Technology in Society 84 (2026) 103089

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

AI can be cyberbullying perpetrators: Investigating individuals' perceptions and attitudes towards AI-generated cyberbullying

Weiping Pei [a], Fangzhou Wang [b,*], Yi Ting Chua [a]

[a] School of Cyber Studies, The University of Tulsa, United States
[b] Department of Criminology and Criminal Justice, University of Texas at Arlington, United States

ARTICLE INFO

Keywords:
Cyberbullying
Artificial intelligence (AI)
Large language models (LLMs)
AI-Generated content
Harmful content

ABSTRACT

Cyberbullying is a critical social problem that can cause significant psychological harm, particularly to vulnerable individuals. While Artificial Intelligence (AI) is increasingly leveraged to combat cyberbullying, its misuse to generate harmful content raises new concerns. This study examines human perception of AI-generated cyberbullying messages and their potential psychological impact. Using large language models (LLMs), we generated cyberbullying messages across three categories (sexism, racism, and abuse) and conducted a user study ($n = 363$), where participants engaged with hypothetical social media scenarios. Findings reveal that AI-generated messages can be just as or even more harmful than human-written ones in terms of participants' comfort levels, perceived harm, and severity. Additionally, AI-generated messages were almost indistinguishable from human-written ones, with many participants misidentifying AI-generated messages as human-written. Furthermore, participants with prior experience using AI tools consistently demonstrated higher accuracy in identification, while their attitudes towards online harm significantly influenced their comfort levels. This study emphasizes the urgent need for robust mitigation strategies to counter AI-generated harmful content, ensuring that AI technologies are deployed responsibly and do not exacerbate online harm.

1. Introduction

Cyberbullying is a critical social problem that can cause significant mental-health related issues, especially to adolescents (Campbell, 2012; Hinduja & Patchin, 2010, 2019; Monks et al., 2012). Cyberbullying involves deliberate and repeated harm via electronic devices (Patchin & Hinduja, 2015). It encompasses various forms including denigration (spreading harmful rumors), harassment (repetitive insults), and outing (sharing embarrassing information). Recent statistics highlight its growing prevalence, with 38 % of individuals encountering cyberbullying daily on social media (Nikola, 2023) and 26.5 % of students reporting incidents in the past 30 days, up from previous years (Patchin, 2024). To address this widespread problem and mitigate its devastating effects on society, researchers from social science and computer science have explored cyberbullying from different perspectives. Social science research has approached the study of cyberbullying through established socio-criminological theories to identify key contributing factors to victimization. One widely applied framework is Routine Activity Theory (RAT), which posits that cyberbullying occurs when three elements converge: a motivated offender, a suitable target, and the absence of a capable guardian to prevent (Cohen & Felson, 1979). RAT has been used to explain how individuals' online behaviors, such as frequent social media use, sharing personal information, or engaging with unknown users, may increase their exposure to potential offenders and reduce protective oversight (Aizenkot, 2022; Arntfield, 2015; Choi et al., 2019; Navarro & Jasinski, 2013). In parallel, the field of computer science has taken a technological approach, increasingly leveraging Artificial Intelligence (AI) to detect and prevent cyberbullying at scale. These AI-based systems apply machine learning, natural language processing, and sentiment analysis to identify harmful content, flag abusive messages, and intervene before further harm occurs (Ali Talpur & O'Sullivan, 2020; Balakrishnan et al., 2020; Bethany et al., 2023; Murnion et al., 2018). However, while AI has shown promise as a defense mechanism, its capabilities have also raised new concerns.

Increasingly, AI tools are being misused to generate harmful content rather than merely detect it. This is not merely hypothetical – past incidents, such as Microsoft's Tay chatbot, which was manipulated into spreading hate speech within hours of release, illustrate AI's

* Corresponding author. University of Texas at Arlington Department of Criminology and Criminal Justice, 601 S. Nedderman University Hall #362 Box 19595, Arlington, Texas, 76019, United States
E-mail address: fangzhou.wang@uta.edu (F. Wang).

https://doi.org/10.1016/j.techsoc.2025.103089
Received 31 March 2025; Received in revised form 20 June 2025; Accepted 25 September 2025
Available online 29 September 2025

vulnerability to exploitation (Twitter taught microsoft, 2016). The risks have intensified with the rapid advancement and increased accessibility of generative AI, as evidenced by the eSafety Commissioner's 2023 report highlighting the rise of AI-generated child sexual abuse material and deepfakes (First reports of children using, 2023). The emergence and widespread accessibility of generative AI, particularly large language models (LLMs), has enabled the scalable and automated generation of harmful content. The automation of content creation significantly lowers the cost of launching cyber attacks, including phishing attacks (Afane et al., 2024; Saha Roy et al., 2024) and coordinated harassment (Alexander, 2025). Moreover, AI-generated content can be hyper-realistic and emotionally manipulative, for instance, fabricated depictions of students in harmful scenarios have been linked to severe emotional trauma (Alexander, 2025). LLMs are also vulnerable to attacks such as prompt-jailbreaking, which can induce them to generate harmful outputs despite safety constraints (Lin et al., 2024b). Compounding these risks, the nuanced and fluent language produced by LLMs makes AI-generated cyberbullying content difficult to detect, often evading both human moderation and automated filtering systems. Collectively, these factors make AI-generated cyberbullying uniquely harmful by amplifying its scale, concealing its origins, and intensifying psychological impact compared to traditional human-written forms of cyberbullying.

As AI-generated content becomes increasingly prevalent in online interactions, existing cyberbullying theories fall short in capturing the unique dynamics introduced by AI-driven attacks. The Olweus Bullying Framework (Olweus, 1993), for example, defines bullying based on key elements such as power imbalance, intent to harm, repetition, and public exposure, typically within the context of human-to-human interactions. However, when LLMs are weaponized to generate harmful content, these dimensions take on new forms. The power imbalance shifts from interpersonal dominance to structural asymmetry, where individuals can exploit AI to launch large-scale, anonymized attacks. This imbalance is further amplified by the constant possibility of automated threats and the potential for widespread exposure (Dordolo, 2014). Intent becomes more diffuse and difficult to assess, as harmful output may originate from prompt manipulation, model vulnerabilities, or misuse, rather than a clearly identifiable aggressor. Similarly, the notion of repetition is redefined, as a single prompt can mass-produce harmful content that mimic persistent behavior. These changes also challenge the core assumptions of RAT, particularly the roles of motivated offenders and capable guardians. In AI-driven contexts, the concept of a motivated offender becomes ambiguous, as harmful content may be generated by automated systems without direct human intent or oversight. Meanwhile, traditional guardianship mechanisms, such as content moderation, reporting, or parental supervision, often fail to keep pace with the speed, scale, and contextual subtlety of AI-generated content. Together, all these divergences suggest that while traditional theories provide a solid foundation, there is a critical need for updated theoretical frameworks that account for the anonymity, automation, lack of human empathy, and systemic reach that characterize AI-driven cyberbullying.

While these theoretical shortcomings have been noted, empirical studies examining how people perceive AI-generated cyberbullying remain scarce. This study addresses that gap by examining how people perceive, evaluate, and distinguish AI-generated from human-written content. Specifically, we seek to answer four research questions: (1) How do individuals perceive AI-generated versus human-written cyberbullying content in terms of harm and severity (**RQ1**)? (2) To what extent can individuals differentiate between AI-generated and human-written cyberbullying messages (**RQ2**)? (3) What are individuals' attitudes toward AI-generated cyberbullying and potential mitigation strategies (**RQ3**)? (4) What factors influence individuals' ability to recognize AI-generated cyberbullying and their perception of its impact (**RQ4**)? To explore these questions, we conducted a survey experiment (n = 363) featuring AI-generated messages across three categories: sexism, racism, and abuse. Utilizing a combination of exposure-based and vignette-style methods, our study investigates participant responses to AI-generated versus human-written cyberbullying to support future detection and intervention efforts.

Overall, our work makes the following contributions: (1) We conduct the first study to systematically assess the risks posed by AI-generated cyberbullying, focusing on user perceptions and vulnerabilities. (2) We generate three types of AI-generated cyberbullying messages based on LLMs, contributing an AI-generated cyberbullying dataset to facilitate future research. (3) We design and conduct a user study to investigate individuals' perceptions and attitudes toward AI-generated cyberbullying, providing actionable insights for prevention strategies. (4) We highlight the necessity of public awareness regarding the risks of AI-generated content and the need for platform-level safeguards that can detect and mitigate AI-generated cyberbullying content.

2. Related work

2.1. Understanding of cyberbullying

Cyberbullying, as defined by Hinduja and Patchin (Patchin and Hinduja, 2015), involves the intentional and repeated use of digital technologies—such as computers, smartphones, and other electronic devices—to cause harm to others. This broad definition captures various harmful behaviors, including but not limited to denigration, flaming, harassment, and outing (Catherine et al., 2019). Studies on cyberbullying have traditionally focused on adolescents, highlighting perpetration and victimization rates ranging widely due to definitional and sampling inconsistencies (Gohal et al., 2023; Kwan et al., 2020; Menesini & Nocentini, 2009; Zhu et al., 2021). However, cyberbullying among adults is increasingly recognized, with estimates ranging from 10 % to 60 %, depending on country and population (Finn, 2004; Thazin Khine et al., 2020; Zhao et al., 2023). While psychological harms — such as depression, anxiety, and suicidal ideation — are consistent across age groups (Chu et al., 2022; Lee et al., 2023; Pabian & Vandebosch, 2021; Yeop Paek et al., 2022), adults often face cyberbullying in different settings. In workplace settings, adult cyberbullying is influenced by prior face-to-face bullying, health conditions, job tenure, and organizational cultures that discourage confrontation (Kim & Choi, 2021; Zhang et al., 2022).

Key predictors of cyberbullying perpetration and victimization vary across the life course, though some are shared across adolescence and adulthood. Among adolescents, common victimization risks include frequent technology use, prior abuse, impulsivity, excessive online activity, poor relationships, and minority status (Alhaboby et al., 2016; Finn, 2004; Mishna et al., 2012), with females and marginalized groups disproportionately affected (Bauman & Newman, 2013; Lindsay et al., 2016; Wong et al., 2018). Research on age shows mixed results: some report no clear effect (Didden et al., 2009; Patchin & Hinduja, 2006; Smith et al., 2008), though Smith et al. found age-gender differences in perceived harm (Smith et al., 2008); others report increased victimization during middle school (Kowalski & Limber, 2007; Williams & Guerra, 2007; Ybarra et al., 2006). For teen perpetrators, key predictors include gender, low empathy, school dissatisfaction, and prior offline victimization (Ang & Goh, 2010; Cross & Walker, 2012; Hinduja & Patchin, 2008). Educational interventions have been shown to impact both perpetration and victimization among adolescents (L'opez-Castro & Priegue, 2019; Debby Ng et al., 2022; Patterson et al., 2019). In young adults, cyberbullying is shaped by sociodemographic factors (age, gender, ethnicity), perceived anonymity, prior traditional bullying, problematic social media use, and broader psychosocial factors such as job stress, weak digital boundaries, poor family management, or low emotional control (Barlett et al., 2017; Chapell et al., 2006; Hemphill & Heerde, 2014; Kwan & Skoric, 2013; Kırcaburun et al., 2019; Wang et al., 2019b; Yubero et al., 2017).

Importantly, the roles of victim, perpetrator, and bystander are not static. While this traditional framework in cyberbullying research offers clarity, it has been critiqued for overlooking the fluidity of roles in online aggression (Boccio & Leal, 2023; Jian et al., 2025). Increasingly, evidence suggests that individuals often shift between roles depending on context and time. For example, victims of traditional bullying may engage in online retaliation under the veil of anonymity (Slonje & Smith, 2008), and many individuals simultaneously occupy both perpetrator and victim roles, commonly referred to as "bully-victims" (Gradinger et al., 2009; Lazuras et al., 2017; Wang et al., 2019a). In addition, perpetrator-victims are more likely to engage in negative bystander behaviors, a tendency influenced by lower levels of empathy and reduced perceptions of incident severity, as demonstrated in a simulation study by Zhao and colleagues (Zhao et al., 2023).

This overlap is influenced by factors such as risky lifestyles, deviant peer networks, low self-control, and weak social bonds, with social learning mechanisms mediating the effects of these traits on cyberbullying behaviors (Xu & Tu, 2024). Gender differences are also notable: males are more often perpetrators (Li, 2006), while females are more frequently victimized, though the theoretical basis for these gendered patterns remains insufficiently examined (Lin et al., 2024a). Furthermore, research show that individuals may transition from victims to passive bystanders or vice versa, depending on situational and psychological factors (Jian et al., 2025; Boccio & Leal, 2023; Weulen Kranenbarg et al., 2019; Balakrishnan & Fernandez, 2018; Matamoros-Fernández & Farkas, 2021; Macaulay et al., 2022).

