photo: Martin Kormaňák

Zhu Luo

An AI-Media Literacy Competency Framework for AI-Augmented Adolescent Cyberbullying: A Scoping Review

ABSTRACT

Artificial intelligence is altering how cyberbullying manifests among adolescents. Generative tools now make it much easier to create fake photos, cloned voices, and fabricated screenshots, reducing the cost of impersonation and deception. At the same time, ranking and recommendation systems on digital platforms can exacerbate the harm when they promote embarrassing content to larger, more widespread audiences and keep harmful content visible for longer periods. Those developments make it more difficult to trace blame and protect usable evidence, and to act quickly. But many of today's prevention efforts and media literacy efforts continue to focus on AI-related risks in a narrow sense. To address this gap, the study conducted a PRISMA-ScR-guided scoping review of multidisciplinary studies published between 2018 and 2025. Following deduplication, 2,249 records were reviewed and 58 were added to the final synthesis. This paper clusters the evidence into four emergent AI-enabled threats: fabrication and impersonation, amplification driven by visibility dynamics, automation at scale, and gaps in governance and adjudication. Based on this evidence map and a five-stage overview of cyberbullying, the paper develops an AI-Media Literacy Competency Framework that establishes links between the threat areas and teachable competencies, stakeholder responsibilities, and actionable response plans. The framework establishes five teachable domains: synthetic-media verification, platform and algorithm knowledge, privacy and identity preservation, reporting and redress, and prosocial bystander behavior. Rather than collecting new human-subject data, this paper synthesizes the existing evidence into a usable, AI-informed framework for prevention and response in contexts of uncertain authenticity, algorithmically amplified visibility, and continuous cross-platform recirculation.

KEY WORDS

Adolescents. Artificial Intelligence Literacy. Cyberbullying. Deepfakes. Generative AI. Media Literacy. Recommender Systems.

https://doi.org/10.34135/mlar-26-01-01

1 Introduction

Cyberbullying in adolescents persists as a growing concern for schools, families, and policymakers. Smith et al. (2008) define cyberbullying as intentional aggression carried out through electronic communication, typically involving repeated attacks and power imbalances. Later studies have also explored comparable criteria within adolescents (see Zhang et al., 2022). The harm done online can be much more rapid and long-term in visibility than bullying done in person. Networked media can spread the information easily, be widely shareable, readily replicated, and reach a huge audience (Dailey et al., 2025). It can be even more devastating in adolescence when peer acceptance and social standing are critical factors. Shaming in the public domain and harm to reputation is therefore particularly damaging at this age (Silk et al., 2024). Previous research has linked engagement in cyberbullying to depression, anxiety, social withdrawal, and decreased school attachment (Kasturiratna et al., 2025).

Over the past few years, AI has been changing the nature of cyberbullying in ways that reach beyond simply enabling more activity online. Generative tools enable the creation of persuasive and malicious fake content; they enhance the impersonation of others. Meanwhile, recommendation and ranking systems can amplify harmful content, exposing victims again and again to public humiliation in front of a wider audience. Automation might aggravate these abuses by magnifying the scale and duration of harassment, while also making it more difficult to attribute accountability, as well as making responsibility more opaque. Such changes are important because they make it more difficult to determine what constitutes usable evidence, shift the onus of monitoring and verification to targets and schools and offer new challenges for monitoring, moderation and redress (Milosevic et al., 2022).

This article asserts that cyberbullying prevention should be adapted to AI. Practical steps can help students, schools, families, and platforms to mitigate harm and respond in ways that are more effective when authenticity is in doubt, visibility is determined by algorithms and harmful content in social media continues to propagate. To address this, the study conducts a scoping review of multidisciplinary literature and high-quality public reports from 2018 – 2025 to examine AI mechanisms and to develop recommendations for intervention. Arksey and O'Malley (2005) argue that scoping reviews are helpful, especially when sources and perspectives are broad and methodologically heterogeneous and recommendations after such review reaffirm the utility of scoping reviews to guide the exploration of areas of interest (Peters et al., 2021). Based on this map of evidence in its entirety, the paper presents an AI-Media Literacy Competency Framework for mitigating and effectively responding to the practice of adolescent cyberbullying. Since this study relies on prior literature, instead of new human-subject data, curriculum development and school-level response planning are informed by it.

1.1 Why AI Changes the Cyberbullying Problem

AI is reshaping cyberbullying by changing the way in which harmful content is created, perceived, circulated, and managed online. One major shift is that AI makes it much easier to create harmful material and impersonate others. Images, audio clips, videos, and text can all be generated or manipulated in a way to make them look credible to peers, known adults, and other readers. This creates new opportunities for harassment through fabricated or manipulated "evidence", including fake screenshots, cloned voice messages, edited photos, and non-consensual sexualized content presented as authentic (Umbach et al., 2024). This harm does not disappear when the validity of something like such material is called into question. More often than not, the damage done to reputation is through visibility, not truth (Ecker et al., 2022).

In many cases, harm is intensified not only by the content itself, but by how widely it is circulated and seen. AI can also shape how harmful content spreads. Today's teens consume online material through algorithmically curated feeds where engagement is frequently the measure for visibility. In an environment like that, sensational, embarrassing or humiliating posts can be easily shared, followed up on, or remixed, causing users to continue to experience the same harm (Costello et al., 2024). As a result, cyberbullying is no longer just a one-to-one face-to-face contact between attacker and victim. It has become more public and collective, with audience size, repetition, and visible social endorsement shaping both the severity and duration of harm (Chan et al., 2023).

