College Students' Perceptions of AI-based Bullying Prevention: Ethical Safeguards, Human Oversight, and Responsible System Design

Suhair Amer
Department of Computer Science
Southeast Missouri State University
Cape Girardeau, Missouri, USA

Seth Turner
Department of Computer Science
Southeast Missouri State University
Cape Girardeau, Missouri, USA

ABSTRACT
Artificial Intelligence (AI) systems increasingly influence communication, education, and decision-making processes. While these systems can promote positive interaction, they also risk enabling or amplifying bullying behaviors—either by generating harmful content, reinforcing social biases, or allowing harassment on AI-driven platforms. This paper examines analytical and ethical strategies to prevent bullying in AI systems. It explores design considerations, dataset ethics, moderation mechanisms, and alignment methods aimed at fostering empathetic, safe, and equitable AI interactions. Recommendations include ethical data curation, multi-layered safety filters, human-in-the-loop moderation, and digital literacy initiatives to ensure AI supports rather than harms users.

Keywords
Artificial Intelligence; Bullying Prevention; Ethical AI; Human Oversight; Digital Well-being

1. INTRODUCTION
AI technologies now mediate a significant portion of human communication, from chatbots to recommendation engines. However, poorly designed systems may unintentionally enable bullying, whether through biased outputs, toxic responses, or unmoderated user interactions. Bullying in AI contexts encompasses user-to-user harassment facilitated by algorithms, AI-generated harmful language, and systemic bias that disadvantage vulnerable groups. Therefore, ensuring safety, empathy, and fairness is central to responsible AI design and deployment.

In traditional contexts, bullying refers to repeated aggressive behavior that causes harm or distress. Within AI, this can manifest as: AI-to-user bullying: when a system outputs harmful, mocking, or manipulative responses, User-to-user bullying via AI: when algorithms amplify harassment or expose targets to abusive content or algorithmic or systemic bullying: when biases in datasets or models reinforce discrimination or marginalization. Preventing these outcomes requires both technical alignment and ethical design principles. This paper analyzes how to prevent bullying in AI through ethical data handling, algorithmic safeguards, and proactive governance.

2. LITERATURE REVIEW
Perla et al. presented the results of a research/training project performed in collaboration with two comprehensive Apulian schools. The research was aimed at primary and secondary teachers and used various methodologies and tools. These included AI, hackathon, and storytelling. The goal was to acquire the practical skills necessary to identify, prevent, and manage bullying scenarios within the teaching environment [1]. Gao and Ye developed an automatic physical and verbal method for detecting bullying in the field of AI. Its features were extracted from acceleration and gyro data in order to train the physical detection method while the mean value of each frame of data samples was calculated for verbal detection. The authors utilized the *k*-NN algorithm as the classifier with final accuracies of 70.4% and 78.0% for physical and verbal detection, respectively. These results indicate that activity recognition can be used to detect bullying as an AI technique and that speech recognition appeared better than activity recognition [2].

Milosevic et al. designed a set of AI-based interventions in cyberbullying that built upon the proactive take-down of bullying-related content based on social learning and social norm theories. This was done by solicitating children's views through focus groups and interviews about their perceived effectiveness and the impact said interventions on the rights of children (from Ireland, aged 12 – 17, $N = 59$). This is a first-of-its-kind study that solicits children's views on the topic and its implications for their privacy and freedom of speech [3]. Mendoza-Pinto addressed the creation of an AI-based chatbot implemented with a topic- and equity-sensitive design. Its goal is to provide emotional support to victims of bullying using a combination of exploratory and descriptive approaches. The chatbot proved capable of providing basic emotional support and guidance to victims. It has the potential to empower victims and improve school safety with the use of integrated ChatGPT software. It also features intuitive interaction through Telegram, providing relevant and appropriate responses to user queries [4].

Asrifan stressed the importance of AI-based moderation tools for online interaction management while also requiring ethics, transparency, and equity. He states that effective cyberbullying prevention includes protection of privacy, algorithmic bias analysis, and education for digital citizens. Systems are further advanced by improved emotional intelligence, tailored well-being assistants, and decentralized safety protocols that help AI create safer online spaces [5]. Chaudhary examined the creation and application of AI-based bullying prevention systems as cutting-edge solutions to the issue. He highlights the ubiquity of online harassment and its effects on individuals and their communities as well as the difficulties preventing bullying in the digital realm and the limitations of conventional techniques. Ai plays a major role in content analysis, user profiling, real-time monitoring, and alert protection to detect and prevent bullying [6].

Chaudhary et al. developed a system with a detection accuracy of 92.5%, with 89% precision and 95% recall. This system surpasses traditional bullying prevention methods by 15% in

accuracy and 10% in recall. Response times exhibited included an average 0.5 seconds for flagging and classifying cyberbullying incidents. A 30% reduction in cyberbullying escalation and a 25% increase in utilized support resources by victims were made possible by proactive prevention strategies including counter-narrative generation and positive interventions. The system maintained consistent performance across datasets over a six-month period, showing a 5% increase in detection accuracy [7]. Singh and Parker highlight how different AI technologies, including natural language processing (NLP) techniques such as sentiment analysis, keyword detection, and contextual interpretation, and well as machine learning models, are advancing bullying detection and prevention. Real-world scenarios demonstrate such applications in real-time monitoring and user behavior analysis, which offer valuable insights [8].