Building on this evolving understanding of cyberbullying roles and mechanisms, recent technological developments have introduced a novel and understudied phenomenon: AI-generated cyberbullying. While prior work has largely utilized AI for the detection and mitigation of human-authored abuse (Ayofe Azeez et al., 2021; Kumar Chaudhary et al., 2024; Milosevic et al., 2023), the rapid advancement of generative AI such as LLMs and conversational agents raises the possibility that AI itself may become a source of harmful content. AI-generated messages, produced at scale with human-like fluency, pose unique challenges by potentially amplifying psychological harm while escaping traditional frameworks for detection, moral responsibility, and intervention. Given the current research gap, our study investigates how individuals perceive and respond to cyberbullying content authored by AI versus humans, focusing on emotional reactions, perceived harm, and likelihood of reporting or intervening. This work aims to extend the conceptual boundaries of cyberbullying by addressing the implications of non-human perpetrators within digital aggression frameworks.

2.2. AI-generated vs. human-written content

This rising concern is further compounded by the increasing difficulty in distinguishing AI-generated from human-created content. As generative AI systems produce outputs with near-human fluency, both researchers and the public face significant challenges in accurately identifying the origins of digital messages. Most existing research has focused on technological approaches for AI-generated content detection, such as the OUTFOX framework, which differentiates between LLM-generated and human-written essays (Koike et al., 2024). At the same time, some studies have explored the human ability to discern AI-generated content. Frank et al. found that most participants struggled to accurately identify AI-generated media content, often relying on mere guesses (Frank et al., 2024) while Lu et al. demonstrated similar difficulties in distinguishing between real photos and those generated by AI (Lu et al., 2024).

With textual content, Köbis and Nossink found that participants performed no better than chance when GPT-2's poems of the highest qualities were selected and presented alongside human-written poems (Köbis & Mossink, 2021). However, when random sample of GPT-2's poems were presented, participants' detection rates improved above chance. In their large-scale research with 4600 participants, Jakesch and colleagues also found that participants generally were unable to tell the difference between personal profiles written by humans or by GPT-3 (Jakesch et al., 2023). In addition, the authors identified several intuitive but flawed heuristics people used when judging text authenticity. Some of these heuristics include first-person pronouns, emotional language, and grammatical issues. If participants had only relied on functional cues such as nonsensical and repetitive text, detection accuracy was greatly improved. Another emerging area of research is understanding individuals' perceptions and attitudes towards AI-generated content. Graefe and colleagues showed that participants rated news articles declared as human-written more favorably, regardless of the actual authorship. However, participants rated AI-generated articles as more credible and higher in expertise (Graefe et al., 2018). Lim and Schmälzle found notable differences in how individuals perceive AI-generated vs. human-written messages in the context of vaping prevention messaging (Lim & Schmälzle, 2024). Similarly, Brigham et al. showed that AI-generated non-consensual intimate imagery (AIG-NCII) is widely perceived as harmful and unethical, with harm assessments varying based on whether the target was a public figure or an individual (Grace Brigham et al., 2024).

The current study addresses both areas of research on AI-generated and human-written content. First, this study examines individuals' ability to recognize AI-generated content but also investigates the criteria and factors influencing their judgments. Second, our study examines individuals' attitudes towards AI-generated cyberbullying, a critical area that has not been thoroughly studied. This is especially important in the context of cyberbullying. It has been shown that participants failed to detect sophisticated bots, with accuracy rate dropped to near 50 % (Kolomeets et al., 2024). This means that malicious actors can employ AI bots to engage in cyberbullying without detection and amplifying abuse via automation.

3. Design of the study

3.1. Overview of the study

To explore the four research questions, we conducted a survey study to examine participants' perceptions towards AI-generated cyberbullying. Our study design follows established methodologies in online harm research. We considered a similar exposure approach, commonly used in cyberbullying dataset construction, in which participants engage with cyberbullying messages for annotation (Samory et al., 2021; Vishwamitra et al., 2021). Additionally, we incorporated a hypothetical scenario design, drawing from prior studies employing vignette methodologies to examine human perceptions in online harm contexts (Chan et al., 2013; Ireland et al., 2020). The effectiveness of these approaches demonstrated that such exposure in a hypothetical scenario, when ethically managed, contributes to improved detection frameworks, policy recommendations, and platform interventions aimed at reducing harm.

Fig. 1 provides an overview of our study, which consists of two stages:

a) AI-generated cyberbullying message creation and b) a user study. In the first stage, we used LLMs to generate cyberbullying messages representing three categories: sexism, racism, and abuse (detailed in Section 3.2). Rather than generating messages completely entirely from scratch, we provided real-world cyberbullying messages as examples in the prompts for two reasons. First, providing concrete examples helps enhance the quality and contextual relevant of AI-generated content by guiding the model toward more realistic outputs. Second, using human-written messages as references allows for the generation of semantically similar AI-generated messages, enabling a more meaningful comparison between human- and AI-generated content in the subsequent user study.

In the second stage, we designed and conducted a user study in which participants were shown two messages, one human-written and one AI-generated. They were instructed to assume they had encountered the messages on social media platforms such as Twitter or Facebook. They

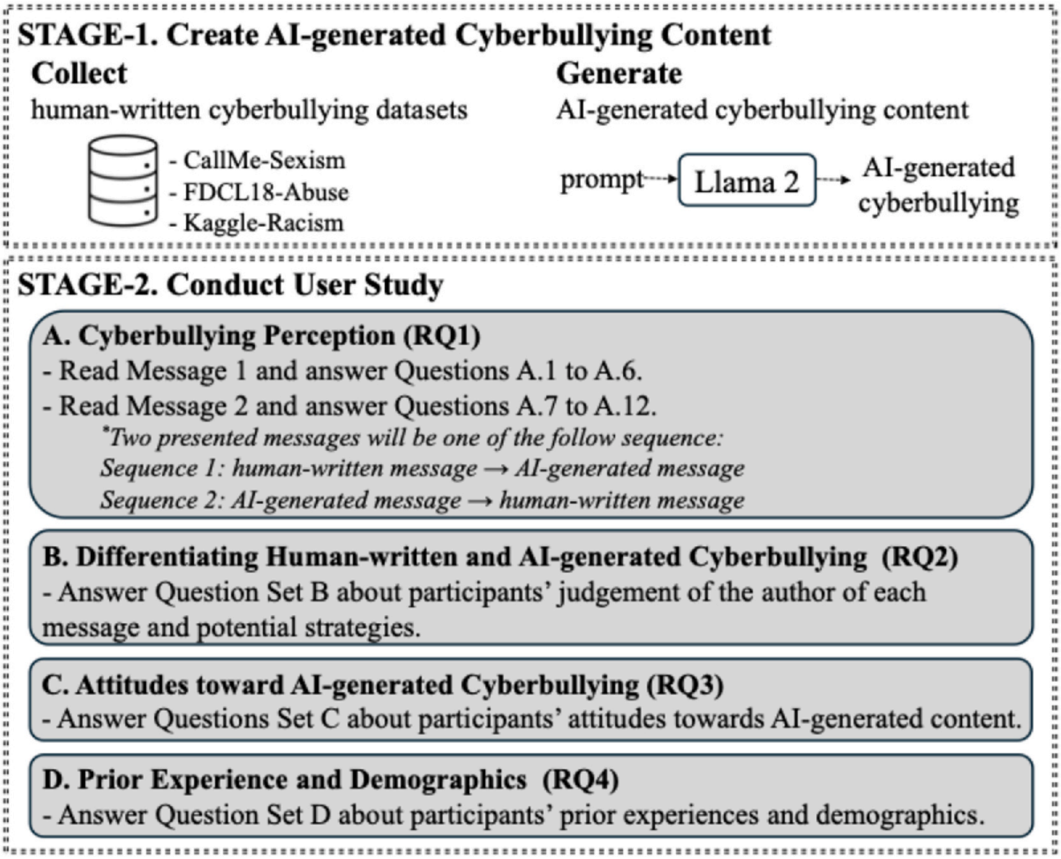

STAGE-1. Create AI-generated Cyberbullying Content
Collect
human-written cyberbullying datasets
- CallMe-Sexism
- FDCL18-Abuse
- Kaggle-Racism
Generate
AI-generated cyberbullying content
prompt
Llama 2
AI-generated cyberbullying
STAGE-2. Conduct User Study
A. Cyberbullying Perception (RQ1)
- Read Message 1 and answer Questions A.1 to A.6.
- Read Message 2 and answer Questions A.7 to A.12.
*Two presented messages will be one of the follow sequence:
Sequence 1: human-written message → AI-generated message
Sequence 2: AI-generated message → human-written message
B. Differentiating Human-written and AI-generated Cyberbullying (RQ2)
- Answer Question Set B about participants' judgement of the author of each message and potential strategies.
C. Attitudes toward AI-generated Cyberbullying (RQ3)
- Answer Questions Set C about participants' attitudes towards AI-generated content.
D. Prior Experience and Demographics (RQ4)
- Answer Question Set D about participants' prior experiences and demographics.

Fig. 1. An overview of our study.

then answered a series questions designed to address our RQs. The questions are organized into four sets, consisting of 32 open-ended and close-ended questions (as shown in Appendix A). In **Question Set A (RQ1)**, participants answered a series of questions about their feelings towards two cyberbullying messages: one that was human-written and another that was AI-generated, without being informed of the authorship (*A.1-A.12*). **Question Set B (RQ2)** involved presenting both the human-written and AI-generated cyberbullying messages, where participants answered questions regarding their judgments about the authors of each message (*B.1-B.6*). **Question Set C (RQ3)** focused on participants' attitudes towards AI-generated cyberbullying, including their comfort levels with known AI-generated cyberbullying messages and their opinions on detecting AI-generated cyberbullying (*C.1-C.4*). Finally, **Question Set D (RQ4)** gathered data on participants' prior experiences (*D.1-D.6*) and demographics (*D.7-D.10*). An attention check question was also included for quality control.

To minimize potential effects of the order in which messages were presented, we considered two presentation sequences: (1) human-written message first, followed by the AI-generated message, and (2) AI-generated message first, followed by the human-written message.

3.2. Cyberbullying datasets

Collect human-written cyberbullying content. We selected human-written cyberbullying datasets based on the following criteria: (1) English-written content, (2) publicly available labels, (3) specific focus on cyberbullying, and (4) content created by real humans. To ensure conceptual and topical diversity, we focus on sexism (Jha & Mamidi, 2017), racism (Matamoros-Fernández & Farkas, 2021), and abuse (Salawu et al., 2021) as representative types of cyberbullying. These categories are widely recognized in the literature as among the most pervasive and harmful forms of online aggression (Nobata et al., 2016; Waseem & Hovy, 2016), and they capture both structural and interpersonal dimensions of online harm (Megarry, 2014). Moreover, they align with sociological and psychological theories related to marginalization, identity-based harassment, and power asymmetries, which are central to understanding cyberbullying as a complex social phenomenon. Based on these considerations, we selected three datasets corresponding to these three types of cyberbullying.

- CallMe-Sexism (Samory et al., 2021): This dataset contains 13k tweets collected from multiple studies (Jha & Mamidi, 2017; Samory et al., 2021; Waseem & Hovy, 2016). Each tweet was annotated in 2021 via MTurk, categorized into one of four sexist content categories using psychological scales, or labeled as non-sexist. We randomly selected 20 sexist tweets from each category to ensure diverse representation of sexist content.
- Kaggle-Racism (Cyberbullying dataset, 2020): This dataset includes 159k messages from various social media platforms such as Kaggle and Twitter, encompassing different types of cyberbullying. We focused on racist tweets and randomly sampled 100 tweets for our study.
- FDCL18-Abuse (Founta et al., 2018): This dataset consists of over 100k tweets annotated for abusive behaviors. The dataset was created on CrowdFlower platform in 2017 and merged abusive, offensive and aggressive annotations into a single category. We randomly selected 100 abusive tweets.

All cyberbullying messages sampled from the above datasets served two purposes: (1) they were presented as human-written cyberbullying messages in the user study, and (2) they were used as input prompts for LLMs to generate cyberbullying content.

Create AI-generated cyberbullying content. We utilized Meta's pre-trained Llama 2 model (Touvron et al., 2023) to create AI-generated cyberbullying content. Specifically, we prompted the model to produce 10 cyberbullying messages similar to a given human-written cyberbullying message, following the prompt template shown in Fig. 2. Using human-written messages sampled from the three datasets, we generated three types of AI-generated cyberbullying messages: sexism, racism, and abuse. Accordingly, we conducted a user study with three independent groups, each focusing on one type of cyberbullying: the sexism group, the racism group, and the abuse group.

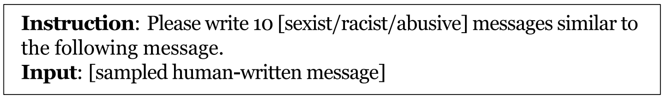
3.3. Data analysis methodology

To comprehensively examine both the measurable and interpretive dimensions of participants' responses, this study employed a mixed-methods design. In this framework, quantitative and qualitative data were collected simultaneously and analyzed independently, with findings integrated during the interpretation phase. This design was selected to provide both statistical validation of behavioral patterns and a rich contextual understanding of participant reasoning, essential for addressing our research questions regarding AI-generated cyberbullying.