Another concern is scale. AI-assisted writing tools, automated messaging, can also increase both the quantity and coordination of abusive communication. For schools, which are frequently underresourced, and for parents and guardians of children and young people, that makes it more difficult to intervene. With so many such messages to choose from, it is much harder to know whether a case involves a single individual, a collective or a larger network of messages, and intimidation can escalate (Cai et al., 2023).

At the same time, AI creates new governance- and decision-related issues. Content moderation and automated detection are still fledgling and reporting systems can be slow, nebulous or not relevant to younger users. Re-uploaded (and sometimes duplicate) synthetic or altered content can spread on different accounts well beyond what was posted first. In many cases, schools do not have clear procedures for preserving evidence, protecting privacy and due process, or coordinating responses with families and platforms (Milosevic et al., 2022; eSafety Commissioner, 2025a). In that light, prevention does not depend on one-size-fits-all warnings about bullying. It also requires practical skills: recognizing risk early, documenting harm, filing reports accurately, and knowing when and where to seek help.

1.2 Why Media Literacy Is the Right Lever – Yet Currently Under-Specified

In general terms, media and information literacy includes both a critical evaluation of messages and sources of information, as well as responsible engagement with online spaces (MediaNet, 2025). However, AI-augmented cyberbullying suggests that these more traditional methods of digital literacy are no match for it.

The trouble is students now have to confront a much tougher question of authenticity. In most cases, they are no longer solely evaluating whether the information is generally true or misleading. They also might have to establish whether a super personal image, message or recording was deliberately and artfully manufactured or manipulated to target a specific person. But it is not just about proving the identity, consent, privacy and emotional security, these move to the forefront in such cases (NIST, 2024).

The second problem is that online visibility is increasingly conditioned by platform algorithms. What seems popular, trustworthy or routine online may have an effect, depending on how feeds rank and circulate information. Therefore, preventing cyberbullying may no longer be just based on usual, source-based criticism. It also has to bear in mind the role of ranking signals, engagement cues, and social approval as a whole, in influencing perceptions and additional violence (Tayie, 2025).

A third problem is that successful resolution tends to be procedural, not knowledge-based. In practice, outcomes often depend on whether victims or bystanders can preserve records, retain evidence, access reporting channels, and seek help. But majority of current programs still follow broad advice, such as "block", "report" or "tell an adult", but they do not translate this into practical methods for teaching children and teens how to deal with messages they receive online (Lan et al., 2022).

Later attempts are still focusing on social norms, digital citizenship and assessing the reliability of online messages. Those are still relevant, but AI-augmented cyberbullying is bringing new functionality challenges that these models fail to fully address: for instance, fabricated evidence, identity misuse, algorithmic amplification, automated, large-scale harassment, and opaque accountability across platforms. Generalizable AI literacy frameworks are helpful, but they are seldom dedicated to examining how school-based incidents play out in reality. Thus, they pay little heed to the procedural skills needed, such as evidence preservation, reporting and redress, privacy protection, due process and who carries the responsibilities, including students, educators, families and platforms. Existing frameworks do not sufficiently address these procedural and school-based challenges; this paper therefore seeks to address that gap by proposing a more structured framework.

1.3 Aim, Research Questions, and Contributions

This article explores how AI-enabled mechanisms are affecting adolescent cyberbullying and converts that evidence into actionable competencies for adolescents and schools. These discussions have been supported by interdisciplinary literature and rigorous public reports published between 2018 and 2025 regarding school-associated trends including synthetic media, ranking and recommendation systems, automation at scale, and persistent shortcomings in detection and moderation. Based on these insights, the paper offers an AI–Media Literacy Competency Framework which connects dynamic risk patterns with competency clusters, learning outcomes, and actions undertaken by actors (i.e., students, teachers, parents or guardians, schools, and platforms) to reduce such changing risks. The research questions put forward during the review are:

RQ1: Which AI-enabled mechanisms are most commonly associated with adolescent cyberbullying and other youth harassment?

RQ2: In which ways do these mechanisms reconfigure key processes of cyberbullying such as production, circulation, interpretation, escalation, and recovery, especially in school settings?

RQ3: Which competencies and intervention points appear in, or are implied by, the literature, and how can they be organized into an actionable AI–media literacy framework?

This paper makes two primary contributions. First, it provides a cross-disciplinary overview of changing AI-related adolescent cyberbullying and related youth harassment. Second, it provides a practice-based AI-media literacy competency matrix, correlating emergent risks with lessons learned and shared responsibilities. This study thus assists curriculum design and the improvement of schools' response protocols without gaining new human-subject data.

2 Conceptual Background and Key Definitions

The key concepts discussed in the paper are clarified in this section, and insight regarding the analytical framework applied to link AI-enabled mechanisms to cyberbullying processes and media literacy interventions is also described. Rather than provide a comprehensive review of the literature, it aims to set up a common set of terms and a process-based framework for the analysis that follows.

2.1 Adolescent Cyberbullying as a Process

Cyberbullying among adolescents in the context of electronic communications is typically considered intentional harm, most often involving repetition or an imbalance of power (Zhang et al., 2022). In contrast, this article conceptualizes cyberbullying as not one post or single

event, but as a continuing process. This distinction matters because the scale of harm is often determined by how content circulates over time among these peer networks and platforms. Networked media, as Boyd (2010) points out, can render damaging content persistent and at times especially visible, but research later pointed out that audience participation and platform conditions also contribute to this (Macaulay et al., 2022). The analysis uses a five-phase method to investigate: production; circulation; interpretation; escalation; and recovery. Production is making harmful content or abusive behavior. Circulation is when harmful content spreads by sharing, reposting, and exposure on the platform. Interpretation focuses on how people judge credibility as well as their role in participating, ignoring, or intervening. Escalation involves repetition, expanding audiences, and the possibility that online harm may spill over into offline spaces. Recovery consists of reporting, adjudication, support and the longer-term handling of reputational damage (Chew et al., 2025). This process-based lens is especially useful in AI-embedded contexts because AI can shape every stage of the process, from case creation to evaluation and resolution (Jaidka et al., 2025).

