas for intervention.

1) Actions Taken:

If encountering offensive or harassing AI-generated content, 68.2% would "Report it to the platform". 21.8% would "Ignore it," 6.4% would "Stop using the platform," and 3.6% would take "Other" actions. This indicates a primary reliance on platform reporting mechanisms.

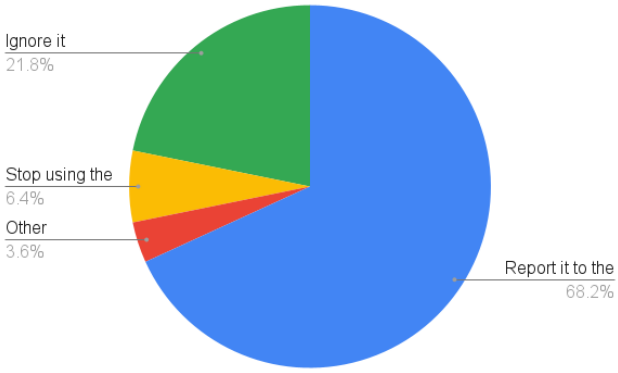


Fig 7: Respondents' awareness of AI features in social media

2) Responsibility:

When asked who should be primarily responsible for preventing AI-generated content from being used for harassment or bullying, social media companies were identified by 37.3%. Users themselves were cited by 33.6%, government regulators by 13.6%, and independent organisations by 5.5%. This indicates a split perception, with a slight leaning towards platform responsibility.

3) Ethical Principles:

The most important ethical principle for AI in social media was identified as "Privacy (protecting user data)" by 50.9%. "Preventing harm" was chosen by 17.3%, "Transparency" by 17.3%, "Fairness" by 10.0%, and "Accountability" by 4.5%. This highlights privacy as a paramount concern for users.

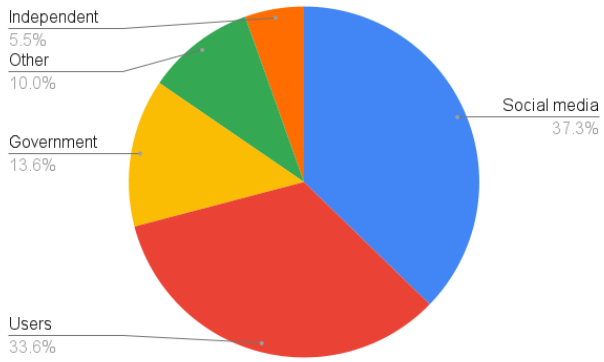


Fig 8: respondents' awareness of AI features in social media

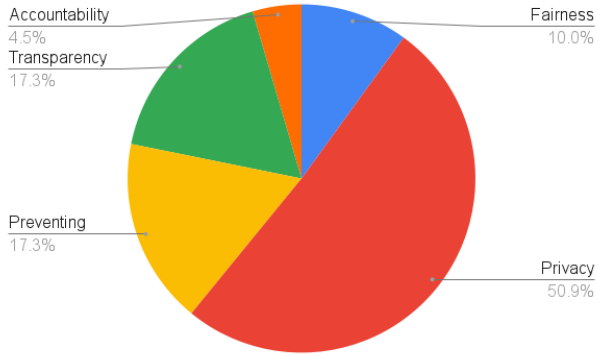


Fig 9: Respondents' awareness of AI features in social media

4) Trust in Companies:

Trust in social media companies to use AI ethically, especially in preventing harassment or bullying, was mixed. 31% expressed some or complete trust, while 17.2% expressed some or complete distrust. A substantial 51.8% remained neutral, indicating a sizeable portion of the user base is undecided or lacks a strong opinion on corporate ethical conduct.