Quantitative analysis of close-ended responses. Quantitative data consisted of responses to close-ended Likert-scale items assessing participant comfort levels, perceived harm and severity of cyberbullying messages, and judgments about message authorship. We first calculated descriptive statistics to summarize response distributions across the three message categories. To assess within-subject differences, particularly in comfort level and harm perception between human-written and AI-generated messages, we used the Wilcoxon signed rank test, a nonparametric alternative to the paired *t*-test suitable for ordinal data. All p-values were reported following APA style guidelines (Number and Statistics Guide, 2025). To ensure the adequacy of statistical inference, we conducted a power analysis (Akter et al., 2020; Li et al., 2017), which determined that a minimum of 24 participants per group was necessary to achieve 80 % power with a medium-to-large effect size (Cohen's d = 0.8) and a 95 % confidence level. Our sample exceeded these requirements: 159 participants in the sexism group, 99 in the racism group, and 105 in the abuse group, thereby ensuring sufficient statistical power across analyses.

Qualitative analysis of open-ended responses. Qualitative data were derived from participants' narrative explanations to open-ended questions, including justifications for their authorship judgments and reflections on AI-generated content. We analyzed these responses using inductive thematic analysis (Braun & Clarke, 2006), guided by best practices for open coding, iterative refinement, and theme development (Braun & Clarke, 2022; Nowell et al., 2017). The thematic coding was performed by the first three authors. Each cyberbullying category was assigned to two coders: the sexism group was coded by the second and third authors, the racism group by the first and second authors, and the abuse group by the first and third authors. The coding process involved multiple stages.

- **Initial Independent Coding:** For each question, two coders independently reviewed and coded a stratified subset of participant responses (typically 20–30 %) to identify emergent codes.
- **Codebook Construction:** Coders then collaborated to synthesize their findings into a preliminary codebook. This included agreed-upon code definitions, exemplar quotes, and inclusion/exclusion criteria.



Instruction: Please write 10 [sexist/racist/abusive] messages similar to the following message.
Input: [sampled human-written message]

Fig. 2. Prompt template used for creating AI-generated cyberbullying messages.

- **Full Dataset Coding:** Using the preliminary codebook, coders independently analyzed the remaining responses within their assigned groups. New or ambiguous responses that could not be categorized using the initial codebook led to iterative codebook updates.
- **Final Reconciliation:** After all responses were coded, the coders met to discuss discrepancies, merge overlapping codes, and refine thematic categories. The final codebook, including detailed definitions and examples, is available on Zenodo.[1]

To assess inter-coder reliability, we computed Cohen's Kappa (Fleiss et al., 2013; Jacob, 1960) for each open-ended question. The average Kappa coefficient across all questions was 0.87 (SD = 0.06), indicating a high level of agreement and consistency in thematic interpretation across coders (see Appendix B).

3.4. Participants and ethical consideration

We recruited participants through MTurk (Amazon mechanical turk, 2024), a widely used crowdsourcing platform, restricting participants to U.S. adults due to the compensation requirement.

Pilot study. Before the formal user study, we conducted a pilot study with 20 participants to evaluate the clarity of the survey questions, the correctness of the procedures, and the time taken to complete the survey. While no major issues were identified, participants took an average of 8.7 min to complete the survey, which exceeded the expected time of 5 min. As a result, we increased the compensation to $2.00 USD for the formal study. Only minor wording refinements were made to finalize the survey. Participants from the pilot study were excluded from the formal study, and their responses were not included in the formal result analysis.

Participant recruitment and compensation. The formal study was conducted on MTurk in May and June 2024. It is important to note that we intentionally did not set any quality-related qualifications, such as crowd workers' approval rate, during pre-selection process. This decision was made to recruit a diverse group of participants regardless of their performance on MTurk. The formal survey took approximately 10 min to complete, and we compensated each qualified participant with $2.00 USD. This results in a projected hourly wage of $12.00 USD, which exceeds the required minimum wage rate of $7.25 USD per hour in the United States (Hara et al., 2018).

Ethical considerations and content warning. Given the sensitive nature of the distributed survey, we took several steps to mitigate ethical risks. First, we submitted our application to the Institutional Review Boards (IRBs) at each author's institution and received approval from all before starting our study. We meticulously adhered to IRB guidelines to minimize any risks to potential participants throughout the study. We ensured that no identifiable information was collected and used an informed consent process that did not require signed consent. Second, we provided potential participants with detailed information about the risks involved, particularly regarding the potential exposure to harmful content. We followed MTurk guidelines to include a clear warning, "WARNING: This HIT may contain negative content," in the task title. In the informed consent form, we highlighted the potential risks and emphasized the voluntary nature of participation as well as the right to withdraw at any time without penalty. Third, we conducted participant screening to exclude individuals under 18 years old, recognizing that minors are particularly vulnerable to cyberbullying content (Hinduja & Patchin, 2010; Kamar et al., 2022; Rebecca, 2015; Yeop Paek et al., 2022). Fourth, we offered debriefing and assess to mental health support by providing participants with links to resources, including: 1) StopBullying.gov (StopBullying.gov. Get help now, 2022), 2) Cyberbullying Research. Center (Cyberbullying Research Center, 2024), 3) National

[1] The final codebook can be found at https://zenodo.org/records/15107233.

Domestic Violence Hotline (National Domestic Violence Hotline), and 4) RAINN's resources for survivors of Stalking and Cyberstalking (RAINN, 2024). Finally, following the widely used methodology in cyberbullying studies (Chan et al., 2023; Conway et al., 2016; Perren & Gutzwiller-Helfenfinger, 2012), we employed a vignette methodology to explore participants' perceptions through hypothetical scenarios of cyberbullying. Additionally, participants were invited to provide feedback (such as additional comments or concerns) at the end of the study, but we did not receive any comments or complaints regarding harmful content.

4. Results analysis

In this section, we present participant demographics and analyze findings to address our four RQs. As mentioned in Section 3.1, participants were assigned one of two message presentation sequences. However, a Chi-square test of independence revealed no statistically significant relationship between message order and participants' comfort level, perceived harm, or severity (Question Set A). Therefore, this factor was excluded from further analysis.

Demographics. A total of 794 participants completed our survey study.[2] After filtering out participants who failed the attention check question or provided completely irrelevant answers to open-ended questions,[3] we were left with 363 valid participants. These participants were distributed across three groups: 159 in the sexism group, 99 in the racism group, and 105 in the abusive group. Table 1 summarized the demographics of the participants in each group. In brief, the majority of participants were between 25 and 39 years old, comprising 76.73 %, 83.84 %, and 80.00 % of the sexism, racism, and abuse groups, respectively. As for the gender, the majority of participants in each group were male, with 74.84 % in the sexism group, 63.64 % in the racism group, and 66.67 % in the abuse group. As for education level, most participants in each group held a bachelor degree, with 73.58 % in the sexism group, 71.72 % in the racism group, and 69.52 % in the abuse group. Regarding race, most participants identified as white, with 89.94 % in the sexism group, 94.95 % in the racism group, and 94.29 % in the abuse group. We noticed demographic biases and discussed their potential impact in Section 5.

Table 1
Demographics of participants.

	sexism (159)	racism (99)	abuse (105)
Age			
18–24	3 (1.89 %)	2 (2.02 %)	3 (2.86 %)
25–39	122 (76.73 %)	83 (83.84 %)	84 (80.00 %)
40–60	33 (20.75 %)	13 (13.13 %)	15 (14.29 %)
60 or older	1 (0.63 %)	1 (1.01 %)	3 (2.86 %)
Gender			
Male	119 (74.84 %)	63 (63.64 %)	70 (66.67 %)
Female	37 (23.27 %)	34 (34.34 %)	34 (32.38 %)
Prefer not to say	3 (1.89 %)	2 (2.02 %)	1 (0.95 %)
Education			
Highschool	6 (3.77 %)	1 (1.01 %)	2 (1.90 %)
College	4 (2.52 %)	3 (3.03 %)	0 (0.00 %)
Bachelor	117 (73.58 %)	71 (71.72 %)	73 (69.52 %)
Master	31 (19.50 %)	24 (24.24 %)	30 (28.57 %)
Doctoral	1 (0.63 %)	0 (0.00 %)	0 (0.00 %)
Race			
White	143 (89.94 %)	94 (94.95 %)	99 (94.29 %)
Asian	7 (4.40 %)	2 (2.02 %)	2 (1.90 %)
Black	4 (2.52 %)	1 (1.01 %)	2 (1.90 %)
American Indian	3 (1.89 %)	1 (1.01 %)	0 (0.00 %)
Other	2 (1.26 %)	1 (1.01 %)	2 (1.90 %)

Prior experience. We explored participants' prior experiences with cyberbullying and AI usage, as summarized in Table 2. More than half of the participants in each group reported having experienced cyberbullying: 62.26 % in the sexism group, 50.51 % in the racism group, and 59.05 % in the abuse group. In terms of cyberbullying incidents over the past 12 months, approximately 25 %–30 % of participants reported being targeted at least once across three groups. Additionally, 21.38 %, 19.19 %, and 27.62 % of participants in the respective groups experienced cyberbullying more than once during the same period. Given that most participants were over the age of 24, these findings suggest that cyberbullying remains prevalent even among adults. Regarding AI usage, the majority of participants reported prior experience with AI tools, and over 70 % indicated that they use such tools on a daily or weekly basis. These findings underscore the growing integration of AI technologies into everyday life.

4.1. *RQ1: Perception of Cyberbullying Content*

To answer RQ1, we presented one human-written and one AI-generated cyberbullying message to participants without disclosing the authorship of messages. Participants were asked to indicate their comfort level, perceptions of harm and severity, and their reactions to both messages.

4.1.1. *Comfort level*

Participants rated their comfort level on a 5-point Likert scale (A.1/A.7) for each message. Fig. 3a displays the distributions of participants' comfort levels when receiving a human-written message compared to an AI-generated message. In the sexism (n = 159) and abuse (n = 105) groups, participants reported a higher level of discomfort with AI-generated content compared to the human-written message. Specifically, in the abuse group, discomfort rose from 60.95 % for human-written messages to 76.19 % for AI-generated messages, with a statistically significant difference based on the Wilcoxon signed rank test. However, in the racism (n = 99) group, there were no differences with the reported comfort level.

4.1.2. *Perceived harm*

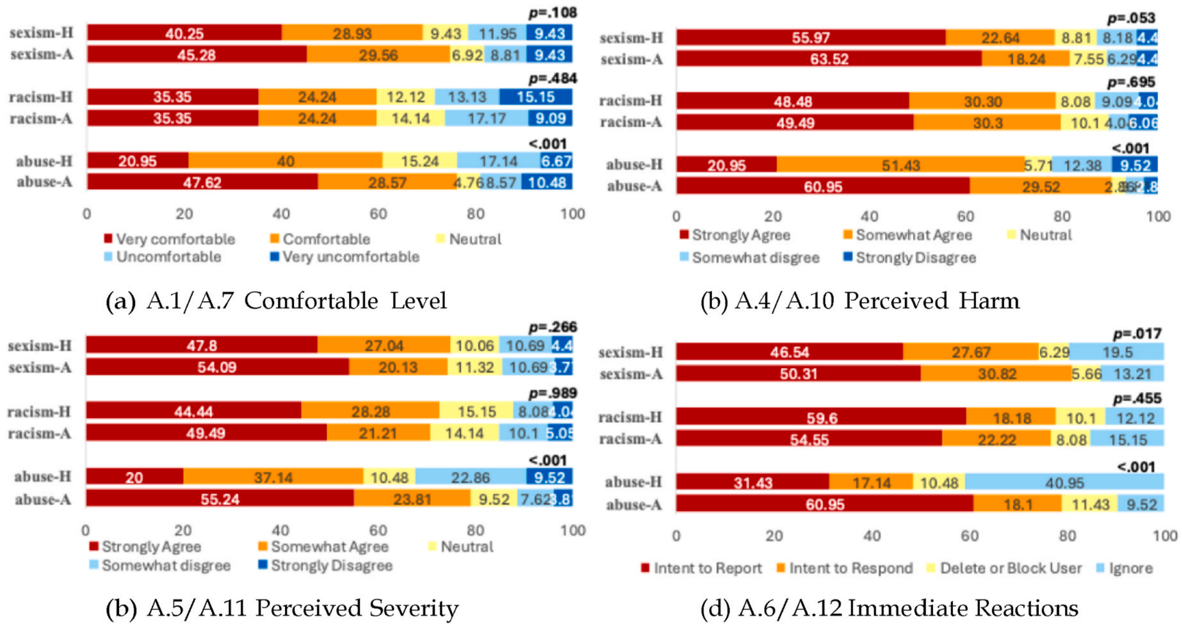
Participants were asked to rate their agreement with the statement, "the above message is harmful," for both human-written and an AI-generated messages (A.4/A.10). Fig. 3b presents the distribution of

Table 2
Participants' prior experiences with cyberbullying and AI usage.

	sexism (159)	racism (99)	abuse (105)
Prior Experience with Being Cyberbullied			
Yes	99 (62.26 %)	50 (50.51 %)	62 (59.05 %)
No	57 (35.85 %)	43 (43.43 %)	39 (37.14 %)
Prefer Not to Say	3 (1.89 %)	6 (6.06 %)	4 (3.81 %)
Frequency of Cyberbullying in Past 12 Months			
None	79 (49.69 %)	45 (45.45 %)	42 (40.00 %)
One	41 (25.79 %)	29 (29.29 %)	31 (29.52 %)
More than one	34 (21.38 %)	19 (19.19 %)	29 (27.62 %)
Prefer Not to Say	5 (3.14 %)	6 (6.06 %)	3 (2.86 %)
Prior Experience with Using AI Products			
Yes	148 (93.08 %)	82 (82.83 %)	99 (94.29 %)
No	6 (3.77 %)	10 (10.10 %)	4 (3.81 %)
Prefer Not to Say	5 (3.14 %)	7 (7.07 %)	2 (1.90 %)
Frequency of Using AI Products for Text Generation			
Never	4 (2.52 %)	3 (3.03 %)	2 (1.90 %)
Daily	62 (38.99 %)	37 (37.37 %)	36 (34.29 %)
Weekly	65 (40.88 %)	38 (38.38 %)	46 (43.81 %)
Monthly/Yearly/Other	28 (17.61 %)	21 (21.21 %)	21 (20.00 %)

[2] Incomplete and duplicate submissions are excluded from analysis.