2.2 What AI-Augmented Means in this Study

In this paper, AI-augmented cyberbullying encompasses types of cyberbullying perpetuated, accelerated, or otherwise transformed under the influence of AI systems and AI-mediated platform infrastructures. In this paper, the term AI is used broadly to refer not only to generative tools but also to the platform-level systems that shape how content is distributed and managed. On the one hand, generative tools make it easier to create, modify, or replicate content, including 'evidence' that appears plausible but is false for a particular audience (NIST, 2024). Conversely, platform infrastructures such as ranking and recommendation mechanisms, distribution mechanisms, moderation processes, and reporting and redress systems configure how harmful content appears, is retrieved, is maintained, and is addressed (Milosevic et al., 2022). Framed this way, AI is not one-sidedly described as a causal force. It is rather the interplay of emergent affordances and infrastructures within the context of peers' culture, platform incentives, and the differing degrees of power given by schools to respond (Shelby et al., 2023)

2.3 Four AI-Enabled Threat Vectors

This review organizes AI-related mechanisms into four overlapping threat vectors to provide a practical working typology to relate it more directly to educational responses. The first is fabrication and impersonation, which uses synthetic media and generative editing to create false "proof", misuse the identity of others, and stage events that did not actually occur (Diel et al., 2025). The second category is amplification and visibility dynamics, which can be used to illustrate the way ranking and recommendation systems can spread and amplify harmful content, and subject individuals to repeated encounters, potentially exacerbating harm through visibility that is platform-driven (Office of the Surgeon General, 2023). Third, automation and scale also imply that harassment may be more frequent, persistent, and coordinated. And the volume can increase intimidation and further put pressure on school systems responding to harassment (Wei et al., 2023). The fourth one, governance and adjudication gaps, pertain to problems with detection, moderation, reporting and institutional response particularly when the material is synthetic, repeatedly remixed or disseminated across platforms (Alexander et al., 2022).

2.4 Linking Threat Vectors to Media Literacy and Intervention Logic

Research has established that media and information literacy is a practical capability as opposed to merely the ability generally, to access, analyze, evaluate, create, and respond responsibly to media messages (NAMLE, 2024). For AI-augmented cyberbullying, that is the shift from analysis to action. When authenticity of communication is suspect, students and schools should know to discern signs of manipulation, make deliberate decisions, refrain from impulsive sharing, protect the privacy and identity of participants, preserve evidence safely, and utilize reporting channels effectively (Freed et al., 2025). This is why media and information literacy entails more than cognitive abilities; it is also crucial to ethical judgment, socio-emotional awareness, and procedural know-how (Vuorikari et al., 2022). Existing frameworks of AI literacy are useful starting points (Miao & Cukurova, 2024), but school-based incidents warrant further emphasis of applying these principles to competencies that reflect the case situations in a meaningful manner.

The synthesis model of this review is anchored in this concern. It weaves together mechanisms of AI, processes of cyberbullying, and possible reactions. AI-facilitated mechanisms may affect incidents across multiple phases, including production, circulation, interpretation, escalation, and recovery. Thus, all relationships between mechanisms and stages can be interpreted as leading to potential intervention points and teachable competencies. These are then mapped onto the stakeholders most directly involved, including students, teachers, parents or guardians, schools, and, where relevant, platforms. In charting the evidence, the review records the mechanism identified in each source, the phase or phases it affects, and any intervention or competency implications that are explicitly discussed or reasonably inferred from the findings (Pollock et al., 2023).

3 Methodology: Scoping Review Protocol

This study applies a scoping review to examine how AI-related mechanisms are framed within adolescent cyberbullying and closely related forms of youth online harassment, and to synthesize that evidence into a practice-driven AI-media literacy competency framework. This makes a scoping review especially appropriate, given the continually evolving body of knowledge surrounding concepts, terminology, and evidence. Arksey and O'Malley (2005) note that this approach can be particularly valuable when the field is emergent rather than settled. It is also well suited to research areas marked by varied terminology and diverse study designs, where the goal is to map patterns in the literature, identify clusters of evidence, and highlight gaps for future work, rather than to calculate a single pooled effect size (Munn et al., 2022).

3.1 Review Objectives and Questions

The review followed the research questions introduced in Section 1.3.

3.2 Eligibility Criteria

Eligibility criteria were established in advance and applied consistently throughout screening in order to keep the review focused on adolescent safety and educational relevance. Records were included if they examined adolescents or youth, understood broadly as school-age populations. Studies using the term "youth" were retained when the participants were mainly minors. When records referred only to "students" without giving clear age information, they were included only if the setting was explicitly K-12 or otherwise clearly school-based (Sorrentino et al., 2023).

The review focused on cyberbullying and closely related forms of youth online harassment directed at individuals or small peer groups. This included behaviors such as repeated humiliation, intimidation, exclusion, impersonation, and reputational attacks carried out through digital communication (Tokunaga, 2010). Conversely, records addressing only general hate speech or ideological extremism were excluded unless they provided evidence relevant to AI-enabled amplification or automated abuse in youth settings (National Academies of Sciences, Engineering, and Medicine, 2024).