Orrù et al. produced the BullyBuster system, which integrates AI techniques with psychological models to comprehensively understand and combat bullying issues. This technology allows automatic identification of potentially harmful content by analyzing linguistics and behaviors in various sources of data such as photos and videos. The system then alerts these threats to authorities or moderators, which allows for rapid interventions and mitigation [9]. Aggarwal and Guar highlight how scholars and experts have turned to AI techniques as proactive strategies for prevention and intervention in bullying. Their research focuses on women's safety in the digital realm. The authors state that online harassment has detrimental effects on the welfare and online encounters of women. Their method of prevention includes automated identification and response to harmful content, promoting more secure online spaces for women [10].

Sinaga et al. state that, at an early stage, AI can play a role in preventing extreme behavior development in children. Later on, a program can be utilized as part of parent counseling for the mental health of children. An early learning function should be deployed and permanent protection must also be added. The results of this study show that AI solutions can reduce bullying occurrences, lessen their impact on children, and provide victim protection. AI-enhanced video games have also been used to identify motivations behind bullying behavior by analysis of audio/visual data. They can play calming music and animations or mentor behavior discussions when attempts at bullying are detected [11]. Saputra et al. utilized a qualitative method with a case study approach, focusing on in-depth exploration of specific phenomena. This allows researchers to understand how AI can detect bullying. The data was collected via a questionnaire with 572 responses consisting of students, teachers, lecturers, and educational personnel in Sumedang Regency. The data was analyzed using the three-stage, Miles and Huberman model. The results indicate that AI has potential, but human supervision is necessary to verify results [12].

Sidhu and Sidhu utilized NLP algorithms and computer vision techniques to explore AI in early detection of bullying patterns, allowing for swift intervention. They also state that chatbots offer immediate support and guidance for victims and perpetrators while real-time sentiment analysis helps educators in gauging emotional climates. Personalized learning platforms and peer support networks also use AI to provide support and rehabilitation for everyone involved [13]. González et al. proposed an ICT tool to prevent bullying. The tool integrates several validated tests to determine potential risks of individuals becoming a bully or a victim thereof. It also provides a notification mechanism for anonymously reporting bullying situations. With both components working in tandem, this tool can determine what action should be considered by teachers and educational staff to mitigate bullying and prevent future incidents [14]. Angelis and Perasso proposed machine learning (ML) as a method for bullying detection and prevention. ML enables systems to quickly access and learn from data in order to make decisions. A systematic review of ten literature reviews shows that most models use content-based features. The most commonly used algorithms are Support Vector Machine, Naïve Bayers, and Convolutional Neural Networks. ML has the potential to optimize and integrate educational programs for children and can be a starting point for automated bullying detection systems [15].

3. DATA COLLECTION

Data were collected from undergraduate and graduate college students using a structured online questionnaire designed to examine perceptions of artificial intelligence (AI) in preventing bullying within educational environments. A convenience sampling method was employed, with participants recruited through university email lists, student organizations, and social media platforms. Participation was voluntary, and respondents were informed about the purpose of the study before completing the survey.

The questionnaire consisted of two sections. The first section gathered demographic information, including age, gender, year of study, and field of study. The second section measured students' perceptions of AI-based bullying prevention systems, including their awareness of AI technologies, perceived effectiveness, trust in AI, privacy concerns, and willingness to support the implementation of AI tools for bullying prevention. Responses were measured using a five-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Prior to the full-scale data collection, the survey instrument was pilot-tested with a small group of students to assess clarity, reliability, and validity. Feedback from the pilot study was incorporated to improve the wording and organization of the questionnaire.

Figure 1 shows the age distribution of the respondents indicates that the majority of participants were young adults. Most respondents were between 18 and 24 years old (56.9%), followed by those aged 25–34 years (39.7%). Only a small proportion of participants were in the 35–44 age group (1.7%) or were 45 years and older (1.7%), while no respondents were under the age of 18. These findings suggest that the sample primarily represents the perspectives of college students in the traditional and early adult age ranges. Figure 2 shows the gender distribution of the respondents indicates that male participants constituted the largest proportion of the sample, followed by female participants. A small number of respondents chose not to disclose their gender by selecting "Prefer not to say." Overall, the distribution reflects representation from multiple gender groups, with a greater proportion of male respondents contributing to the study on the use of artificial intelligence for bullying prevention among college students.

Figure 3 shows that the respondents generally demonstrated a high level of familiarity with artificial intelligence. The majority reported being very familiar with AI (59.6%), while a substantial proportion indicated that they were somewhat familiar (31.6%). Only a small percentage of respondents reported that they were not familiar with AI (8.8%). These findings suggest that most participants possessed at least a basic understanding of artificial intelligence, making them well-positioned to provide informed perspectives on its potential role in preventing bullying in college settings.

Figure 4 shows how the respondents reported varying levels of exposure to bullying or harmful content on digital platforms. The largest proportion indicated that they rarely encountered such content (38.6%), followed by those who reported encountering it sometimes (29.8%). A smaller percentage of respondents stated that they encountered bullying or harmful content very often (19.3%), while the fewest respondents indicated that they had never encountered it (12.3%). These findings suggest that although frequent exposure is not common for most participants, bullying and harmful content remain a recurring issue on digital platforms.