[3] All authors independently reviewed each participant's responses and held discussions to reach final decisions on excluding low-quality responses.



(a) A.1/A.7 Comfortable Level

(b) A.4/A.10 Perceived Harm

(b) A.5/A.11 Perceived Severity

(d) A.6/A.12 Immediate Reactions

Fig. 3. Participants' Perception of Cyberbullying Content (RQ1). For each group, the upper bar (i.e., sexism-H, racism-H, and abuse-H) is for human-written messages, and the lower bar (i.e., sexism-A, racism-A, and abuse-A) is for AI-generated messages.

participants' agreement levels. Similar to reported comfort level, participants in the sexism (n = 159) and abuse (n = 105) groups somewhat or strongly agreed that the AI-generated message was more harmful than human-written message, with the difference in the abuse group reaching statistical significance. In the racism (n = 99) group, comparable percentages of participants rated both messages as harmful.

4.1.3. *Perceived severity*

Participants also rated their agreement level with the statement "the above message is severe" for both a human-written message and an AI-generated message (A.5/A.11). The response distribution is presented in Fig. 3c. For both the sexism (n = 159) and racism (n = 99) groups, there were no significant differences in the perceived severity between human-written and AI-generated messages. The one exception is the abuse (n = 105) group where participants significantly viewed AI-generated message as more severe than human-written message. Specifically, 57.14 % of participants considered the human-written message as more severe compared to 79.05 % for the AI-generated message.

4.1.4. *Immediate reaction to cyberbullying messages*

Participants reported their immediate reactions to both human-written and AI-generated messages (A.6/A.12), with response distribution shown in Fig. 3d. In the sexism (n = 159) and racism (n = 99) groups, the most common response was "intent to report" for both message types: 46.54 % and 59.60 % participants chose to report the human-written message, respectively, while 50.31 % and 54.55 % did so for the AI-generated message. However, the abuse group (n = 105) exhibited a distinct trend: while 40.95 % of participants ignored the human-written message, a significant majority (60.95 %) chose "intent to report" for the AI-generated message. This aligns with the prior findings, where AI-generated message elicited higher discomfort, greater perceived harm, and increased severity compared to human-written ones. These results suggest that participants are more likely to take active measures, such as reporting, when they perceive a message as particularly harmful or severe.

4.1.5. *Key RQ1 Takeaways*

First, participants across all groups felt uncomfortable with both human-written and AI-generated cyberbullying messages. Second, the majority perceived both messages as harmful and severe. Third, while participants in the sexism and racism groups were more likely to report both human-written and AI-generated messages, those in the abuse group were more inclined to ignore human-written messages but report AI-generated ones. Finally, when comparing human-written and AI-generated messages in terms of comfort level, perceived harm and severity, we found similar distribution in the sexism and racism groups, but a significant difference in the abuse group. Overall, these findings highlight that AI-generated cyberbullying content can be just as, if not more, harmful than human-written content.

4.2. *RQ2: differentiating AI-generated and human-written cyberbullying*

To answer RQ2, participants were first informed that the presented messages could be either human-written or AI-generated. They were then asked to identify the author of each message and describe the aspects they considered in making their judgments.

4.2.1. *Identification accuracy*

Participants were asked to indicate who they believed wrote the presented message (Questions B.1 and B.4). The overall accuracy rates for the sexism, racism, and abuse groups were 44.97 %, 39.39 %, and 41.90 %, respectively, suggesting that participants struggled to distinguish between human-written and AI-generated cyberbullying messages. To delve into the details, we obtained the confusion matrix for

Table 3
Confusion matrix for each group.

Type of Cyberbullying	Ground Truth	Human Prediction		
		Human	AI	Not_Sure
sexism	Human	127	9	23
	AI	119	16	24
	Not_Sure	0	0	0
racism	Human	68	13	18
	AI	71	10	18
	Not_Sure	0	0	0
abuse	Human	83	11	11
	AI	90	5	10
	Not_Sure	0	0	0

each group, as shown in Table 3. The results showed that a majority of participants mistakenly attributed the AI-generated messages to a human author. These findings highlight the remarkable ability of LLMs to generate content that closely mimics human writing, making it challenging for individuals to differentiate between the two.

4.2.2. Aspects Considered in judging the authorship of human-written messages

Participants were asked to identify the aspects they focused on when considering the author of the human-written message (Question B.2), with results summarized in Fig. 4. Additionally, an open-ended question (Question B.3) is designed for participants to provide explanations to their responses. Overall, we found that for all groups, participants primarily focused on the emotion and the wording of the message when considering the author of the human-written message.

Sexism group (n = 159). Participants primarily based their judgments on emotions (42.77 %) and wording (32.70 %) when identifying human-written messages. ***Among the 68 participants who chose emotions***, 15 participants judged based on emotional intensity, associating strong language with human authorship, e.g., *“The use of strong language and personal frustration suggests a human wrote it”* (PS-46); 13 participants linked emotions expressed through personal opinions with human authorship, e.g., *“The message conveys a personal opinion and emotional judgment, which seems more characteristic of a human writer. Additionally, the phrasing and word choice reflect a subjective viewpoint that AI might not typically express as naturally”* (PS-3). Others cited casual and informal tone, personal expressions and feelings, and biased or stereotypical expressions. While two participants doubted on AI’s capability to generate emotional content, another two found the emotions too ambiguous to judge. ***Among 52 participants who focused on wording***, nine noted that words used for personal beliefs suggested human authorship, e.g., *“Message contains personal anecdotes and subjective experiences, which are often characteristics of human-written content”* (PS-125); eight participants highlighted the use of slang and informal language, e.g., *“Slang terms like ‘jus’ and ‘tht,’ casual profanity (‘shit’), and informal phrasing (‘holla back’) are all more typical of human conversation than the typically neutral and formal language used by AI models”* (PS-10). Others pointed to human-like communication patterns, use of biased language, and simplicity of word choices, with some citing the use of sarcastic and sensitive words as indicator of human authorship.

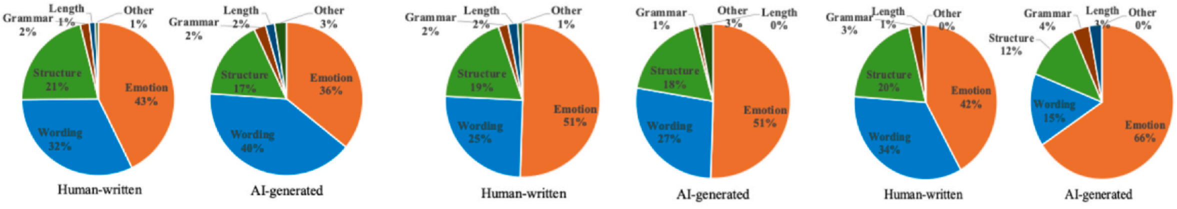
Racism group (n = 99). Participants primarily based their judgments on emotions (50.51 %) and wording (25.25 %) when distinguishing human-written messages. ***Among 50 participants who focused on emotions***, 17 associated strong emotional intensity with human authorship, e.g., *“The message contains strong opinions and controversial statements, which are typically more indicative of a human writer”* (PR-73); nine pointed to biases and hostile intent as indicative of human authorship, e.g., *“The message has a specific and hostile intent that seems characteristic of a human’s biased thinking”* (PR-93). Others highlighted negative emotions (e.g., xenophobic and discriminatory tone), sophisticated emotional expressions and personal beliefs to human authorship. These responses suggest that participants perceived AI-generated messages as generally robotic and emotionless. ***Among 25 participants who focused on wording,*** seven cited human-specific word choices, e.g., *“AI tends to have a much more formal and structured way of presenting text. Especially the ‘Ha’ at the end makes it come across as a human message instead of AI”* (PR-62); six linked sensitive and emotionally charged words to human authorship, e.g., *“The use of phrases like ‘being duped’ and the implication of a conspiracy suggest a certain emotional intensity and possibly a biased perspective”* (PR-46). Participants also identified provocative, opinionated wording, as well as the use of historical references, as indicative of human authorship, whereas formal language was more commonly associated with AI-generated content.

Abuse group (n = 105). Participants primarily based their judgments on emotions (40.00 %) and wording (34.29 %). ***Among 42 participants who considered emotions***, 14 recognized emotional expression that seemed distinctly human, e.g., *“The emotional expression and frustration in the sentence suggest a human author. AI-generated text often lacks this level of personal emotion and context-specific frustration”* (PA-10); nine noted strong emotional content such as anger e.g., *“Message 1 contains strong emotional language and derogatory terms, indicating a personalized expression of frustration or anger”* (PA-31). Others associated introspection, personal experiences, and nuanced emotions with human authorship. ***Among 38 participants who focused on wording***, 21 cited the informal tone and colloquial language as strong indicators of human authorship, e.g., *“The sentence wording in Message 1 includes colloquial phrases and slang, which are typically used in informal, human conversations”* (PA-53); four highlighted emotional wording, e.g., *“The emotional intensity and the choice of words looks more indicative of human expression rather than AI-generated content”* (PA-54). Participants also explained the use of offensive words, humor, and sentimental expressions as human-like traits.

4.2.3. Aspects Considered in judging the authorship of AI-generated messages

Participants were also asked to identify the aspects they focused on when assessing the authorship of the AI-generated message (B.2/B.5). Across all groups, emotions and wording were the most frequently considered aspects.

Sexism group (n = 159). Participants primarily based their judgments on wording (39.62 %) and emotions (36.48 %). ***Among 63 participants who focused on wording,*** 34 associated wordings reflecting personal opinion/bias/belief with human authorship, recognizing that AI-generated content is typically not subjective, e.g., *“The statement reflects traditional and biased views on gender roles, suggesting a personal belief that is likely human-generated. AI-generated content typically avoids such direct and subjective opinions”* (PS-101); 10 noted complexity and clarity of wording, suggesting that AI-generated content is often more complex, e.g., *“I’m leaning toward it being by a human because it’s simple, understandable, and would fit in in a conversation or post on social media”* (PS-11). Others cited confrontational and casual tones in wording as human traits, while five provided other reasons. ***Among 58 participants who considered emotions***, 23 identified stereotype and personal bias as indicators of human authorship, e.g., *“The message uses strong gender stereotypes and traditionalist language that suggests a deeply ingrained belief system. The direct and confrontational wording is more characteristic of human”* (PS-19); 16 pointed to personal experiences and thoughts, e.g.,



(a) Sexism Group

(b) Racism Group

(c) Abuse Group

Fig. 4. Aspects considered by individuals when judging the author of the presented message.

"Message 2 seems to include personal experiences and emotions that might be more authentically expressed by a human" (PS-73). Others noted emotional depth, complexity, and provocativeness as features of AI-generated messages, with four recognizing distinct human emotional expressions and two believing AI models could not generate certain emotions.