To be considered, a record likewise had to cover at least one AI-related mechanism which could plausibly alter the pattern of bullying or harassment. Relevant mechanisms were synthetic media and deepfakes; generative AI usage for fabrication, impersonation, or message production; ranking and recommendation systems used to define visibility; algorithmic systems that automate or orchestrate harassment; and governance activities (e.g., detection, moderation, reporting, and redress) (Prem & Krenn, 2023). Records that used only "technology" or "social media" more generally, and did not detail a particular mechanism related to AI, were excluded. This review focused on studies from environments relevant to schools, adolescent peer groups, or youth-utilized platforms. Adult-only samples were excluded from analysis unless they provided essential evidence that is clearly established and that could potentially be readily transferred to youth cyberbullying; if established, the transfer was documented in the synthesis. Studies focused on intimate partner violence or on workplace harassment were disregarded (Chicote-Beato et al., 2024).

Sources that met criteria included peer-reviewed empirical studies and peer-reviewed conceptual papers with well-described mechanisms. Furthermore, reports were eligible if they were high-quality public reports that contained evidence and an analysis directly relevant to schools and youth. Report quality for this review was evaluated on four dimensions – authority, transparency, verifiability, and relevance. Reports were also filtered out if they were largely opinion articles, press releases, vendor marketing materials, or web-based content that was not high on credibility (Duma et al., 2023).

To account for the era when deepfakes, recommender-system governance controversies, and contemporary generative AI became especially prevalent, the review covered records published between January 2018 and December 2025. Given the scope of the available material and the constraints of the journal, only English-language records were included (Helbach et al., 2022).

3.3 Information Sources and Search Strategy

The searches were conducted on 26 January 2026 to capture literature from education, psychology, communication, and information science. The bibliographic databases included ERIC, PsycINFO, Scopus, and Web of Science Core Collection. Likewise, Google Scholar was used as an additional source in this search period. We conducted targeted searches in repository report sources and the websites of respected organizations focused on youth online safety and AI (intergovernmental institutions, policy research institutes, and non-profit safety organizations). The review also relied on backward reference checking and forward citation tracking for further relevant literature.

The search strategy was organized around three broad concept categories: youth/adolescent-related terms, terms related to cyberbullying and online harassment, and AI-related mechanisms. The AI was intentionally broad to cover variation in terminology in different areas and sources. The term AI-level descriptions were constructed to encapsulate issues identified across the literature, which involved synthetic media and deepfakes, generative AI and AI-assisted content creation, impersonation, recommender systems, algorithmic amplification, and automated moderation. To each database the search syntax was modified and controlled vocabulary where applicable used. Complete search strategies were subsequently recorded. One example of the

Boolean search structure used in the review was: (adolescen* OR teen* OR youth OR student*) AND (cyberbull* OR "online harassment" OR "digital harassment" OR "online abuse" OR "peer harassment") AND (deepfake* OR "synthetic media" OR "generative AI" OR "AI-generated" OR "algorithm* recommendation" OR ranking OR "algorithmic amplification" OR "automated moderation" OR "content moderation").

3.4 Screening and Selection Process

The retrieved records were exported to a reference management program for deduplication and screened in two stages. Title and abstract screening comprised the first stage, during which clearly irrelevant records were removed according to the predefined eligibility criteria. The second stage was full-text review for final inclusion. Reasons for exclusion at the full-text stage were recorded, and the overall process is summarized in a PRISMA-style flow diagram covering identification, screening, eligibility assessment, and final inclusion (Page et al., 2021).

Several steps were built into the screening process to improve consistency and reduce the risk of selection bias. Before formal screening began, the decision rules were tested on an initial set of records so that recurring ambiguities could be addressed in advance, including cases where sources referred to general "technology" rather than a clearly identifiable AI-related mechanism, or where school relevance was not immediately clear. During title and abstract screening, a random 20 percent subset of the records, approximately 450 out of 2,249, was screened again in a second pass after a washout interval. Any differences between the first and second pass were resolved through documented refinements to the screening rules. Because full-text screening carried greater consequences for inclusion, all retained full texts reviewed at that stage (n=177) were checked again in a second pass. Particular attention was paid to borderline cases and to the consistent coding of exclusion reasons. When refinements to the rules affected earlier decisions, previously screened records were revisited to preserve internal consistency across the review (Hadie, 2024).

3.5 Data Charting and Coding Scheme

A standardized data-charting form was developed and piloted on an initial sample of the included records. The form consisted of basic bibliographic parameters like disciplinary field, study type, research method, population and setting, platform or media context, and the AI-related mechanism or mechanisms described in the source.

The identified records were then analyzed using three interconnected coding dimensions. First, AI-based mechanisms were organized based on the four threat vectors enumerated in Section 2. Second, records were coded by a specific stage or stages where they addressed the cyberbullying process, utilizing the five-phase process lens. Third, the review tracked intervention implications. These included not only recommendations put forward directly by the source, but also implications inferred when they pointed clearly to a competency gap or a procedural obstacle. To maintain analytic transparency, all inferred items were marked as such (Pollock et al., 2023).

Alongside this, the charting process tracked what authors referred to as literacy-related concepts: media literacy, digital literacy, AI literacy, digital citizenship, bystander intervention, resilience, and help-seeking. Where applicable, it also noted any educational, institutional, or policy interventions found in the source.

3.6 Synthesis Approach

The synthesis was carried out in two phases. At the first stage, an evidence map was developed to indicate the distribution of included records by year of publication, disciplinary field, research method, setting, AI-related mechanism, and reported outcomes or key claims. The second phase employed an integrative qualitative method following the four threat vectors. For each vector, the review identified key mechanisms already covered in the literature, the phases of the cyberbullying process most affected, themes that emerged as common gaps or tensions across sources, and specific implications relevant to the school-based context.

The resulting framework was developed from this synthesis by linking each threat vector to corresponding competency clusters, learning outcomes, and stakeholder actions. The framework was designed to ensure that practical, on-the-ground skills remain directly aligned with the tasks and responses most likely to arise in incident situations. As Biggs (1996) suggests, learning outcomes are most beneficial when they are meaningfully linked to action, and the framework was developed based on this assumption so that each competency remains linked to patterns seen in the evidence reviewed (Pollock et al., 2023).