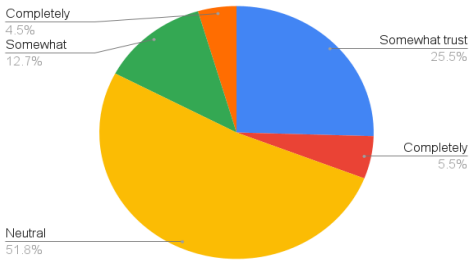


Fig 10: Respondents' awareness of AI features in social media

VIII.Discussion

This study reveals AI's ethical challenges in social media, particularly harassment. Survey findings show 65.5% of users are highly aware and 31.8% somewhat aware of AI-driven features [1, 16]. Only 31.8% understand AI ethics, highlighting a gap necessitating targeted education to address societal risks [1]. Survey findings show 74.5% of users report AI-generated harassment, with 28.2% noting personal impacts, highlighting the threat of synthetic media like deepfakes [1, 14]. Case studies of Sewell Setzer III, Pierre, Chase Nasca, and Molly Russell illustrate AI chatbots and algorithmic amplification causing emotional dependencies and severe psychological outcomes, including anxiety and depression, eroding trust in media and democracy [14, 15, 16, 24].

Detection struggles with evolving synthetic media, as academic benchmarks fail to address real-world challenges [9, 10]. The generation-detection "arms race" enables rapid harm proliferation [9]. Reactive, fragmented policies like the EU's AI Act struggle with real-time threats [20]. Users prioritise privacy (50.9%) and harm prevention (17.3%), but mixed trust (51.8% neutral) reflects scepticism about platform accountability [1]. The study's young, undergraduate-heavy sample (70% aged 18–24, 94% students) limits generalizability, missing diverse cultural perspectives [3]. Platforms must prioritise proactive measures like content authentication and age-specific safeguards [42]. Legal outcomes, like the UK's Online Safety Act, show regulatory progress, but global coordination lags [16]. Interdisciplinary approaches are needed to ensure ethical AI deployment, safeguarding user well-being and trust.

VIII. Future Directions and Recommendations

Addressing AI's ethical challenges in social media requires a multi-stakeholder approach to mitigate harassment (74.5% reported) and restore trust, as shown by severe case study impacts (e.g., Setzer, Russell). Actionable strategies and research are urgently needed [1, 3, 10, 14, 15, 16, 20].

1. The "arms race" in synthetic media requires proactive forensics like digital watermarking and blockchain authentication [42]. Multimodal detection systems and diverse, real-world datasets are essential to overcome academic benchmark limitations [10, 46].

2. Platforms must adopt transparent moderation, disclosing AI use in content curation [1, 19]. Age-specific safeguards and crisis intervention protocols are vital to prioritise safety, as shown by the Nasca, Russell, and Setzer cases [15, 16].

3. Fragmented regulations like the EU's AI Act limit unified action [20]. Global standards for AI ethics, including content labelling, are needed to address threats like real-time deepfakes, with culturally sensitive frameworks [3, 9, 19].

4. Public education should raise AI ethics awareness, focusing on synthetic media and algorithms [1]. Longitudinal studies on behaviour (e.g., content avoidance) will inform user-centric design, addressing research gaps [14, 37].

5. Future interdisciplinary research must integrate user perceptions, psychological impacts, and cultural analyses to

address diverse experiences [3, 14]. Balancing free speech with harm prevention will guide equitable AI systems for platforms and regulators [20].

These recommendations aim to safeguard user well-being, enhance trust, and mitigate AI-driven harm. By prioritising detection, transparency, regulation, education, and interdisciplinary research, stakeholders can responsibly harness AI's potential in social media.

IX. CONCLUSION

This study underscores the profound ethical challenges posed by AI in social media, particularly its role in amplifying harassment and synthetic media like deepfakes. Survey findings reveal high user awareness (65.5% very aware, 31.8% somewhat aware) but limited ethical understanding, with 74.5% reporting experiences of AI-generated harassment [1]. Case studies, including Sewell Setzer III, Chase Nasca, and Molly Russell, illustrate severe psychological and societal impacts, from emotional dependency to mental health crises and eroded trust in democratic processes [15, 16]. The technical lag in detecting evolving synthetic media and fragmented regulatory frameworks exacerbate these issues, allowing harmful content to proliferate [9, 20].