Racism group (n = 99). Participants primarily based their judgments on emotions (50.51 %) and wording (27.27 %). ***Among 50 participants who considered emotions***, 17 emphasized emotional intensity as a human trait, e.g., *"The emotional intensity and specific accusations suggest a human perspective, but AI models can mimic such tone and structure"* (PR-72); ten identified hostile emotions as indicative of human authorship, e.g., *"The message contains strong, inflammatory language and a clear expression of hatred, which are more typical of human-generated content"* (PR-6). Others cited more specific emotional cues, such as authenticity of emotions, strong biases and directness as human-like characteristics. ***Among 27 participants who focused on wording***, eight pointed to the choice of specific words, e.g., *"choice of words seem more aligned with human communication"* (PR-58); another eight associated strong biases in the wording with human authorship, e.g., *"The message contains specific accusations and a narrative that implies a strong bias and intentional harm, which seems more characteristic of human writing"* (PR-61). Participants also recognized the following as indicators of human writing: (a) the use of inflammatory language, (b) the lack of nuances in wording, and (c) comparable wording style to other AI-generated texts.

Abuse group (n = 105). Participants primarily relied on emotions (65.71 %) and wording (15.24 %). ***Among 69 participants who considered emotions***, 27 associated that the harsh and direct language with human authorship, e.g., *"The first message reads as a very direct, emotionally charged insult that seems much more characteristic of how a human would lash out angrily"* (PA-64); 14 noted strong negative emotions as indicative of human authorship, e.g., *"I focus on the emotions resented in the sentences as the strong negative sentiments expressed suggest genuine human emotion rather than a calculated response"* (PA-14). Others linked hostility (8), emotional tone (8), and authenticity of emotions (5) to human authorship. ***Among 16 participants who considered wording,*** six identified derogatory language and strong emotional tone as human traits, e.g., *"The use of strong derogatory language and the hostile tone convey intense emotions."* (PA-2); four associated direct insults and emotionally charged language with human authorship, e.g., *"The direct insult and lack of context make it seem like something a human might impulsively write"* (PA-10); two considered formality and complexity of sentence wording, while four cited other reasons.

4.2.4. Key RQ2 Takeaways

First, for all groups, the majority of participants struggled to differentiate between human-written and AI-generated cyberbullying messages, with accuracy rates below 50 %. Specifically, in most cases, participants incorrectly identified AI-generated messages as human-written. Second, when assessing messages authorship, participants primarily focused on emotions and wording. Strong emotional intensity, personal biases, and subjective language were key factors that led participants to attribute both human-written and AI-generated cyberbullying messages to human authors. These findings highlight the deceptive realism of AI-generated content, reinforcing concerns about the potential misuse of AI in generating persuasive and emotionally charged messages.

4.3. RQ3: attitudes towards AI-generated cyberbullying

To answer RQ3, participants were explicitly informed that the presented message was generated by AI. They were then asked to respond to Question Set C, which examined their attitudes towards AI-generated cyberbullying.

4.3.1. Comfort level with AI-generated cyberbullying

Participants were asked to indicate their comfort level upon receiving an AI-generated cyberbullying message (Question C.2). The response distribution is presented in Table 4. The results show that a majority of participants remained uncomfortable with AI-generated cyberbullying messages, even when aware of their AI origin. Notably, participants had previously rated their comfort level with the same message in Question A.7 (Section 4.1), without knowing whether it was written by AI or a human. A comparison of responses indicates a reduction in discomfort across all three groups once participants were informed of the AI authorship. However, the decrease in discomfort was statistically significant only in the sexism group, as indicated by the Wilcoxon signed rank test ($p < 0.05$). This suggests that source disclosure can to some extent reduce perceived discomfort, though the effect is limited in some cases.

4.3.2. Opinions on Protection Against AI-generated cyberbullying

Fig. 5 presents the distributions of participants' agreements on the statement, "Protecting internet users against cyberbullying launched by AI is as important as that by humans" (Question C.3). We found a vast majority of participants supported this notion, with 96.23 % in the sexism group, 91.92 % in the racism group, and 95.24 % in the abuse group somewhat or strongly agreeing. Participants further elaborated on their reasoning in Question C.4.

Sexism group (n = 159). ***Among 157 participants who agreed or strongly agreed*** on the importance of protecting users against AI-generated cyberbullying, 86 cited its serious psychological effects on individuals' well-being and mental health, e.g., *"Victims may experience stress, anxiety, depression, and a decrease in self-esteem. In severe cases, it can lead to suicidal thoughts or actions"* (PS-92); 27 participants emphasized AI's potential for greater harm than human-written cyberbullying due to its scalability and anonymity, e.g., *"AI can be programmed to spread hate speech and harassment at an alarming rate, overwhelming victims much faster than a human bully could"* (PS-10); 18 argued that protection from cyberbullying should be prioritized regardless of whether the content is AI-generated or human-written; 11 explicitly emphasized the equal importance of safeguarding users from both sources; seven highlighted the human role in mitigating AI-generated bullying, and eight provided other reasons.

Racism group (n = 99). ***Among 93 participants who agreed with the statement,*** 61 explained that both human-written and AI-generated cyberbullying inflict significant harms, e.g., *"Cyberbullying, regardless of whether it originates from AI or humans, can have equally harmful consequences on individuals' mental health and well-being."* (PR-47), 14 participants pointed out that the source of cyberbullying does not diminish the need for protection, e.g., *"Addressing both forms of cyberbullying with equal attention and resources is essential to ensure a safe and inclusive online environment for everyone"* (PR-26); six highlighted AI-generated content's broader impact on online safety; two explicitly emphasized equal importance of protecting users from both AI-generated and human-written cyberbullying messages; two mentioned AI's potential role in cyberbullying detection; and seven cited other reasons such as ethical violations.

Abuse group (n = 105). ***Among 101 participants who agreed with the statement,*** 61 emphasized the significant emotional and psychological harm caused by AI-generated cyberbullying, e.g., *"AI-generated cyberbullying can still cause harm and distress to individuals, even if it lacks the emotional depth of human-generated content"* (PA-62); 12 highlighted AI's potential for greater, more damaging impact due to automation and scalability, e.g., *"AI-driven cyberbullying could be highly automated, scalable, difficult to detect and more damaging"* (PA-95); eight recognized AI's benefits; four participants critiquing the unethical use of AI to engage in cyberbullying; eight asserted that cyberbullying is harmful regardless of authorship; eight had other reasons. The remaining four participants felt neutral about the statement for various reasons, including discomfort when confronted with AI-generated cyberbullying, e.g., *"feel better if*

Table 4
Comfort Level on Receiving AI-generated Cyberbullying Messages. Percentage changes in parentheses are calculated between the responses in Questions A.7 shown in Fig. 3a and Question C.2. (C.2: How comfortable would you feel if you receive presented AI-generated message and be informed that this message is generated by an AI model?).

Types of Cyberbullying	Very Comfortable	Comfortable	Neutral	Uncomfortable	Very Uncomfortable
sexism	6.92 % (−2.52 %)	14.47 % (+5.66 %)	14.47 % (+7.55 %)	29.56 % (+0.00 %)	34.59 % (−10.69 %)
racism	8.08 % (−1.01 %)	18.18 % (+1.01 %)	15.15 % (+1.01 %)	25.25 % (+1.01 %)	33.33 % (−2.02 %)
abuse	7.62 % (−2.86 %)	13.33 % (+4.76 %)	8.57 % (+3.81 %)	37.14 % (+8.57 %)	33.33 % (−14.29 %)

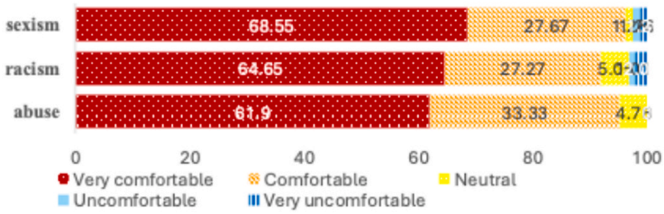


Fig. 5. Participants' agreement level on the importance of protection against AI-generated cyberbullying.

written by ai. if it wasnt someone doesnt know how to correctly use words" (PA-103).

4.3.3. Key RQ3 Takeaways

First, changes in participants' comfort levels were generally insignificant except in the sexism group. Second, most participants agreed that protecting internet users from AI-generated cyberbullying is equally important, regardless of whether it is initiated by AI or humans. This consensus was largely driven by concerns over the significant harm and scalability of AI-generated cyberbullying, which led to a greater negative impact. These findings highlight the urgent need for careful consideration of AI's role in cyberbullying and implementing protective measures to mitigate its widespread effects.

4.4. RQ4: Factors influencing judgment and perception

To investigate factors influencing participants' judgments and perceptions (RQ4), we analyzed independent variables across demographics, prior experience, and attitudes towards freedom of speech. Appropriate regression models were applied based on data types, and Odds Ratios (OR) were computed to improve interpretability (Frik et al., 2022). Additionally, we used bidirectional stepwise regression for variable selection, guided by the corrected Akaike Information Criterion (AIC), which adjusts the traditional AIC to accommodate small to moderate sample sizes (Hurvich & Tsai, 1989).

4.4.1. Factors influencing the identification accuracy

Each participant evaluated one human-written and one AI-generated cyberbullying message, with accuracy coded as 1 (correct) or 0 (incorrect). Given the categorical nature of the dependent variable, we applied logistic regression analysis.

Identifying human-written messages. Table 5 presents the results of the stepwise regression analysis. In the ***sexism*** group, participants who reported frequent use of AI tools (e.g. daily or weekly) were 3.494 times more likely to correctly identify human-written sexist messages ($p = 0.011$, OR = 3.494). Two additional variables demonstrated marginal significance: participants who had ever used AI ($p = 0.063$) and white participants ($p = 0.066$) tended to perform slightly better in correctly identifying such messages. In the ***racism*** group, no variables reached conventional significance levels; however, both prior AI use and lower being-bullied frequency showed positive trend. In the ***abuse*** group, prior experience with AI had a substantial effect: participants who had used AI were over 10 times more likely to correctly identify human-written abusive messages ($p = 0.011$). Overall, these findings suggest that AI experience, particularly having ever used AI tools, consistently improves participants' ability to identify human-written content, especially in the abuse and sexism groups. Demographic factors such as race and age may also moderate identification performance, though their effects did not reach statistical significance and warrant further investigation.

Table 5
Logistic regression on whether participants correctly identified human-written cyberbullying message (Question B.1).

	Coef.	Std. Err.	z	P-value	95 % CI	Odds Ratio
Sexism Group						
Intercept	−1.184	0.821	−1.442	0.149	[-2.794, 0.425]	0.3059
use_ai_freq	1.251	0.491	2.547	0.011*	[0.288, 2.214]	3.494
use_ai	1.365	0.733	1.862	0.063	[-0.072, 2.801]	3.9145
age	−2.100	1.293	−1.624	0.104	[-4.635, 0.434]	0.1224
race	1.148	0.625	1.836	0.066	[-0.078, 2.373]	3.1503
being_bullied	−0.820	0.500	−1.64	0.101	[-1.799, 0.160]	0.441
Racism Group						
Intercept	−0.484	0.620	−0.782	0.435	[-1.699, 0.731]	0.6161
bullied_freq	0.787	0.488	1.612	0.107	[-0.170, 1.743]	2.1964
use ai	0.865	0.554	1.563	0.118	[-0.220, 1.950]	2.376
Abuse Group						
Intercept	−1.342	0.985	−1.363	0.173	[-3.273, 0.588]	0.261
use ai	2.364	0.924	2.558	0.011*	[0.553, 4.175]	10.632
bullied_freq	0.761	0.523	1.454	0.146	[-0.265, 1.786]	2.140

$^*p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Identifying AI-generated messages. Table 6 presents the stepwise regression results examining participants' ability to correctly identify AI-generated cyberbullying messages. In the ***sexism*** group, gender showed a marginal effect, with male participants being less likely to correctly identify AI-generated sexist messages ($p = 0.079$, OR = 0.386), suggesting a possible gender difference in perception or sensitivity to AI-generated gendered content. In the ***racism*** group, two variables reached statistical significance. Individuals who had used AI tools were significantly less likely to correctly identify AI-generated racist content ($p = 0.005$, OR = 0.126), potentially reflecting overconfidence or familiarity bias in AI users. Additionally, age showed a marginally significant effect ($p = 0.050$, OR = 18.999), with younger participants (i.e., < 40) being substantially more likely to correctly identify AI-generated racism messages. In the ***abuse*** group, gender effect was retained in the model but did not reach significance ($p = 0.216$, OR = 0.314), though the odds ratio again suggested that male participants may be less accurate. In general, these results suggest that both demographic factors (gender and age) and AI experience influence the ability of individuals to detect AI-generated harmful content, particularly in the context of racism. The

Table 6
Logistic regression on whether participants correctly identified AI-generated cyberbullying message (Question B.5).