Figure 5 demonstrates that respondents showed mixed views regarding the potential of artificial intelligence systems in reducing bullying. The largest proportion of participants expressed uncertainty about AI's effectiveness in addressing bullying (43.9%). A slightly smaller group believed that AI systems can help reduce bullying (26.3%), while a comparable proportion did not believe that AI systems would be effective in doing so (29.8%). Overall, these responses indicate a divided perception among participants, with a notable level of uncertainty about the role of AI in mitigating bullying behaviors.

In figure 6, the respondents' confidence in the ability of current AI systems to appropriately handle bullying was generally moderate to low. The largest proportion of participants expressed a neutral stance (50.9%), indicating uncertainty about the effectiveness of existing AI systems in addressing bullying. Smaller but equal proportions of respondents reported being very confident (14.0%) and somewhat confident (14.0%) in these systems. In contrast, a notable portion of participants reported being not confident (19.3%), while only a very small percentage indicated that they were not at all confident (1.8%). Overall, these findings suggest that most respondents neither strongly trust nor fully distrust current AI systems in managing bullying appropriately, reflecting considerable uncertainty in this area.

From figure 7, respondents identified several key concerns regarding the use of artificial intelligence in detecting or responding to bullying. The most frequently reported concern was the violation of privacy (38.6%), indicating that many participants are worried about how personal data may be collected or used in AI-driven systems. This was followed by concerns about inaccurate judgment of tone or text (25.0%), suggesting apprehension that AI systems may misinterpret communication context and intent. A smaller proportion of respondents highlighted false detection or algorithmic bias (12.5%) as a major risk, reflecting concerns about fairness and reliability in AI decision-making. Overall, these findings suggest that privacy and accuracy are the primary issues influencing student perceptions of AI-based bullying detection systems.

In figure 8, respondents showed a strong overall agreement that AI systems should prioritize ethical safeguards when addressing bullying. The majority of participants either agreed or strongly agreed with this statement (89.5%), indicating broad support for the inclusion of ethical considerations in AI-based bullying detection and response systems. A small proportion of respondents remained neutral (8.8%), suggesting limited uncertainty or lack of strong opinion on the issue. Only a very small percentage strongly disagreed (1.8%), and no respondents selected disagreement. Overall, these results highlight a clear consensus among participants on the importance of ethical safeguards in the development and implementation of AI systems for bullying prevention.

In Figure 9, respondents identified several important features that an AI system should include to effectively prevent bullying. The most commonly supported feature was user reporting tools (32.2%), highlighting the importance of allowing individuals to actively report harmful behavior. This was closely followed by transparency in AI decisions (27.0%) and real-time content moderation (26.1%), indicating a strong preference for systems that are both responsive and understandable. Emotion or sentiment analysis (25.2%) was also considered important, suggesting interest in AI's ability to interpret tone and intent in communication. Educational feedback to users (18.3%) received comparatively lower support, though it still reflects recognition of the value of preventive and awareness-based approaches. Overall, the findings suggest that respondents prefer AI systems that combine responsiveness, transparency, and user involvement in addressing bullying.

From figure 10, respondents generally viewed human oversight as a critical component in AI systems that manage bullying or hate speech. The majority of participants rated it as very important (50.9%), indicating strong support for maintaining human involvement in AI-driven moderation processes. A smaller proportion considered it somewhat important (24.6%), while another group remained neutral (22.8%), suggesting some uncertainty about the extent of its necessity. Only a very small percentage of respondents believed that human oversight is not important (1.8%). Overall, these findings highlight a strong consensus that human supervision plays an essential role in ensuring the fairness, accuracy, and ethical use of AI in addressing harmful online behavior.

From figure 11, respondents overwhelmingly supported the idea that AI developers should be trained in anti-bullying ethics and social impact. The vast majority of participants indicated agreement with this requirement (84.2%), reflecting strong concern for the ethical responsibilities of developers working on AI systems that may be used to detect or prevent bullying. A smaller proportion of respondents were uncertain and selected "maybe" (14.0%), suggesting that some participants recognize the importance of ethical training but may question its implementation or scope. Only a very small percentage of respondents opposed the idea (1.8%). Overall, these findings demonstrate a strong consensus on the need for ethical and socially responsible training for AI developers in this context.

From figure 12, respondents suggested several ways in which AI education programs can effectively incorporate bullying prevention topics. The most commonly recommended approach was the use of case studies and simulations (37.8%), indicating a preference for practical, scenario-based learning experiences that illustrate real-world situations. This was followed by the inclusion of courses or modules on AI ethics (26.8%), highlighting the importance of formal instruction on ethical principles in AI development and use. Collaboration with psychology or education experts (24.4%) was also identified as an important strategy, reflecting the value of interdisciplinary input in understanding and addressing bullying behavior. Overall, these findings suggest that participants favor a combination of theoretical, practical, and interdisciplinary approaches to integrating bullying prevention into AI education programs.

Figure 13 shows that respondents support several technical strategies for addressing bullying with AI. The most favored approach was using natural language processing and sentiment analysis (78.9%). This was followed by updating models with verified, diverse data (64.9%), regularly auditing AI systems

(61.4%), and improving understanding of tone, sarcasm, and cultural context (54.4%). Other strategies included identifying repeat offenders through behavioral analysis (49.1%) and ensuring AI decisions are transparent and reviewable by humans (45.6%). Overall, respondents favor adaptive, accurate, and explainable AI systems for bullying detection and prevention.