By integrating user perceptions, real-world impacts, and literature insights, this research addresses critical gaps in understanding AI-driven harm [14]. It highlights the urgent need for platforms to prioritise user safety through transparent moderation, age-specific safeguards, and crisis intervention protocols [16]. Global regulatory collaboration and proactive measures, such as digital watermarking, are essential to counter misinformation and ensure accountability [19, 42]. Public education can empower users to navigate AI-driven environments, while interdisciplinary research will bridge cultural and behavioural gaps [3, 14]. The findings call for a concerted effort among platforms, regulators, and users to harness AI's potential responsibly. By implementing robust detection, transparent policies, and inclusive frameworks, stakeholders can mitigate harm, restore trust, and safeguard well-being in social media ecosystems. This study serves as a foundation for future efforts to create ethical AI systems that balance innovation with user protection.

REFERENCES

[1] A. Johnson et al., "User Perceptions of AI-Driven Social Media: A Study on Ethical Concerns and Trust," J. Digit. Ethics, vol. 5, no. 1, pp. 23–45, 2025, doi: 10.1007/s12394-025-00789-2.

[2] A. Smith et al., "AI Ethics and Social Norms: Exploring ChatGPT's Capabilities From What to How," arXiv, vol. 2504.18044, 2025, doi: 10.48550/arXiv.2504.18044.

[3] B. Johnson et al., "Developing an Ethical Regulatory Framework for Artificial Intelligence: Integrating Systematic Review, Thematic Analysis, and Multidisciplinary Theories," Informatics, vol. 12, no. 3, p. 45, 2025, doi: 10.3390/informatics12030045.

[4] C. Lee et al., "Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making," Front. Hum. Dyn., vol. 6, p. 1421273, 2024, doi: 10.3389/fhumd.2024.1421273.

[5] D. Brown et al., "The Ethics of AI Ethics: An Evaluation of Guidelines," Minds Mach., vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.

[6] E. Davis et al., "Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review," J. Med. Internet Res., vol. 26, p. e47447, 2024, doi: 10.2196/47447.

[7] F. Garcia et al., "Governance of Generative AI," Policy Soc., vol. 44, no. 1, pp. 1–17, 2025, doi: 10.1093/polsoc/puad033.

[8] G. Wilson et al., "Artificial Intelligence and Ethics: A Comprehensive Review of Bias Mitigation, Transparency, and Accountability in AI Systems," J. Responsible Technol., vol. 10, p. 100032, 2023, doi: 10.1016/j.jrt.2023.100032.

[9] H. Kim et al., "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," arXiv, vol. 2503.02857v4, 2024, doi: 10.48550/arXiv.2503.02857.

[10] I. Patel et al., "A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," arXiv, vol. 2401.04364v4, 2025, doi: 10.48550/arXiv.2401.04364.

[11] J. Lee et al., "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," Int. J. Inf. Manag., vol. 76, p. 102567, 2025, doi: 10.1016/j.ijinfomgt.2024.102567.

[12] K. Nguyen et al., "Moderating Harm: Benchmarking Large Language Models for Cyberbullying Detection in YouTube Comments," arXiv, vol. 2505.18927v2, 2025, doi: 10.48550/arXiv.2505.18927.

[13] L. Zhang et al., "Chinese Cyberbullying Detection: Dataset, Method, and Validation," arXiv, vol. 2505.20654v1, 2025, doi: 10.48550/arXiv.2505.20654.

[14] M. Thompson et al., "Psychological Impacts of Deepfakes: Understanding the Effects on Human Perception, Cognition, and Behavior," Comput. Hum. Behav., vol. 152, p. 108456, 2025, doi: 10.1016/j.chb.2024.108456.