3.7 Ethical Considerations

This review did not include new human-subject data and no identifiable material involving minors was reproduced. Instead, it relied on previously published scholarly research and reputable public reports. This secondary synthesis generally does not require formal ethical approval. According to Suri (2019), the review followed the principles of harm-minimisation by avoiding identifiable cases, graphic material, and unnecessary information related to specific incidents.

3.8 Reporting and Transparency

To enhance transparency, the review contains a description of the search methods employed, inclusion criteria, screening process, and data-charting categories in the manuscript and is augmented by supplementary materials when available. The outcomes of the review are threefold: a PRISMA-style flow diagram, an evidence map table, and a threat-vector-by-competency matrix linking the findings to implications for education and governance.

4 Results: Evidence Map

Here, we describe the evidence base in descriptive terms. It describes the research scope, explains how included studies were separated into the disciplines and research methods, and explains the main mechanisms related to artificial intelligence that are found to be often associated with adolescent cyberbullying and closely related youth online harassment.

4.1 Search and Selection Outcomes

During the search process, 3,215 records were retrieved from the identified databases and supplementary sources. After removing duplicates, there were 2,249 unique records left for title and abstract screening. Of these, 177 records were retained for full-text review; 58 met all eligibility criteria and were included in the final synthesis. Figure 1 presents the PRISMA-style flow diagram, and the reasons for full-text exclusion are reported alongside it. This format is

widely recommended in order to make the screening decisions transparent and easy to follow (Tricco et al., 2018).

At the full-text stage, the most frequent reason for exclusion was a mismatch in AI-related mechanism. Most records mentioned social media or online harms in general terms, but no specific AI-related mechanism relevant to the scope of this review was identified. Population and phenomenon mismatches were also common. These exclusions are reported in aggregate in Figure 1 and summarized further in the Methods supplement.

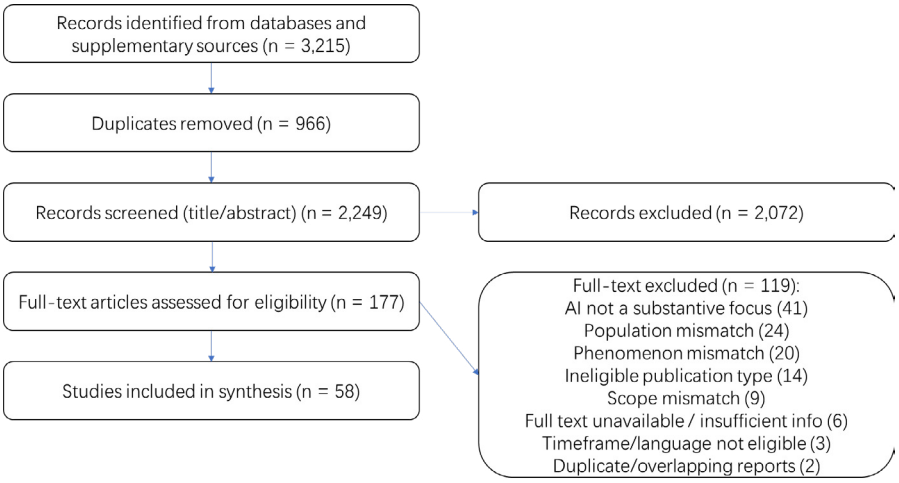

Figure 1: *PRISMA-style flow diagram of study selection*

Source: own processing, 2026

4.2 Characteristics of the Included Evidence Base

The included records span 2018 to 2025, and most of the literature appears in the later years of that period. Table 1 shows that 44 of the included records were published between 2022 and 2025, with the highest annual counts in 2023 and 2024. Indeed, this pattern echoes the escalating attention paid to synthetic media risks and today's generative AI, both in policy debates and school safety conversations. Early work like Chesney and Citron (2019) helped to highlight issues relating to deepfakes, and the recent youth safety reporting indicates that this attention has been growing (Raghuvanshi et al., 2024).

The corpus is rich in peer-reviewed empirical studies, peer-reviewed conceptual or theoretical papers with explicit mechanisms, and relatively few high-quality reports (public or organizational) with regards to issues of youth and schools. The empirical studies are mixed-method, quantitative, and qualitative designs. These all point to the fact that the field is not monolithic in terms of research focus.

The literature also reflects a broad cross-disciplinary approach that includes both communication and media studies, education, psychology and behavioral science, information science and computing, and interdisciplinary studies on youth safety and policy. But that distribution is uneven. Some aspects of the literature touch psychosocial outcomes and coping, while others seem more concerned with platform governance, detection, and policy responses. That unevenness highlights the need for a later synthesis that bridges these strands and links evidence of mechanisms to competence needs in the field.

Also, the context in which the evidence is presented is heterogeneous. 20 studies focus on the school-based setting, 25 youth-facing platforms, and 13 mixed or poorly defined settings. Platform focus is broad as well. Eighteen records deal with social media in higher-level terms

without specifying a platform type, fifteen social networking or messaging mediums, and ten short-video or livestream. Narrower groups of research look at gaming environments, community forums, and several platform types.