From figure 14, respondents emphasized several ethical and governance strategies that should guide the development and use of AI systems for addressing bullying. The most frequently supported measure was the implementation of strong privacy safeguards (68.4%), highlighting concerns about protecting user data and maintaining confidentiality. This was followed by combining AI decisions with human moderators (59.6%) and involving diverse users and communities in AI design (57.9%), reflecting the importance of inclusivity and human oversight in system development. Requiring AI designers to follow established ethics codes (54.4%) and clearly explaining how AI detects and responds to bullying (54.4%) were also strongly supported, indicating a need for transparency and accountability in AI decision-making processes. Overall, these findings suggest that respondents prioritize privacy protection, ethical compliance, transparency, and human involvement as key components of responsible AI governance in bullying prevention systems.

Figure 15 shows that respondents support multiple educational and preventive strategies for addressing bullying in AI systems. The most supported approach was educating users about AI moderation, privacy, and respectful online behavior (70.2%). This was followed by enabling users to appeal or influence AI decisions (63.2%), promoting ethical AI use (61.4%), and educating engineers about the social impacts of AI (59.6%). Interdisciplinary collaboration between fields such as computer science, psychology, and education (52.6%) was also valued. Overall, the results highlight strong support for combining user education, ethical awareness, user empowerment, and cross-disciplinary design in AI-based bullying prevention.

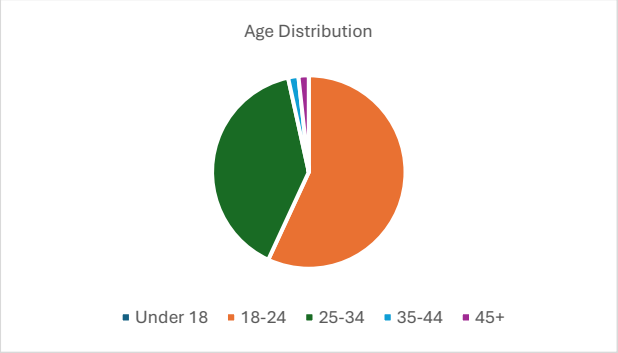

Fig 1. Age distribution

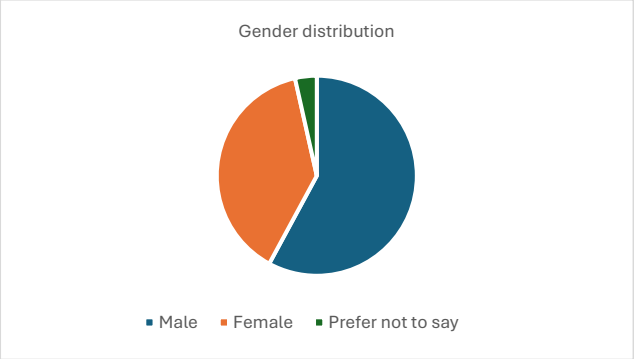

Fig 2. Gender distribution

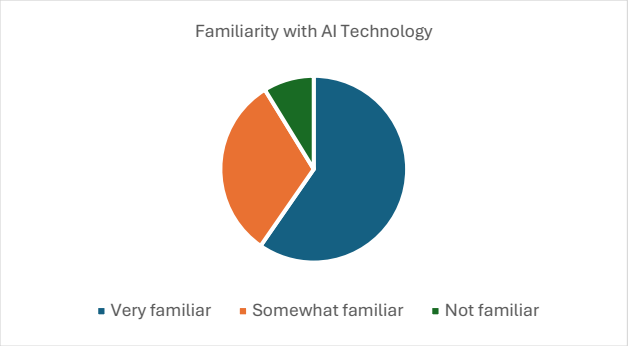

Fig 3. Familiarity with AI Technology

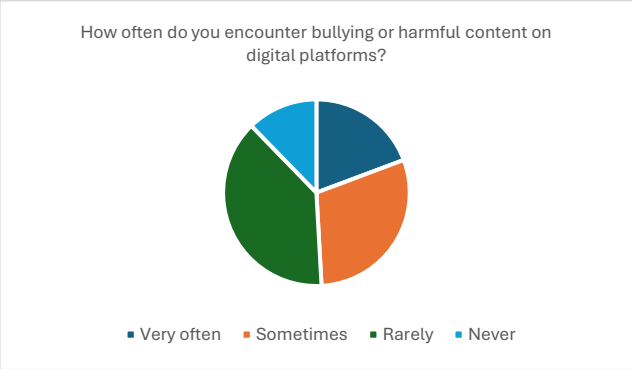

Fig 4. Data regarding 'How often do you encounter bullying or harmful content on digital platforms?'