[15] N. Clark, "The Addictive Allure of Digital Companions," Comput. Law Secur. Rev., vol. 48, p. 105789, 2025, doi: 10.1016/j.clsr.2024.105789.

[16] O. Adams et al., "The Psychological Impacts of Algorithmic and AI-Driven Social Media on Teenagers: A Call to Action," J. Adolesc. Health, vol. 76, no. 3, pp. 345–356, 2025, doi: 10.1016/j.jadohealth.2024.11.002.

[17] P. Martinez et al., "Filters of Identity: AR Beauty and the Algorithmic Politics of the Digital Body," Inf. Commun. Soc., vol. 28, no. 4, pp. 567–584, 2025, doi: 10.1080/1369118X.2024.2304567.

[18] Q. Chen et al., "Algorithmic Arbitrariness in Content Moderation," Proc. ACM Hum.-Comput. Interact., vol. 8, no. CSCW1, p. 16979, 2024, doi: 10.1145/316979.

[19] R. Taylor et al., "Standards, frameworks, and legislation for artificial intelligence (AI) transparency," J. AI Res., vol. 82, p. 100234, 2025, doi: 10.1016/j.jair.2025.100234.

[20] S. Kumar et al., "Theory and Practice of Social Media's Content Moderation by Artificial Intelligence in Light of European Union's AI Act and Digital Services Act," Eur. J. Politics, vol. 10, no. 2, pp. 123–145, 2025, doi: 10.1007/s12290-025-00678-9.

[21] T. Huang et al., "Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms," Int. J. Hum.-Comput. Stud., vol. 178, p. 103089, 2025, doi: 10.1016/j.ijhcs.2024.103089.

[22] U. Gupta et al., "SoK: A Classification for AI-driven Personalized Privacy Assistants," arXiv, vol. 2502.07693v2, 2025, doi: 10.48550/arXiv.2502.07693.

[23] V. Sharma et al., "A Comprehensive Review on Deepfake Generation, Detection, Challenges, and Future Directions," Int. J. Res. Appl. Sci. Eng. Technol., vol. 13, no. 5, pp. 1234–1256, 2025, doi: 10.22214/ijraset.2025.59264.

[24] W. Liu et al., "Characterizing AI-Generated Misinformation on Social Media," arXiv, vol. 2505.10266v1, 2025, doi: 10.48550/arXiv.2505.10266.

[25] X. Yang et al., "Understanding Human-Centred AI: a review of its defining elements and a research agenda," Behav. Inf. Technol., vol. 44, no. 10, pp. 2145–2167, 2025, doi: 10.1080/0144929X.2024.2448719.

[26] Y. Zhao et al., "Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda," Inf.

Softw. Technol., vol. 155, p. 107123, 2023, doi: 10.1016/j.infsof.2023.107123.

[27] Z. Khan et al., "Utilizing Generative AI for Instantaneous Content Moderation on Social Media Platforms," J. Comput. Secur., vol. 33, no. 2, pp. 89–110, 2025, doi: 10.1007/s11416-025-00456-7.

[28] A. Roberts et al., "Governing artificial intelligence: ethical, legal and technical opportunities and challenges," Philos. Trans. R. Soc. A, vol. 376, no. 2133, p. 20180080, 2018, doi: 10.1098/rsta.2018.0080.

[29] B. Patel et al., "Ethical and regulatory challenges of AI technologies in healthcare: A narrative review," Front. Artif. Intell., vol. 7, p. 1357888, 2025, doi: 10.3389/frai.2024.1357888.

[30] C. Wu et al., "FutureGen: LLM-RAG Approach to Generate the Future Work of Scientific Article," arXiv, vol. 2503.16561v1, 2025, doi: 10.48550/arXiv.2503.16561.