Characteristic	Category	N	%
Publication year	2018	2	3.4
Publication year	2019	3	5.2
Publication year	2020	4	6.9
Publication year	2021	5	8.6
Publication year	2022	8	13.8
Publication year	2023	12	20.7
Publication year	2024	14	24.1
Publication year	2025	10	17.2
Evidence type	Peer-reviewed empirical study	33	56.9
Evidence type	Peer-reviewed conceptual or theoretical paper	18	31.0
Evidence type	High-quality public or organizational report	7	12.1
Study approach	Quantitative	14	24.1
Study approach	Qualitative	12	20.7
Study approach	Mixed methods	7	12.1
Study approach	Non-empirical (conceptual or report)	25	43.1
Primary disciplinary area	Communication and media studies	14	24.1
Primary disciplinary area	Education	10	17.2
Primary disciplinary area	Psychology or behavioral science	12	20.7
Primary disciplinary area	Information science or computing	13	22.4
Primary disciplinary area	Interdisciplinary youth safety or policy	9	15.5
Primary setting	School-based context	20	34.5
Primary setting	Youth-facing platform context	25	43.1
Primary setting	Mixed or not clearly specified	13	22.4
Primary platform focus	Platform not specified or general social media	18	31.0
Primary platform focus	Social networking or messaging services	15	25.9
Primary platform focus	Short-video or livestream platforms	10	17.2
Primary platform focus	Gaming or community forums	6	10.3
Primary platform focus	Multiple platform types	9	15.5

Table 1: *Characteristics of included records (N = 58)*

Source: own processing, 2026

4.3 AI Mechanisms Addressed in the Literature

The records used can be generally categorized into the four AI-enabled threat vectors used in this study. The most frequent ones are fabrication and impersonation (see Table 2), followed by amplification and visibility dynamics, governance and adjudication gaps, and automation and scale. The coding allows a single record to be assigned to more than one threat vector, so the totals are not expected to match the overall number of included studies, and the percentages may add up to more than 100 percent.

AI-enabled threat vector	n (records coded to vector)	% of included records
Fabrication and impersonation	36	62.1
Amplification and visibility dynamics	28	48.3
Governance and adjudication gaps	22	37.9
Automation and scale	14	24.1

Note: Records may be coded to more than one threat vector; therefore, totals can exceed N = 58 and percentages can sum to more than 100%.

Table 2: *Distribution of AI-enabled threat vectors across included records (multi-label coding; N = 58)*

Source: own processing, 2026

The fabrication and impersonation cluster collects studies and reports on the use of synthetic media, deepfakes, and other types of AI-assisted manipulation with harmful reputational, sexualized, and identity-based misuse outcomes in youth contexts. This literature tends to also highlight the challenge of inferring authenticity, plausible deniability, and social salience of content that can be construed as "proof" even if not proven to be true (Jackson et al., 2025).

The amplification and visibility cluster focuses on the ranking, recommendation, and the distribution driven by engagement. In this literature, harm is seen not only in the initial act of an aggressor, but also in terms of how widely content spreads, how frequently it resurfaces, and how much of it is positively reinforced through social symbols that one can see (National Academies of Sciences, Engineering, and Medicine, 2024).

While not as common in youth-related studies, the automation and scale cluster is important to recognize patterns of persistent, high-impact, and sometimes orchestrated harassment which can impose substantial strain on the system of school response (Cai et al., 2023).

The governance and adjudication cluster focused on the boundaries of detection, the friction in reporting systems, the persistence of damaging content across platforms and the administrative hurdles faced by schools and parents. When the content is readily synthesized, remixed and quickly re-uploaded after removal, these issues become severe (Singapore MOE, 2025).

4.4 Mapping AI Mechanisms onto Cyberbullying Process Phases

When reading the literature regarding a five-phase process, the processes of production and circulation emerge as the most highlighted phases. The majority of literature looks at how synthetic content is produced and disseminated and the factors that shape how widely it spreads. Conversely, less attention has been given to interpretation, escalation, and recovery even though such phases are particularly critical when developing school-based interventions. That imbalance indicates that at present, the evidence base is stronger on the creation and distribution of content than on how incidents are evaluated, resolved, and repaired through school workflows (Espino et al., 2023).

A subset of records discusses recovery-oriented processes including documenting evidence, reporting, restorative responses, and psychological support. These records are especially valuable for constructing frameworks because they bridge mechanism-level risks with actionable institutional and pedagogical responses.

4.5 Intervention Implications Discussed in the Evidence Base

Intervention outcomes were organized into three common categories, listed in Table 3 using multi-label coding. Learner-facing implications are: verification; judgment under uncertainty; and responsible sharing – typically put forth as media literacy, digital citizenship, or bystander

response. School-facing implications relate to documentation, reporting pathways, investigation procedures, and coordination with parents or guardians. Platform and governance-facing implications also focus upon reporting usability, detection and moderation limits, provenance or labeling approaches, and cross-platform coordination.

Intervention implication type	n (records coded to type)	% of included records
Learner-facing (verification, judgment under uncertainty, responsible sharing, bystander response)	34	58.6
School-facing (documentation, reporting pathways, investigation practices, parent coordination, restorative responses)	29	50.0
Platform/governance-facing (reporting usability, detection and moderation limits, provenance/labeling, cross-platform coordination)	21	36.2

Note: Records may be coded to more than one implication type; therefore, totals can exceed N = 58 and percentages can sum to more than 100%.

Table 3: *Distribution of intervention implications across included records (multi-label coding; N = 58)*

Source: own processing, 2026

The specific references to AI literacy appear less frequent in the literature than the skillsets that such cases inherently require. Most sources do refer to problems driven by AI, but do not explicitly label those problems as issues of "AI literacy". This pattern highlights the importance of an integrated approach that combines the language of media literacy, the practical orientation of digital citizenship, and the structured skill-building focus of AI competency frameworks (Miao et al., 2024).