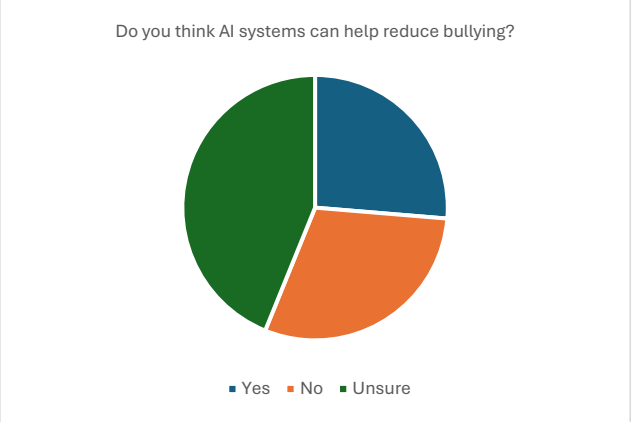

Fig 5. Data regarding 'Do you think AI systems can help reduce bullying?'

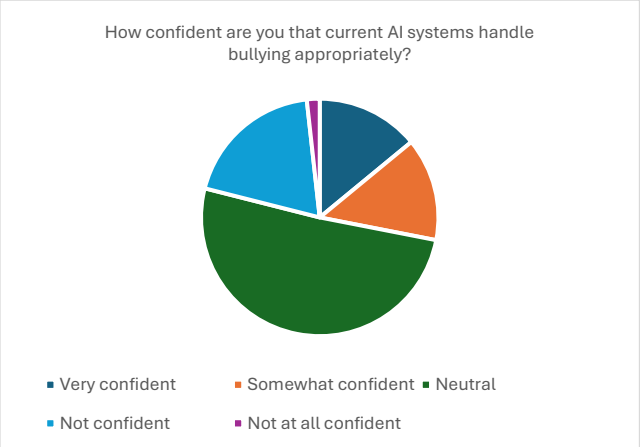

Fig 6. Data regarding 'How confident are you that current AI systems handle bullying appropriately?'

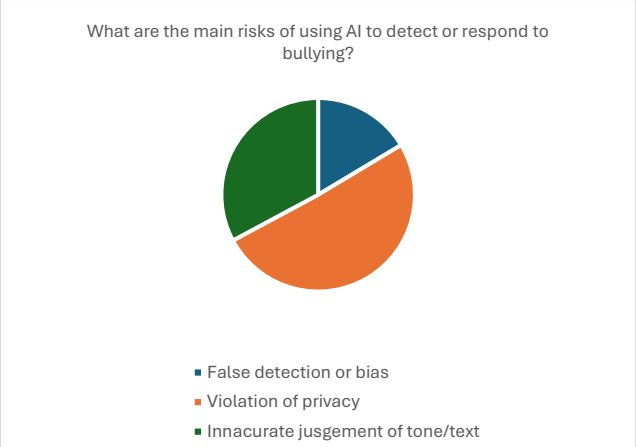

Fig 7. Data regarding 'What are the main risks of using AI to detect or respond to bullying?'

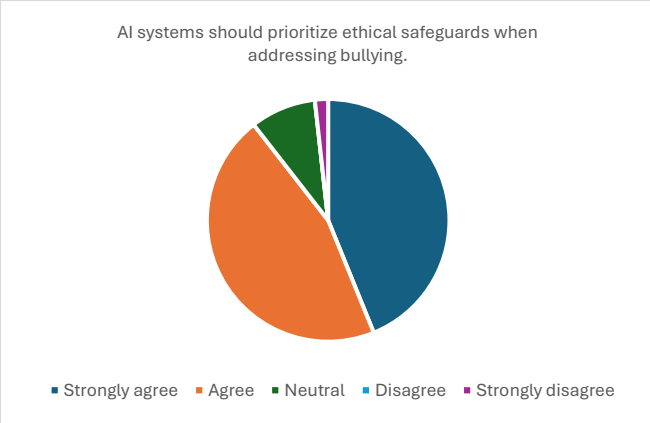

Fig 8. Data Regarding 'AI systems should prioritize ethical safeguards when addressing bullying.'

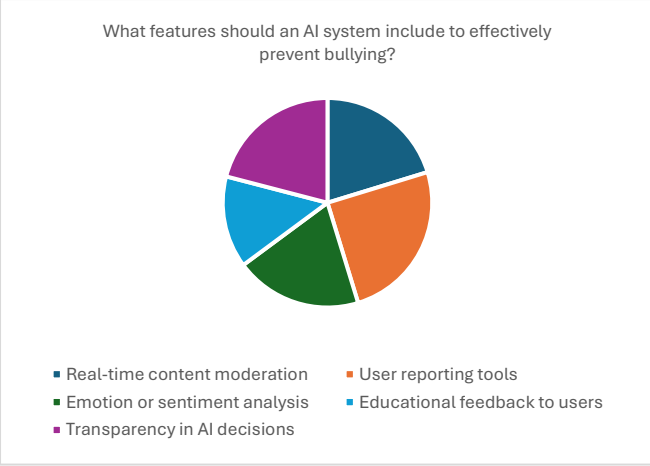

Fig 9. Data regarding 'What features should an AI system include to effectively prevent bullying?'