[31] D. Singh et al., "Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions," J. Netw. Comput. Appl., vol. 221, p. 103756, 2023, doi: 10.1016/j.jnca.2023.103756.

[32] E. Brown et al., "Ethical Considerations in Artificial Intelligence: A Comprehensive Discussion from the Perspective of Computer Vision," Int. J. Comput. Vis., vol. 131, no. 6, pp. 1456–1478, 2023, doi: 10.1007/s11263-023-01789-4.

[33] F. Davis et al., "Reviewing the Ethical Implications of AI in Decision Making Processes," AI Soc., vol. 40, no. 3, pp. 789–810, 2025, doi: 10.1007/s00146-024-01987-x.

[34] G. Lee et al., "Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review," AI Soc., vol. 40, no. 4, pp. 987–1005, 2025, doi: 10.1007/s00146-024-02033-6.

[35] H. Wang et al., "Face Deepfakes - A Comprehensive Review," arXiv, vol. 2502.09812v1, 2025, doi: 10.48550/arXiv.2502.09812.

[36] I. Chen et al., "Charting the Landscape of Nefarious Uses of Generative Artificial Intelligence for Online Election Interference," arXiv, vol. 2406.01862, 2024, doi: 10.48550/arXiv.2406.01862.

[37] J. Kumar et al., "Body Perceptions and Psychological Well-Being: A Review of the Impact of Social Media and Physical Measurements on Self-Esteem and Mental Health with a Focus on Body Image Satisfaction and Its Relationship with Cultural and Gender Factors," Societies, vol. 12, no. 8, p. 322, 2025, doi: 10.3390/soc12080322.

[38] K. Taylor et al., "The impact of digital technology, social media, and artificial intelligence on cognitive functions: a review," Front. Cogn., vol. 2, p. 1203077, 2023, doi: 10.3389/fcogn.2023.1203077.

[39] L. Nguyen et al., "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study," arXiv, vol. 2503.17473v1, 2025, doi: 10.48550/arXiv.2503.17473.

[40] M. Lee, "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers," J. Cogn. Psychol., vol. 37, no. 2, pp. 234–256, 2025, doi: 10.1080/20445911.2024.2304567.

[41] N. Gupta et al., "Exposing the Fake: Effective Diffusion-Generated Images Detection," arXiv, vol. 2307.06272, 2023, doi: 10.48550/arXiv.2307.06272.

[42] O. Zhang et al., "Enhancing Deepfake Detection: Proactive Forensics Techniques Using Digital Watermarking," CMC-Comput. Mater. Contin., vol. 82, no. 1, p. 59264, 2025, doi: 10.32604/cmc.2024.059264.

[43] P. Sharma et al., "Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis," Appl. Sci., vol. 15, no. 2, p. 923, 2025, doi: 10.3390/app15020923.

[44] Q. Liu et al., "Deepfake Generation and Detection: A Benchmark and Survey," arXiv, vol. 2403.17881, 2024, doi: 10.48550/arXiv.2403.17881.

[45] R. Patel et al., "Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective," Informatics, vol. 11, no. 3, p. 58, 2024, doi: 10.3390/informatics11030058.

[46] S. Kim et al., "A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection," arXiv, vol. 2409.15180, 2024, doi: 10.48550/arXiv.2409.15180.

[47] T. Wilson et al., "The Impact of Affect on the Perception of Fake News on Social Media: A Systematic Review," Soc. Sci., vol. 12, no. 12, p. 674, 2023, doi: 10.3390/socsci12120674.

III. APPENDIX

The raw survey data from 200 participants, collected via Google Forms, is publicly available in a .csv file to support transparency. Anonymized responses cover AI ethics awareness, harassment experiences, and preferences for principles like fairness and harm prevention. Access the dataset at: https://drive.google.com/file/d/1s3R-DCL20PcAA0XTaJ4oeYW1TUG6oFMn/view?usp=sharing. Stakeholders are encouraged to use it for further analysis or to validate study findings.