4.6 Evidence Gaps and Concentration Patterns

There are clear gaps in the literature that have implications for education and school practice. The first point is that little research is available evaluating specific curricula for remedying synthetic-media harms in the context of cyberbullying. Second, relatively little is known about the procedural realities of school adjudication – such as due process, evidentiary standards, and the handling of false accusations when synthetic artifacts are present. Third, literature on recommendation-driven amplification in youth settings is usually indirect, with interpretations based more on generalized conclusions than direct measurement in school-linked samples (Cristello et al., 2024). These limitations taken together suggest that the conceptual framework proposed here should be considered evidence-informed and practice-oriented and not a group of interventions already validated via outcome testing.

In general, the existing evidence provides strong descriptive support on fabrication, visibility, and governance issues. At the same time, very little empirical research is currently testing targeted educational responses. This pattern shifts into the next element, wherein mechanistic insight of the four modes with respect to the four threat vectors is consolidated, along with identification of the domains where skills and capabilities development could enhance prevention and response-related capacity.

5 Results: Synthesis by AI-Enabled Threat Vector

This section integrates the results using the four AI-enabled threat vectors introduced in Section 2 as the main organizing framework. It identifies (by way of each vector) the most important mechanisms discussed in the literature, the stages of the cyberbullying process most affected, and the recurring implications these patterns carry for prevention and response.

5.1 Fabrication and Impersonation: Synthetic Evidence Harms

Across the literature, fabrication and impersonation emerge as the most frequently discussed AI-related shift. Many of the studies featured describe harms constructed around what can be called synthetic evidence – manipulated images, synthetic audio, falsified screenshots, and deepfake-style impersonation. In youth settings, these materials can be leveraged to shame victims, sexualize their image, or generate plausible "evidence" that promotes peer denunciation (Jackson et al., 2025).

While generating and distributing such content is fast, the primary pressure point is often how information and content are interpreted. The issue is not only that media can be falsified, but also that vivid, apparently well-documented artifacts are often treated as credible within peer groups. As Walter and Tukachinsky (2020) and Lee et al. (2021) suggest, judgments about credibility are influenced by the content, but also visible social endorsement, peer pressure, and prior conflict. Under these conditions, denials or later corrections often have less impact than the first artifact that begins circulating.

Fabrication complicates both escalation and recovery as well. Harm can become more acute when such artifacts are recycled, remixed, or recast as jokes, which may normalize recirculation and subject targets to successive episodes of degrading behaviour. Recovery is also often more difficult, since targets are likely coping not only with the loss of reputation but also with the procedural load of documenting the evidence, explaining manipulation to adults over a longer time frame, and operating within reporting systems that are not very well-structured to deal with issues of authenticity (Wei et al., 2025). The stakes are therefore not simply messages of kindness or courtesy. Competencies related to this focus include, but are not limited to: identifying synthetic media, justifying provenance of sources, and judicious decisions when uncertainty exists. By contrast, some organizations (such as the eSafety Commissioner) warn against relying exclusively on visual inspection. Practical guidance also emphasizes pausing before sharing, contextualizing the content, seeking help from trusted adults or institutions, and preserving evidence without further circulating harmful material (eSafety Commissioner, 2025a).

5.2 Amplification and Visibility Dynamics: Algorithmically Shaped Circulation

As the second set of findings suggests, the gravity of cyberbullying does not stem from the sheer output of damaging messaging, but from the visibility mechanisms embedded in the platforms themselves. Distribution based on ranking, recommendation and engagement can further widen exposure, speed circulation and prolong harm through frequent re-visiting of the same content. In that case, the impacts are most apparent in the stages of circulation and escalation, as harm tends to intensify as users increase and as targets struggle to break free from ongoing cycles of visibility (Ofcom, 2024).

In this literature, cyberbullying is often viewed less as an adversarial interaction between an aggressor and victim and more like a platformed social event. The size and repetition of the audience is integrated into the means of injury. They can compound humiliation, signal social

validation and encourage participation. Other papers also suggest that with sensationalist content being algorithmically amplified and recirculated in response to algorithmic feedback it can become tempting to create blurred demarcating lines between bullying and entertainment (Regehr et al., 2024).

Amplification is also responsible for the interpretation of dangerous material. A few other users will come to believe that seeing repeated exposure may in principle itself mark visibility as a marker for significance or popularity or truth. Kim (2015) and Wang et al. (2023) suggest that being tempted to understand visibility as credibility may exacerbate reputational harm, whether based on the content itself or faked. Implications include knowledge of platforms and algorithms in pragmatic considerations of such media literacy which prioritizes safety. What is needed is not merely technical instruction, but a deeper understanding of engagement loops as mechanisms that perpetuate social signals. This would require an opportunity for students to rehearse not engaging with harmful posts, use reporting tools strategically, and develop norms among peers that can prevent recirculation. Response policies work most effectively on the school level when recirculation is perceived as part of the incident itself and not as a neutral aftereffect. Lloyd (2020) and eSafety Commissioner (2024) claim this.

5.3 Automation and Scale: Volume Harassment and Coordinated Attacks

Third, in published literature, AI can scale up harassment, persistence and coordination. Although youth-specific evidence has been scarce in the literature compared to synthetic media or visibility dynamics literature, some records indicate possible approaches by which AI has been used to enhance the prevalence, persistence and coordination of abuse. This includes automated or semi-automated messaging, the rapid generation of abusive variants, and the reduced effort required to sustain harassment over time (Hinduja, 2023).

This vector seems to have the most acute effects during escalation and recovery phases. Harsh harassment in large numbers may corrode coping resources, provoke anxiety about the constant threat of return, and present obstacles for targets to accurately document what is occurring. It is also a strong potential pressure on school response systems, particularly when the scale of staff and procedural timing cannot keep up with the pace and volume of reports arriving (Lechner et al., 2023).