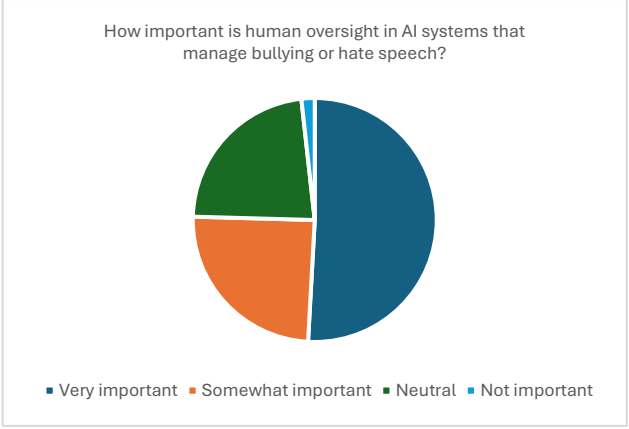

Fig 10. Data regarding 'How important is human oversight in AI systems that manage bullying or hate speech?'

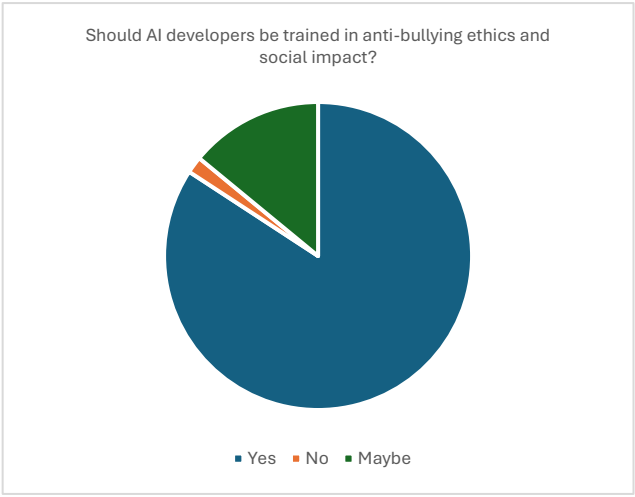

Fig 11. Data Regarding 'Should AI developers be trained in anti-bullying ethics and social impact?'

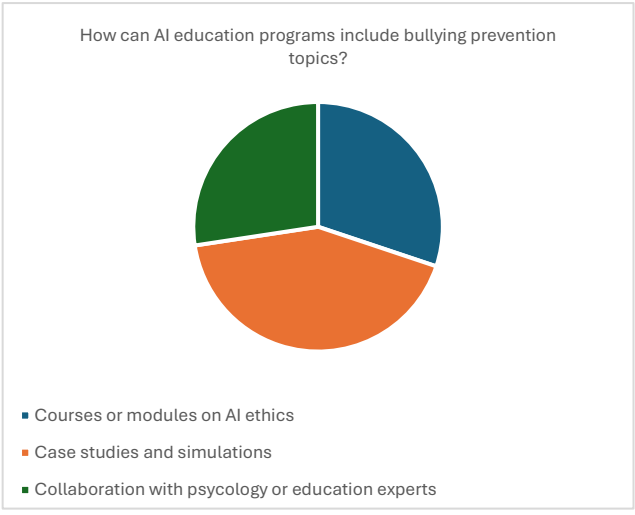

Fig 12. Data regarding 'How can AI education programs include bullying prevention topics?'

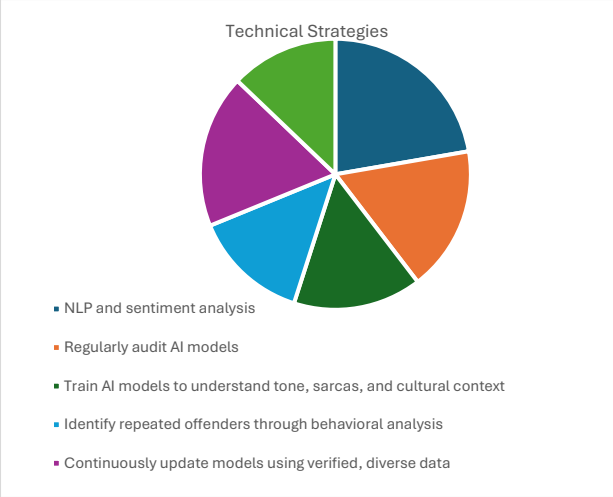

Fig 13. Data regarding choices about Technical Strategies

Fig 14. Data regarding choices about Ethical and Governance Strategies

Fig 15. Data regarding choices about Educational and Preventive Strategies

4. RESULTS AND FINDINGS

This section presents the findings of the study based on respondents' demographic characteristics, awareness of artificial intelligence (AI), and perceptions of AI-based systems for bullying detection and prevention.

4.1 Demographic Profile of Respondents

The age distribution of respondents indicates that the sample was primarily composed of young adults, with the majority falling within the 18–24 age range, followed by those aged 25–34. Only a small proportion of participants were in older age groups, while no respondents were under 18 years of age. This suggests that the study largely reflects the perspectives of traditional college-aged and early adult learners. The gender distribution shows that male participants formed the largest proportion of the sample, followed by female participants. A small number of respondents preferred not to disclose their gender. Overall, the sample included representation from multiple gender categories, although it was moderately male-dominated.

4.2 Familiarity with Artificial Intelligence

Findings indicate that respondents generally had a strong awareness of artificial intelligence technologies. Most participants reported being very familiar with AI, while a substantial proportion indicated moderate familiarity. Only a small minority reported limited or no familiarity. This suggests that respondents were sufficiently informed to provide meaningful insights into AI applications in bullying prevention.