	Coef.	Std. Err.	z	P-value	95 % CI	Odds Ratio
Sexism Group						
Intercept	−1.551	0.416	−3.726	0.000***	[-2.366, −0.735]	0.2121
gender	−0.953	0.542	−1.759	0.079	[-2.014, 0.109]	0.3857
Racism Group						
Intercept	−0.876	0.532	−1.645	0.100	[-1.919, 0.168]	0.4167
use_ai	−2.069	0.739	−2.799	0.005**	[-3.518, −0.620]	0.1263
age	2.944	1.504	1.957	0.050*	[-0.004, 5.893]	18.999
Abuse Group						
Intercept	−2.367	0.604	−3.920	0.000***	[-3.551, −1.184]	0.094
gender	−1.159	0.938	−1.236	0.216	[-2.997, 0.679]	0.314

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

findings highlight possible perceptual or experiential differences that shape how users interpret AI-generated cyberbullying across different categories.

4.4.2. Factors influence Individual's perception

To examine factors influencing participants' comfort levels in receiving AI-generated messages, we employed ordered logistic regression. The results are presented in Table 7. In both the ***sexism*** and ***abuse*** groups, participants who prioritized preventing harmful messages over protecting freedom of speech reported significantly lower comfort levels (sexism: $p < 0.001$, OR $= 0.258$; abuse: $p < 0.001$, OR $= 0.220$), indicating that greater concern for online harm is associated with increased sensitivity to these types of content. In contrast, in the ***racism*** group, three variables showed significant associations with comfort level. Participants who had experienced bullying more than once had a higher level of comfort with AI-generated racist messages ($p = 0.019$, OR $= 2.887$), while those who had previously used AI ($p = 0.025$, OR $= 0.351$) and male participants ($p = 0.035$, OR $= 0.439$) reported significantly lower comfort. Overall, the findings suggest that participants' attitude toward online harm strongly influences their emotional responses to AI-generated sexist and abusive messages, whereas prior personal experiences, including AI use, bullying history, and gender play a more prominent role in shaping responses to racist content.

In addition, to explore how individual factors affect participants' changes in comfort level after discovering that a message was generated by AI, we conducted linear regression analyses. The change in comfort level was determined by calculating the difference between participants' reported comfort levels before and after learning that the message was AI-generated, which is a continuous variable. The results are presented in Table 8. In the ***sexism*** group, education emerged as a significant factor ($p = 0.004$, OR $= 2.052$), with higher-educated participants showing a greater increase in comfort after discovering the message was AI-generated. In the ***racism*** group, age was included in the model but was not a significant predictor ($p = 0.501$, OR $= 1.649$), though younger participants (< 40) tended to show a higher comfort change. In the ***abuse*** group, both attitude and gender significantly influenced change in comfort level. Participants who prioritized preventing harmful messages over protecting freedom of speech reported greater increases in comfort after learning the message was AI-generated ($p = 0.001$, OR $= 2.236$), while male participants reported smaller comfort changes compared to females ($p = 0.037$, OR $= 0.606$). Overall, the findings suggest that education level and attitudes toward online harm are key factors in how participants emotionally recalibrate after discovering a message's AI authorship, particularly in the contexts of sexist and abusive content. Gender also appears to shape these shifts, with males being less responsive in their comfort adjustment for abusive content.

Table 7
Ordered logistic regression on participants' comfort level on receiving AI-generated message and being informed that the message is generated by AI (Question C.2).

	Coef.	Std. Err.	z	P-value	95 % CI	Odds Ratio
Sexism Group						
attitude	−1.355	0.329	-4.114	0.000***	[-2.001, −0.709]	0.258
Racism Group						
bullied_freq	1.060	0.450	2.354	0.019*	[0.177, 1.943]	2.887
use_ai	−1.048	0.468	-2.238	0.025*	[-1.966, −0.130]	0.351
gender	−0.823	0.391	-2.108	0.035*	[-1.589, −0.058]	0.439
Abuse Group						
attitude	−1.514	0.412	-3.674	0.000***	[-2.322, −0.706]	0.220

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

4.4.3. Key RQ4 Takeaways

First, participants with prior experience using AI tools consistently demonstrated better ability to identify human-written cyberbullying messages, particularly in the sexism and abuse categories. Second, gender and age were associated with participants' ability to detect AI-generated messages, especially in the racism category, suggesting possible perceptual or cognitive differences across groups. Third, attitudes toward online harm shape emotional responses. Participants who prioritized the prevention of harmful content over the protection of freedom of speech reported lower comfort levels with AI-generated sexist and abusive messages, indicating a strong relationship between values and emotional sensitivity. Finally, education and attitudes impact comfort level change: upon learning that a message was AI-generated, participants with higher education or stronger harm-prevention attitudes showed greater shifts in comfort, particularly for sexist and abusive content. Male participants, however, appeared less responsive in adjusting their comfort levels, in response to abusive messages.

5. Discussion

5.1. Implications and recommendations

The harm potential of AI-generated cyberbullying. Our findings in Section 4.1 highlight that AI-generated cyberbullying content can be equally or even more harmful than human-written content in terms of users' comfort levels, perceived harm, and severity. These results emphasize the need for serious attention to AI-generated harmful content, as it can evoke strong negative reactions similar to those caused by human-written content. While AI-driven threats are gaining increased attention (Karasavva & Noorbhai, 2021; Pechenik Gieseke, 2020), research on AI-generated cyberbullying remains scarce. Our study assesses the risks associated with AI-generated harmful content by examining human perceptions of AI-generated cyberbullying, particularly focusing on LLMs, which pose an even greater risk due to their accessibility and widespread adoption. A recent tragic case illustrates these risks: a 14-year-old committed suicide after engaging with an AI in a role-playing context (Yang, 2024), underscoring the profound ethical concerns surrounding AI-generated content. Given the rapid integration of AI into online interactions, we urge the research community to expand its efforts in understanding and mitigating malicious use of AI. In particular, cyberbullying remains a pervasive issue among minors (Peebles, 2014; Anderson, 2018; Manuel Ga'mez-Guadix & Estibaliz Mateos-Pérez, 2019; Gohal et al., 2023), necessitating stronger

Table 8
Ordinary least squares regression on participants' comfort level change on receiving AI-generated message and being informed that the message is generated by AI (Questions A.7 and C.2).

	Coef.	Std. Err.	z	P-value	95 % CI	Odds Ratio
Sexism Group						
education	0.7187	0.245	2.929	0.004**	[0.234, 1.203]	2.0518
Racism Group						
age	0.5	0.741	0.675	0.501	[-0.970, 1.970]	1.6487
Abuse Group						
attitude	0.8048	0.239	3.372	0.001***	[0.331,1.278]	2.2362
gender	−0.5007	0.237	-2.113	0.037*	[-0.971, −0.031]	0.6061

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

protective measures. Overall, the potential for AI misuse to escalate into a broader societal threat cannot be underestimated, and proactive measures are essential to safeguard against these dangers.

Challenges in distinguishing AI-generated content from human-written content. As summarized in Section 4.2, most participants struggled to differentiate between human-written and AI-generated cyberbullying messages, with accuracy rates below 50 %. This finding aligns with prior studies in other domains, where detection of AI-generated content also remains low, for example, news articles (52 %) (Brown et al., 2020), media image (49.93 %), and media text (54.48 %) (Frank et al., 2024). In exploring potential reasons for this low accuracy, we discovered that many participants believe that AI-generated content is inherently neutral and devoid of emotion or bias. Therefore, raising public awareness of the potential risks associated with AI-generated content is essential. Our findings also indicate that knowing the origin of AI-generated cyberbullying messages can, to some extent, reduce individuals' discomfort levels (Section 4.3), which is consistent with prior research showing that source disclosure significantly influences perceptions of messages (Lim & Schmälzle, 2024). We recommend that social media platforms implement transparent source disclosure, which will help users better understand the content they encounter. This practice could ultimately reduce the harm associated with AI-generated cyberbullying.

Behavioral responses and limitations of human moderation. Our study provides valuable insights into the roles of human and automated detection in combating AI-generated cyberbullying. We found that participants are more likely to take active measures when they perceive a message as particularly harmful or severe regardless of whether it was authored by a human or AI (Section 4.1). This finding aligns with prior research showing that perceived severity increases the likelihood of intervention, especially from bystanders (Bastiaensens et al., 2014; DeSmet et al., 2012; Macaulay et al., 2022). Our results extend this literature by showing that perceived severity similarly influences behavioral response from the victim's perspective. This behavioral tendency suggests that existing policies on social media platforms, which offer options to block, report, or restrict unwanted interactions (Instagram, 2024a, 2024b; Snapchat, 2024), can be effective when users recognize content as harmful. Encouragingly, a majority of participants across all three groups expressed an intention to report AI-generated cyberbullying messages (Section 4.1.4). However, the scalability and automation of AI-generated content present new challenges that limit the effectiveness of relying solely on user reports and human moderation. To address this emerging threat, social media platforms must recognize AI-generated cyberbullying as a new attack vector and adapt their policies accordingly. This includes refining AI detection mechanisms, strengthening moderation frameworks, and educating users on AI-generated threats to ensure a more proactive and comprehensive approach in online safety. Research has shown that targeted training can improve individuals' ability to identify AI-generated content (Dugan et al., 2023). Thus, educating individuals about linguistic cues and generative patterns could serve as a feasible safeguard against AI-generated cyberbullying.

Predictors of emotional sensitivity to AI-generated cyberbullying. Our findings in Section 4.4 confirm and extend prior research by identifying key demographic and experiential factors such as gender, education, and bullying history as significant influences on individuals' emotional responses to cyberbullying. For example, male participants showed lower emotional sensitivity to harmful content (Section 4.4.2), supporting prior evidence that females and marginalized groups experience greater psychological impact and perceive higher levels of harm in cyberbullying incidents (Schodt et al., 2021; Wong et al., 2018). Similarly, the finding that individuals with repeated bullying experiences were more tolerant of AI-generated racist messages aligns with research suggesting that prior victimization may contribute to desensitization or diminished harm perception in both adolescents and adults (Hemphill & Heerde, 2014; Zhang et al., 2022). Education and age also emerged as important factors, influencing participants' ability to detect AI-generated messages and their emotional recalibration after learning the messages' origin, patterns that mirror their known associations with cyberbullying perpetration and victimization in past studies (Wang et al., 2019b; L'opez-Castro & Priegue, 2019; Patterson et al., 2019; Balakrishnan, 2015). Importantly, we extend the existing literature by uncovering how value-based attitudes and AI familiarity influence emotional responses to AI-generated cyberbullying. Specifically, participants who prioritized preventing harm over protecting free speech reported significantly lower comfort levels with sexist and abusive content, indicating that moral and ethical orientations substantially influence harm sensitivity (Section 4.4.2). Furthermore, those with prior experience using AI tools were less comfortable with AI-generated racist messages, possibly reflecting greater awareness of AI's persuasive capabilities or increased attunement to algorithmic bias. These insights introduce new psychological and technological dimensions to the cyberbullying literature, emphasizing that both normative beliefs and prior technological engagement can shape how individuals process and respond to AI-mediated online harm.

5.2. Limitations and Future work

This study has several limitations that can be overcome by future studies. First, there is a demographic bias in the participant sample, which may not adequately represent the full spectrum of the population. This bias could influence the interpretation of results, particularly when dealing with sensitive topics like sexism, racism, and abuse, as the perspectives and reactions of a more diverse sample might yield different outcomes. Future studies should aim to recruit a more diverse and representative participant sample by utilizing stratified sampling techniques to ensure that various demographic groups are adequately represented. Second, the study's scope was confined to a specific set of sentences focusing on sexism, racism, and abuse, without expanding to other types of potentially harmful or controversial content. This limitation restricts the range of insights that could be drawn, as it overlooks other forms of discrimination, bias, or abuse that might also be relevant. Therefore, future research should include a wider variety of sentence types that cover other forms of harmful or controversial content, such as homophobia, ageism, ableism, and religious discrimination, among others. This will allow for a more thorough assessment of the AI model's ability to handle a broader spectrum of sensitive topics, providing a more complete evaluation of its performance and limitations. Finally, the current analysis was limited to using only the Llama model, without exploring or comparing alternative models. Future studies should incorporate a comparative analysis of multiple AI models. By evaluating different models, researchers can gain insights into the strengths and weaknesses of each model in processing various types of content.

6. Conclusion

In this paper, we investigated risks associated with AI-generated cyberbullying. We conducted a user study to examine individuals' perceptions and attitudes towards AI-generated cyberbullying content. Our findings show that AI-generated cyberbullying messages cause significant discomfort and perceived harm comparable to human-written message. Additionally, AI-generated messages are almost indistinguishable from human-written messages, with many participants misidentifying AI-generated messages as human-written. We also found that participants with prior experience using AI tools consistently demonstrated higher accuracy in identification, while their attitudes towards online harm significantly influenced their comfort levels. Overall, this work represents a significant step in understanding the human perception of AI-generated cyberbullying and its potential psychological impacts. Our findings emphasize the urgent need for continued development of comprehensive strategies to safeguard individuals for AI-generated cyberbullying.