4.3 Exposure to Online Bullying

Respondents reported varying levels of exposure to bullying or harmful content on digital platforms. Most participants indicated that they rarely or sometimes encountered such content. A smaller proportion reported frequent exposure, while a minority stated they had never encountered such experiences. These findings highlight that although exposure levels differ, online harm remains a recurring issue for many users.

4.4 Perceptions of AI in Bullying Prevention

Respondents expressed mixed views regarding the effectiveness of AI systems in reducing bullying. While a portion believed AI could help mitigate bullying, a similar proportion disagreed, and a large group remained uncertain. This reflects a lack of consensus and suggests that confidence in AI's preventive capabilities is still developing. Similarly, respondents demonstrated moderate to low confidence in current AI systems handling bullying appropriately. The majority adopted a neutral stance, while smaller proportions expressed confidence or lack of confidence. Overall, this indicates uncertainty regarding the reliability and effectiveness of existing AI moderation systems.

4.5 Perceived Risks and Ethical Concerns

Privacy violation emerged as the most significant concern regarding AI-based bullying detection systems. Other major concerns included inaccurate interpretation of tone or text and algorithmic bias or false detection. These findings suggest that respondents are primarily concerned with fairness, accuracy, and data protection when evaluating AI systems. Despite these concerns, there was strong agreement that AI systems should prioritize ethical safeguards. Respondents emphasized the importance of fairness, accountability, and responsible design, with minimal opposition to this view.

4.6 Desired Features in AI Systems

Respondents identified several key features necessary for effective AI-based bullying prevention systems. The most supported feature was user reporting tools, followed by transparency in decision-making and real-time content

moderation. Emotion and sentiment analysis were also valued, along with educational feedback mechanisms. Overall, participants favored systems that combine user involvement, interpretive capability, and transparency.

4.7 Human Oversight and Developer Responsibility

Findings strongly indicate that human oversight is considered essential in AI systems dealing with bullying and hate speech. Most respondents rated it as very important, with only a small minority expressing disagreement. In addition, respondents overwhelmingly supported the idea that AI developers should receive training in anti-bullying ethics and social impact. This reflects strong expectations for ethical responsibility in AI development.

4.8 Education and Training Approaches

Respondents suggested that AI education programs should integrate bullying prevention through practical and theoretical methods. Case studies and simulations were the most preferred approach, followed by formal AI ethics coursework and interdisciplinary collaboration with psychology and education experts. This highlights the importance of applied, real-world learning combined with cross-disciplinary perspectives.

4.9 Technical Strategies for AI Systems

Participants supported several technical improvements for AI-based bullying prevention. Natural language processing and sentiment analysis were the most emphasized strategies, followed by continuous model updates, regular auditing, and contextual understanding of tone and sarcasm. Behavioral analysis for identifying repeat offenders and ensuring explainable AI decisions were also considered important. Overall, respondents favored adaptive, transparent, and context-aware systems.

4.10 Ethical, Governance, and Educational Strategies

Respondents strongly supported governance measures such as privacy protection, human–AI collaboration, and ethical compliance. Transparency in AI decision-making and involvement of diverse user groups were also emphasized.

Educational and preventive strategies were equally important to respondents. These included educating users about AI moderation and digital behavior, allowing users to appeal AI decisions, promoting ethical digital citizenship, and integrating interdisciplinary collaboration in system design. Overall, respondents favored a balanced framework combining education, ethics, transparency, and inclusivity.

5. AVOIDING BULLYING IN AI SYSTEMS

Avoiding bullying in artificial intelligence (AI) systems involves preventing technologies from enabling, amplifying, or generating harmful, harassing, or discriminatory interactions. This requires a combined approach of technical safeguards, ethical design, and social responsibility throughout the AI lifecycle as listed next:

1. Defining Bullying in AI Contexts: Bullying in AI systems can take several forms, including user-to-user bullying where platforms fail to moderate harassment, AI-to-user bullying where systems produce harmful or biased outputs, and systemic bullying where algorithms reinforce inequality or unfair treatment of vulnerable groups.

2. Design Principles for Prevention: Ethical Dataset Curation: Training data should be filtered to remove toxic content, clearly labeled for harmful behavior, and designed to ensure demographic diversity to reduce bias.

- Bias Detection and Auditing: AI systems should undergo continuous bias checks, including audits, red-teaming, and fairness metrics such as toxicity and parity scoring.
- Safety and Filtering: Real-time toxicity detection, content filtering, and reinforcement learning from human feedback should be used to reduce harmful outputs.
- Empathy-Based Training: Models should be trained on respectful and supportive communication to encourage positive and constructive interactions.

3. Responsible Deployment Practices: User reporting systems should allow users to flag harmful behavior, with feedback used to improve models. Human-in-the-loop oversight is essential for handling ambiguous cases, while transparency through clear documentation, audit logs, and accountability systems ensures responsible AI behavior.

4. Education and Digital Literacy: Users and developers should be educated about AI limitations, bias, and ethical communication. Emphasis should be placed on empathy, responsible use, and awareness of AI risks.

5. Preventing AI-Generated Harm: AI systems that interact directly with users must include safeguards against insulting or manipulative outputs. They should prioritize de-escalation and use techniques such as sentiment analysis to provide calm and supportive responses.

6. Ethical Frameworks and Standards: AI anti-bullying systems should align with global ethical guidelines such as UNESCO AI Ethics (2021), IEEE Ethically Aligned Design, and OECD AI Principles, which emphasize fairness, transparency, accountability, and human well-being.

6. CONCLUSION

This study examined college students' perceptions of AI in bullying detection and prevention, focusing on awareness, trust, risks, and system design preferences. Findings show that while respondents are generally familiar with AI, their confidence in current systems remains moderate and uncertain, with mixed views on AI's effectiveness in reducing bullying. Key concerns included privacy risks, inaccurate interpretation of tone, and algorithmic bias. Despite this, participants strongly supported ethical safeguards, human oversight, transparency, and user-centered features such as reporting tools and real-time moderation. They also emphasized the importance of interdisciplinary collaboration and ethical training for developers. Overall, results suggest that effective AI-based bullying prevention requires a balanced approach combining technical capability with ethical responsibility and human involvement. While AI is seen as promising, its success depends on trust, accountability, and responsible design.

Future work should evaluate AI systems in real-world educational settings, improve detection of subtle and context-dependent bullying, and address algorithmic bias through diverse datasets and fairness mechanisms. Further research should also compare automated and human-in-the-loop approaches and examine long-term impacts on student behavior, well-being, and digital safety.

7. REFERENCES

[1] Perla, L., Forliano, R., Di Grassi, A., & Santacroce, M. T. (2024). AI-Enhanced education and beliefs: training course on bullying and cyberbullying prevention. In EDULEARN24 Proceedings (pp. 6320-6325). IATED.Gao,

S., & Ye, L. (2019, May). A physical and verbal bullying detecting algorithm based on K-NN for school bullying prevention. In International Conference on Artificial Intelligence for Communications and Networks (pp. 150-157). Springer International Publishing.

[2] Gao, S., & Ye, L. (2019, May). A physical and verbal bullying detecting algorithm based on K-NN for school bullying prevention. In International Conference on Artificial Intelligence for Communications and Networks (pp. 150-157). Springer International Publishing.

[3] Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., & O'Higgins Norman, J. (2023). Effectiveness of artificial intelligence–based cyberbullying interventions from youth perspective. Social Media+ Society, 9(1), 20563051221147325.

[4] Mendoza-Pinto, R. (2023). Artificial Intelligence in the Fight Against Bullying: Integration of ChatGPT in an Emotional Support Chatbot. In CEUR Workshop Proceedings (Vol. 1613, p. 0073).

[5] Asrifan, A. (2026). AI as a Guardian: Leveraging Machine Learning to Detect and Prevent Cyberbullying. In Integrating Parental Consent and Child Engagement With Digital Protection Rules (pp. 55-92). IGI Global Scientific Publishing.

[6] Chaudhary, A. (2024). AI-Based Online Harassment and Cyber Bullying Prevention System. In Impact of AI on Advancing Women's Safety (pp. 24-32). IGI Global Scientific Publishing.

[7] Chaudhary, P. K., Yalamati, S., Palakurti, N. R., Alam, N., Kolasani, S., & Whig, P. (2024, July). Detecting and preventing child cyberbullying using generative artificial intelligence. In 2024 Asia Pacific Conference on Innovation in Technology (APCIT) (pp. 1-5). IEEE.

[8] Singh, T. M., & Parkar, T. V. V. (2025). Developing Intelligent Systems to Combat Cyberbullying: AI-Driven Prevention Techniques. In Combating Cyberbullying With Generative AI (pp. 181-210). IGI Global Scientific Publishing.

[9] Orrù, G., Galli, A., Gattulli, V., Gravina, M., Micheletto, M., Marrone, S., ... & Sansone, C. (2023). Development of technologies for the detection of (cyber) bullying actions: The BullyBuster project. Information, 14(8), 430.

[10] Aggarwal, A., & Gaur, S. (2024). Applying artificial intelligence to explore online harassment and cyberbullying prevention. In Impact of AI on Advancing Women's Safety (pp. 104-120). IGI Global Scientific Publishing.

[11] Sinaga, J. A. br, Sitohang, S., Pardede, F. O. I., Sitorus, E., Sidabutar, R., Tobing, M. T., ... & Sinaga, D. Y. (2025). A socialization of the impact on artificial intelligence (AI) in overcoming bullying for children at school. Journal Of Human And Education (JAHE), 5(1), 172-177.

[12] Saputra, Y. M., Ali, E. Y., & Susilawati, D. (2025). UTILIZATION OF ARTIFICIAL INTELLIGENCE (AI) TO DETECT ELEMENTS OF BULLYING IN THE SCHOOL AND CAMPUS ENVIRONMENT. Jurnal Cakrawala Pendas, 11(1), 33-43.

[13] Sidhu, M. S., & Sidhu, K. K. (2025). AI strategies for handling disciplinary and cyber bullying in schools. In Current and future trends on ai applications: Volume 1 (pp. 207-229). Springer Nature Switzerland.

[14] González, C., Alberola Oltra, J. M., Sánchez Anguix, V., Monfort, G., Molines, S., Soto, M. D., & Díaz, S. (2020). A novel ICT tool to prevent bullying in schools.

[15] Angelis, J. D., & Perasso, G. (2020). Cyberbullying detection through machine learning: Can technology help to prevent internet bullying?. International Journal of Management and Humanities, 4(11), 57-69.