CRediT authorship contribution statement

Weiping Pei: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fangzhou Wang:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **Yi Ting Chua:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Acknowledgements

This research was supported in part by the National Science Foundation under Award No. 2246143 and No. 2521383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Four Sets of Questions

Participants for this survey were recruited from the Amazon Mechanical Turk (MTurk) and all participants were given the same questions. Participants were presented with the following 32 questions. We list them in four sets (i.e., A to D) and added notes in bold font before each set to improve the readability of the paper.

Question Set A: These questions are about participants' perception of cyberbullying messages.

Please assume you see the following message on a social media platform such as Twitter or Facebook, and then answer six questions:
"human-written cyberbullying sentence"

A.1 How comfortable would you feel if you received the above message?

- ○ Very comfortable
- ○ Comfortable Neither
- ○ comfortable nor uncomfortable
- ○ Uncomfortable
- ○ Very uncomfortable

A.2 I believe the above message is an example of cyberbullying.

- ○ Strongly agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Strongly disagree

A.3 Which of the following best describes content of the above message.

- ○ Abuse
- ○ Racism
- ○ Sexism
- ○ Not a cyberbullying message
- ○ Other: can't tell without context

A.4 Do you think the above message is harmful?

- ○ Strongly agree … ○ Strongly disagree (5-level scale)

A.5 Do you think the above message is severe?

- ○ Strongly agree … ○ Strongly disagree (5-level scale)

A.6 What are your immediate reactions after reading the above sentence?

- ○ Intent to report
- ○ Intent to respond
- ○ Delete or block user
- ○ Ignore

Please assume you see the following message on a social media platform such as Twitter or Facebook, and then answer six questions: "human-written cyberbullying sentence"

Questions A.7 to A.12 are identical to Questions A.1 to A.6.

Question Set B: These questions are about participants' judgements regarding authors of provided cyberbullying messages.

Please consider both messages together and answer the following questions:
Message 1: {human-written cyberbullying sentence} Message 2: {AI-generated cyberbullying sentence}

*B.1 Each of the above messages could be written by either a human (human-written) or by an Artificial Intelligence (AI-generated). Who do you think wrote the **message 1**?*

- ○ Human
- ○ Artificial Intelligence
- ○ Not sure

*B.2 Which aspects of the sentence do you focus on when considering the author of **message 1**?*

- ○ Emotions presented in sentences
- ○ Sentence structures
- ○ Sentence wording
- ○ Sentence grammars
- ○ The length of sentences Other (Please specify in the following question)

B.3 Can you please explain your answer to the question above? Please write down at least two sentences: (Your answer should be anonymous and please do not enter any identifiable information in this question.)

B.4 Each of the above messages could be written by either a human (human-written) or by an Artificial Intelligence (AI-generated). Who do you think wrote the **message 2**?

- ○ Human
- ○ Artificial Intelligence
- ○ Not sure

*B.5 Which aspects of the sentence do you focus on when considering the author of **message 2**?*

- ○ Emotions presented in sentences
- ○ Sentence structures
- ○ Sentence wording
- ○ Sentence grammars
- ○ The length of sentences Other (Please specify in the following question)

B.6 Can you please explain your answer to the question above? Please write down at least two sentences: (Your answer should be anonymous and please do not enter any identifiable information in this question.)

Question Set C: These questions are about participants' attitudes towards AI-generated cyberbullying messages.

*C.1 How comfortable would you feel if you receive **message 1** and be informed that this message is generated by an AI model?*

- ○ Very comfortable … ○ Very uncomfortable (5-level scale)

*C.2 How comfortable would you feel if you receive **message 2** and be informed that this message is generated by an AI model?*

○ Very comfortable … ○ Very uncomfortable (5-level scale)

C.3 Protecting internet users against cyberbullying launched by AI is as important as that by humans?

○ Strongly agree … ○ Strongly disagree (5-level scale)

C.4 Can you please explain your answer to the question above? Please write down at least two sentences. (Your answer should be anonymous and please do not enter any identifiable information in this question.)

Question Set D: These questions are about participants' prior experience and demographics.

D.1 Do you have prior experience with being cyberbullied?

○ Yes
○ No
○ I prefer not to say

D.2 How many times have you experienced being cyberbullied over the past 12 months?

○ None
○ One
○ Two
○ More than three
○ I prefer not to say

D.3 Do you have prior experience with using AI products (such as ChatGPT, Llama)?

○ Yes
○ No
○ I prefer not to say

D.4 On average, how often do you use AI products (such as ChatGPT, Llama) to generate textual content?

○ Daily
○ Weekly
○ Monthly
○ Yearly
○ Never
○ Other

D.5 What is your attitude towards freedom of speech?

○ Positive
○ Negative

D.6 If you absolutely have to choose between protecting freedom of speech and preventing harmful messages from spreading, which is more important to you?

○ Protecting Freedom of Speech
○ Preventing Harmful Message from Spreading

D.7 What race best describes you?

○ White/Caucasian
○ Black/African American
○ Indian/Alaskan Native Asian/Pacific Islander
○ Multiracial
○ I prefer not to say Other

D.8 What best describes your gender?

○ Male
○ Female
○ I prefer not to say
○ Other

D.9 Which of the following best describes the highest level of education you have completed?

○ Less than high-school diploma or GED
○ High school diploma or GED

- ○ Some college
- ○ Associate degree/2-year college degree
- ○ Bachelor's degree/4-year college degree
- ○ Master's or professional degree/JD/LLM
- ○ Doctorate degree (e.g., Ph.D., Ed.D., M.D)

D.10 What is your age group?

- ○ 18–24 years
- ○ 25–39 years
- ○ 40–60 years
- ○ 61 years or above

Appendix B. Codebook

Table B.9 lists the inter-coder agreement using Cohen's Kappa for each open-ended question.

Table B.9
Inter-coder Agreement using Cohen's Kappa κ for Each Open-ended Question

Question	Group	κ
B.3 Aspect Used to Judge the Author of Human-written Message	sexism	0.90
	racism	0.93
	abuse	0.87
B.6 Aspect Used to Judge the Author of AI-generated Message	sexism	0.80
	racism	0.98
	abuse	0.80
C.4 Agreement on Protecting Users Against AI-driven Cyberbullying	sexism	0.80
	racism	0.82
	abuse	0.93

Data availability

I have shared the link to my data in the manuscript.

References

Afane, K., Wei, W., Mao, Y., Farooq, J., & Chen, J. (2024). Next-generation phishing: How llm agents empower cyber attackers. In *2024 IEEE international conference on big data (Big-Data)* (pp. 2558–2567). IEEE.

Aizenkot, D. (2022). The predictability of routine activity theory for cyberbullying victimization among children and youth: Risk and protective factors. *Journal of Interpersonal Violence*.

Akter, T., Dosono, B., Ahmed, T., Kapadia, A., & Semaan, B. (2020). "I am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *USENIX security symposium (USENIX security)*.

Alexander, S. (2025). Deepfake cyberbullying: The psychological toll on students and institutional challenges of ai-driven harassment. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 98*(2), 36–50.

Alhaboby, Z. A., al Khateeb, H. M., Barnes, J., & Short, E. (2016). 'the language is disgusting and they refer to my disability': The cyberharassment of disabled people. *Disability & Society, 31*(8), 1138–1143.

Ali Talpur, B., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLoS One*.

Amazon mechanical turk — Access a global, on-demand, 24x7 workforce.(2024).

Anderson, M. (2018). *A majority of teens have experienced some form of cyberbullying*.

Ang, R. P., & Goh, D. H. (2010). Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry and Human Development, 41*, 387–397.

Arntfield, M. (2015). Toward a cybervictimology: Cyberbullying, routine activities theory, and the anti-sociality of social media. *Canadian Journal of Communication*.

Ayofe Azeez, N., Idiakose, S. O., Onyema, C. J., & Van Der Vyver, C. (2021). Cyberbullying detection in social networks: Artificial intelligence approach. *Journal of Cyber Security and Mobility*, 745–774.

Balakrishnan, V. (2015). Cyberbullying among young adults in Malaysia: The roles of gender, age and internet frequency. *Computers in Human Behavior, 46*, 149–157.

Balakrishnan, V., & Fernandez, T. (2018). Self-esteem, empathy and their impacts on cyberbullying among young adults. *Telematics and Informatics*.

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*.

Barlett, C., Chamberlin, K., & Witkower, Z. (2017). Predicting cyberbullying perpetration in emerging adults: A theoretical test of the barlett gentile cyberbullying model. *Aggressive Behavior, 43*(2), 147–154.

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*.

Bauman, S., & Newman, M. L. (2013). Testing assumptions about cyberbullying: Perceived distress associated with acts of conventional and cyber bullying. *Psychology of violence, 3*(1), 27.

Bethany, M., Seong, A., Henrique Silva, S., Beebe, N., Vishwamitra, N., & Najafirad, P. (2023). Towards targeted obfuscation of adversarial unsafe images using reconstruction and counterfactual super region attribution explainability. In *USENIX Security Symposium (USENIX Security 23)*.

Boccio, C. M., & Leal, W. E. (2023). The bully-victim overlap and vaping activity among adolescents. *Crime & Delinquency, 69*(8), 1489–1510.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*.

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative psychology, 9*(1), 3.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Campbell, M. (2012). How research findings can inform legislation and school policy on cyberbullying. In *Principles of cyberbullying research* (pp. 261–273). Routledge.

Catherine, D. M., & Higgins, G. E. (2019). Examining the effectiveness of academic scholarship on the fight against cyberbullying and cyberstalking. *American Journal of Criminal Justice, 44*, 645–655.

Chan, T. K. H., Cheung, C. M. K., Benbasat, I., Xiao, B., & Lee, Z. W. Y. (2023). Bystanders join in cyberbullying on social networking sites: The deindividuation and moral disengagement perspectives. *Information Systems Research*.

Chan, H. L., Kok, Y., Ong, J., & Yuvitasari, F. (2013). *Social cues & cyberbullying in Facebook: The effects of flaming messages, friend count and anonymity on cyberbullying behaviors*.

Chapell, M. S., Hasselman, S. L., Kitchin, T., Lomon, S. N., et al. (2006). Bullying in elementary school, high school, and college. *Adolescence, 41*(164), 633.

Choi, K.-S., Cho, S., & Lee, J. R. (2019). Impacts of online risky behaviors and cybersecurity management on cyberbullying and traditional bullying victimization among korean youth: Application of cyber-routine activities theory with latent class analysis. *Computers in Human Behavior*.

Chu, X., Yang, S., Sun, Z., Jiang, M., & Xie, R. (2022). The association between cyberbullying victimization and suicidal ideation among chinese college students: The parallel mediating roles of core self-evaluation and depression. *Frontiers in Psychiatry, 13*, Article 929679.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608.

Conway, L., Gomez-Garibello, C., Talwar, V., & Shariff, S. (2016). Face-to-face and online: An investigation of children's and adolescents' bullying behavior through the lens of moral emotions and judgments. *Journal of School Violence*.

Cross, D., & Walker, J. (2012). Using research to inform cyberbullying prevention and intervention. In *Principles of cyberbullying research* (pp. 274–293). Routledge.

Cyberbullying dataset.(2020).

Cyberbullying Research Center. (2024). Cyberbullying research center - How to identify, prevent, and respond. https://cyberbullying.org/. (Accessed 5 February 2024).

Debby Ng, E., Chua, J. Y. X., & Shorey, S. (2022). The effectiveness of educational interventions on traditional bullying and cyberbullying among adolescents: A systematic review and meta-analysis. *Trauma, Violence, & Abuse, 23*(1), 132–151.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., & De Bourdeaudhuij, I. (2012). Mobilizing bystanders of cyberbullying: An exploratory study into behavioural determinants of defending the victim. *Annual Review of Cybertherapy and Telemedicine*, 58–63, 2012.

Didden, R., Scholte, E. H., Korzilius, H., De Moor, J. M., Vermeulen, A., O'Reilly, M., & Lancioni, G. E. (2009). Cyberbullying among students with intellectual and developmental disability in special education settings. *Developmental Neurorehabilitation, 12*(3), 146–151.

Dordolo, N. (2014). The role of power imbalance in cyberbullying. *The Undergraduate Journal of Psychology, 3*, 35–41.

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 12763–12771.

First reports of children using ai to bully their peers using sexually explicit generated images. (2023). esafety commissioner says.

Finn, J. (2004). A survey of online harassment at a university campus. *Journal of Interpersonal Violence, 19*(4), 468–483.

Fleiss, J. L., Levin, B., & Cho Paik, M. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*.

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., & Holz, T. (2024). A representative study on human detection of artificially generated media across countries. In *IEEE symposium on security and privacy (SP)*.

Frik, A., Kim, J., Sanchez, J. R., & Ma, J. (2022). Users' expectations about and use of smartphone privacy and security settings. In *Proceedings of the 2022 CHI conference on human factors in computing systems*.

Gohal, G., Alqassim, A., Eltyeb, E., Ahmed, R., Hakami, B., Faqih, A. Al, Hakami, A., Qadri, A., & Mohamed, M. (2023). Prevalence and related risks of cyberbullying and its effects on adolescent. *BMC Psychiatry*.

Grace Brigham, N., Miranda, W., Kohno, T., & Redmiles, E. M. (2024). "violation of my body:" perceptions of ai-generated non-consensual (intimate) imagery. In *Proceedings of the 20th symposium on useable privacy and security (SOUPS)*. Philadelphia, PA: USENIX.

Gradinger, P., Strohmeier, D., & Spiel, C. (2009). Traditional bullying and cyberbullying: Identification of risk groups for adjustment problems. *Zeitschrift für Psychologie/Journal of Psychology, 217*(4), 205–213.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism, 19*(5), 595–610.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon mechanical turk. In *Proceedings of the CHI conference on human factors in computing systems*. CHI).

Hemphill, S. A., & Heerde, J. A. (2014). Adolescent predictors of young adult cyberbullying perpetration and victimization among Australian youth. *Journal of Adolescent Health, 55*(4), 580–587.

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior, 29*(2), 129–156.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research*.

Hinduja, S., & Patchin, J. W. (2019). Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence, 18*(3), 333–346.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297–307.

Instagram. (2024a). How to combat bullying and harassment on Instagram. https://help.instagram.com/464473649316860/?helpref=uf_share.

Instagram. (2024b). Instagram's commitment to bullying prevention—about Instagram. https://about.instagram.com/community/anti-bullying.

Ireland, L., Hawdon, J., Huang, B., & Peguero, A. (2020). Preconditions for guardianship interventions in cyberbullying: Incident interpretation, collective and automated efficacy, and relative popularity of bullies. *Computers in Human Behavior*.

Jacob, C. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), Article e2208839120.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the workshop on NLP and computational social science*.

Jian, Y.-S., Lin, K., Sun, I. Y., & Chen, S. (2025). Cyberbullying victim-offender overlap among Chinese college students: Comparing the predictive effects across criminological factors. *Victims and Offenders*, 1–22.

Kamar, E., Maimon, D., Weisburd, D., & Shabat, D. (2022). Parental guardianship and online sexual grooming of teenagers: A honeypot experiment. *Computers in Human Behavior*.

Karasavva, V., & Noorbhai, A. (2021). The real threat of deepfake pornography: A review of Canadian policy. *Cyberpsychology, Behavior, and Social Networking*.

Kim, Y., & Choi, J. S. (2021). Individual and organizational factors influencing workplace cyberbullying of nurses: A cross-sectional study. *Nursing and Health Sciences, 23*(3), 715–722.

Kırcaburun, K., Kokkinos, C. M., Demetrovics, Z., Király, O., Griffiths, M. D., ba, T., & Çolak, S. (2019). Problematic online behaviors among adolescents and emerging adults: Associations between cyberbullying perpetration, problematic social media use, and psychosocial factors. *International Journal of Mental Health and Addiction, 17*, 891–908.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in human behavior, 114*, 106553.

Koike, R., Kaneko, M., & Okazaki, N. (2024). Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI conference on artificial intelligence*.

Kolomeets, M., Tushkanova, O., Desnitsky, V., Vitkova, L., & Chechulin, A. (2024). Experimental evaluation: Can humans recognise social media bots? *Big Data and Cognitive Computing, 8*(3), 24.

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health, 41*(6), S22–S30.

Kumar Chaudhary, P., Yalamati, S., Ramesh Palakurti, N., Alam, N., Kolasani, S., & Whig, P. (2024). Detecting and preventing child cyberbullying using generative artificial intelligence. In *2024 Asia Pacific conference on innovation in technology (APCIT)* (pp. 1–5). IEEE.

Kwan, I., Kelly, D., Richardson, M., MacDowall, W., Burchett, H., Stansfield, C., Brunton, G., Sutcliffe, K., & Thomas, J. (2020). Cyberbullying and children and young people's mental health: A systematic map of systematic reviews. *Cyberpsychology, Behavior, and Social Networking, 23*(2), 72–82.

Kwan, G. C. E., & Skoric, M. M. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior, 29*(1), 16–25.

Lazuras, L., Barkoukis, V., & Tsorbatzoudis, H. (2017). Face-to-face bullying and cyberbullying in adolescents: Transcontextual effects and role overlap. *Technology in Society, 48*, 97–101.

Lee, J. M., Choi, H. H., Lee, H., Park, J., & Lee, J. (2023). The impact of cyberbullying victimization on psychosocial behaviors among college students during the covid-19 pandemic: The indirect effect of a sense of purpose in life. *Journal of Aggression, Maltreatment & Trauma, 32*(9), 1254–1270.

Li, Q. (2006). Cyberbullying in schools: A research of gender differences. *School Psychology International, 27*(2), 157–170.

Li, Y., Vishwamitra, N., Knijnenburg, B. P., Hu, H., & Kelly, C. (2017). Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of ai-generated messages. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100058.

Lin, Y., He, P., Xu, H., Xing, Y., Yamada, M., Liu, H., & Tang, J. (2024b). Towards understanding jailbreak attacks in llms: A representation space analysis. In *Proceedings of the 2024 conference on empirical methods in natural language processing (EMNLP)* (pp. 7067–7085).

Lin, K., Zhou, Y., Xu, B., & Chang, L. Y. C. (2024a). A latent class analysis of online victim-offender overlap among Chinese youth: Examining overlap risks across online deviance types. *Crime & Delinquency*, Article 00111287241266589.

Lindsay, M., Booth, J. M., Messing, J. T., & Thaller, J. (2016). Experiences of online harassment among emerging adults: Emotional reactions and the mediating role of fear. *Journal of Interpersonal Violence, 31*(19), 3174–3195.

Lu, Z., Huang, Di, Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2024). Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems*.

López-Castro, L., & Priegue, D. (2019). Influence of family variables on cyberbullying perpetration and victimization: A systematic literature review. *Social Sciences, 8*(3), 98.

Macaulay, P. J. R., Betts, L. R., Stiller, J., & Kellezi, B. (2022). Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. *Computers in Human Behavior, 131*, Article 107238.

Manuel Gámez-Guadix and Estibaliz Mateos-Pérez. (2019). Longitudinal and reciprocal relationships between sexting, online sexual solicitations, and cyberbullying among minors. *Computers in Human Behavior*.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*.

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. In *Women's studies international forum, ume 47* pp. 46–55). Elsevier.

Menesini, E., & Nocentini, A. (2009). Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift für Psychologie/Journal of Psychology*.

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

Mishna, F., Khoury-Kassabri, M., Gadalla, T., & Daciuk, J. (2012). Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims. *Children and Youth Services Review, 34*(1), 63–70.

Monks, C. P., Robinson, S., & Worlidge, P. (2012). The emergence of cyberbullying: A survey of primary school pupils' perceptions and experiences. *School Psychology International*.

Murnion, S., Buchanan, W. J., Smales, A., & Russell, G. (2018). Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*.

National Domestic Violence Hotline. Domestic violence support — National domestic violence hotline. https://www.thehotline.org/. (Accessed 5 February 2024).

Navarro, J. N., & Jasinski, J. L. (2013). *Why girls? Using routine activities theory to predict cyberbullying experiences between girls and boys*. Women & Criminal Justice.

Nikola, B. (2023). *Latest cyberbullying statistics – The extent of cyberbullying in 2024*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), Article 1609406917733847.

Number and statistics guide (7th ed.). (2025). APA Style.

Olweus, D. (1993). *Bullying at school: What we know and what we can Do*. Oxford, UK: Blackwell Publishing.

Pabian, S., & Vandebosch, H. (2021). Perceived long-term outcomes of early traditional and cyberbullying victimization among emerging adults. *Journal of Youth Studies*.

Patchin, W. (2024). In *Cyberbullying data*, 2023.

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice, 4*(2), 148–169.

Patchin, J. W., & Hinduja, S. (2015). Measuring cyberbullying: Implications for research. *Aggression and Violent Behavior*.

Patterson, V. C., Closson, L. M., & Patry, M. W. (2019). Legislation awareness, cyberbullying behaviours, and cyber-roles in emerging adults. *Canadian Journal of Behavioural Science/Revue Canadienne des sciences du comportement, 51*(1), 12.

Pechenik Gieseke, A. (2020). The new weapon of choice": Law's current inability to properly address deepfake pornography. *Vand. L. Rev.*

Peebles, E. (2014). Cyberbullying: Hiding behind the screen. *Paediatrics & child health*.

Perren, S., & Gutzwiller-Helfenfinger, E. (2012). Cyberbullying and traditional bullying in adolescence: Differential roles of moral disengagement, moral emotions, and moral values. *European Journal of Developmental Psychology*.

RAINN. (2024). Resources for survivors of stalking and cyberstalking. https://rainn.org/news/resources-survivors-stalking-and-cyberstalking. (Accessed 5 February 2024).

Rebecca, P. A. (2015). Adolescent cyberbullying: A review of characteristics, prevention and intervention strategies. *Aggression and Violent Behavior*.

Saha Roy, S., Thota, P., Naragam, K. V., & Nilizadeh, S. (2024). From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)* (pp. 36–54). IEEE.

Salawu, S., Jo, L., & He, Y. (2021). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *The workshop on online abuse and harms*.

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). "call me sexist, but…": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*.

Schodt, K. B., Quiroz, S. I., Wheeler, B., Hall, D. L., & Silva, Y. N. (2021). Cyberbullying and mental health in adults: The moderating role of social media use and gender. *Frontiers in Psychiatry, 12*, Article 674298.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147–154.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376–385.

Snapchat. (2024). Snapchat safeguards for teens — Parent's guide to Snapchat. https://parents.snapchat.com/safeguards-for-teens.

StopBullying.gov. Get help now — Stopbullying.gov. https://www.stopbullying.gov/resources/get-help-now, (2022)–. (Accessed 5 February 2024).

Thazin Khine, A., Saw, Y. M., Ye Htut, Z., Khaing, C. T., Zaw Soe, H., Kyu Swe, K., Thike, T., Htet, H., Nandar Saw, T., Cho, S. M., et al. (2020). Assessing risk factors and impact of cyberbullying victimization among university students in Myanmar: A cross-sectional study. *PLoS One, 15*(1), Article e0227051.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and finetuned chat models. *arXiv preprint arXiv:2307.09288*.

Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day.(2016).

Vishwamitra, N., Hu, H., Luo, F., & Cheng, L. (2021). Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*.

Wang, C.-W., Masika Musumari, P., Techasrivichien, T., Pilar Suguimoto, S., Tateyama, Y., Chan, C.-C., Ono-Kihara, M., Kihara, M., & Nakayama, T. (2019a). Overlap of traditional bullying and cyberbullying and correlates of bullying among taiwanese adolescents: A cross-sectional study. *BMC Public Health, 19*, 1–14.

Wang, M.-J., Yogeeswaran, K., Andrews, N. P., Hawi, D. R., & Sibley, C. G. (2019b). How common is cyberbullying among adults? Exploring gender, ethnic, and age differences in the prevalence of cyberbullying. *Cyberpsychology, Behavior, and Social Networking, 22*(11), 736–741.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*.

Weulen Kranenbarg, M., Holt, T. J., & Jean-Louis Van Gelder. (2019). Offending and victimization in the digital age: Comparing correlates of cybercrime and traditional offending-only, victimization-only and the victimization-offending overlap. *Deviant Behavior, 40*(1), 40–55.

Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. *Journal of Adolescent Health, 41*(6), S14–S21.

Wong, R. Y., Cheung, C. M., & Xiao, B. (2018). Does gender matter in cyberbullying perpetration? An empirical investigation. *Computers in Human Behavior, 79*, 247–257.

Xu, B., & Tu, X. (2024). The formation mechanism of adolescents' cyber violence. *Youth Study, 43*(2), 27–39.

Yang, A. (2024). Lawsuit claims character.ai is responsible for teen's suicide. https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791. (Accessed 23 October 2024).

Ybarra, M. L., Mitchell, K. J., Wolak, J., & Finkelhor, D. (2006). Examining characteristics and associated distress related to internet harassment: Findings from the second youth internet safety survey. *Pediatrics, 118*(4), e1169–e1177.

Yeop Paek, S., Lee, J., & Choi, Y.-J. (2022). The impact of parental monitoring on cyberbullying victimization in the covid-19 era. *Social Science Quarterly*.

Yubero, S., Navarro, R., Elche, M., Larrañaga, E., & Ovejero, A. (2017). Cyberbullying victimization in higher education: An exploratory analysis of its association with social and emotional factors among Spanish students. *Computers in Human Behavior, 75*, 439–449.

Zhang, S., Leidner, D., Cao, X., & Liu, N. (2022). Workplace cyberbullying: A criminological and routine activity perspective. *Journal of Information Technology*.

Zhao, Y., Chu, X., & Rong, K. (2023). Cyberbullying experience and bystander behavior in cyberbullying incidents: The serial mediating roles of perceived incident severity and empathy. *Computers in Human Behavior, 138*, Article 107484.

Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in Public Health, 9*, Article 634909.