Automation is also worse when it combines with amplification and governance issues. Scale is particularly harmful when platform distribution systems spread the message further out and where instruments for reporting are not designed to deal with repeated, slightly altered re-uploads. A number of sources suggest that the combination of larger-scale abuse and reporting friction can lower people's confidence to seek help, not only targets, but also bystanders (Boulton et al., 2017; O'Higgins Norman et al., 2024). For this reason, the implications of this are not limited to technical knowledge and are a sign of procedural skills and social norms. Students benefit from learning how to preserve evidence while limiting further exposure, when to seek help early, and how peer support can reduce isolation. Such protocols, clear documentation standards, and advice about when cases should be escalated to platform safety teams or appropriate external agencies (Biernesser et al., 2023) are also needed for schools.

5.4 Governance and Adjudication Gaps: Detection Limits, Reporting Friction, and Due Process

The fourth threat vector revolves around governance and adjudication shortcomings. In the literature to date, the most frequently identified issue is the ineffectiveness of detection and moderation, particularly in cases where the abusive material is subtle, context-based, or swiftly reshuffled into new forms. Reporting also can be complicated to use and harmful material can continue to circulate via duplication and cross-platform re-posting. These problems can be exacerbated with synthetic media, where contention on authenticity makes attribution and institutional decision-making more difficult (Chesney & Citron, 2019; Fisher et al., 2024).

The effect of this vector is strongest on recovery; however, it can work even earlier by affecting the expectations of deterrence and trust. A student's own beliefs that reporting might not lead to meaningful action, or that adults will be unable to understand how manipulation works, might reduce the desire for students to ask for help in the first place. Dennehy et al. (2020), Hsieh et al. (2023) indicate weak confidence in response systems can lower the degree to which aggressors feel that they will face real consequences.

A main point repeatedly raised in this literature is the relevance of procedural clarity and fairness in school response. There is a high level of tension regarding the need for speed of response to protect targets, and the right to due process of accused students, especially when relevant artifacts are unclear or disputed. In the face of this, most responses tend to agree on clear protocols that establish how evidence should be handled, how confidentiality should be maintained, when parents or guardians should step in, and what restorative options may be appropriate. Simultaneously, a number of sources acknowledge that schools might not have the technical capability to independently verify synthetic media (eSafety Commissioner, 2025b). Consequently, the implications extend across students, schools, and platforms. Students and educators need the procedural knowledge required to document and report such incidents in such a way that they do not accidentally circulate harmful content further. Schools need formal workflows, accountability structures, mechanisms of communications to minimize moral panic, while still responding with seriousness. Simultaneously, platforms have also been found to enhance the usability of reporting systems and reduce the continuation of deleterious re-uploads (Matias et al., 2015; OECD, 2023).

5.5 Cross-Vector Interactions and Synthesis Summary

Although the four threat vectors are analytically distinct, studies indicate that incidents rarely involve only one of them. For instance, fabrication harms may be exacerbated when visibility systems prioritize sensational content, and if governance systems do not rapidly remove harmful content or do not restrict re-uploads. To make matters worse, automation also harms people when reporting is sluggish and the risks of repeated exposure remain high. Governance gaps generally cut across all four vectors, as they organize not only measures of deterrence, but also the actual possibility of recovery.

Collectively, these patterns signal three cross-cutting shifts that matter for both explanation and intervention. For one thing, AI is redirecting cyberbullying to types of harm more closely matching aspects of credibility instead, making verification, provenance reasoning, and judgment in the face of ambiguity more critical than ever to individuals in an environment of uncertainty. Second, platform visibility systems seem to magnify audience effects and repeated exposure, making circulation management essential for prevention. Third, recovery is reliant less on perfect verification than on procedural capacity. This is the gap on which most current media literacy approaches are limited – they prioritize interpretation rather than documentation, reporting, and peer-support practices that drive the real experience of in-school response (Vissenberg et al., 2022).

So these repeated patterns do more than merely recapitulate the available literature: they justify the framework set forth in the next part of the study, which recontextualizes media literacy as an operational response capacity as opposed to merely an interpretive skill.

AI-enabled threat vector	Primary cyberbullying phases most affected	Core intervention focus (learner, school, platform/ governance)
Fabrication and impersonation	Interpretation, escalation, recovery	Synthetic media recognition, provenance reasoning, uncertainty calibration; evidence handling and authenticity-dispute procedures; improved reporting and reupload control
Amplification and visibility dynamics	Circulation, escalation, interpretation	Algorithm and platform awareness, reduce engagement and recirculation; policies that treat recirculation as incident behavior; safer ranking, de-ranking, and reporting design
Automation and scale	Escalation, recovery	Early help-seeking and evidence capture without repeated exposure; triage and documentation standards; rate limits and controls for repeat-variant abuse
Governance and adjudication gaps	Recovery, with spillover to earlier phases	Procedural literacy for reporting; transparent school workflows and due process; reporting usability, redress, and persistence reduction

Table 4: *Ultra-brief synthesis summary by AI-enabled threat vector (N = 58)*

Source: own processing, 2026

6 Discussion: An AI-Media Literacy Competency Framework for Cyberbullying Prevention and Response

Based on the synthesized threat landscape, this section argues that existing modes of media literacy are too general for the phenomenon of cyberbullying in the context of AI. To this end, it presents an evidence-informed, practice-oriented framework adapted for adolescents' and schools' contexts with terms such as learner capabilities, school routines, and shared stakeholder responsibilities. It offers the framework as a design roadmap, rather than as a claim of validated intervention effects.

6.1 Framework Logic and Design Principles

The framework is based on three design choices. First, it defines competence as task-based, because school incidents are managed through concrete decisions rather than abstract awareness. Second, it breaks down responsibility across students, teachers, parents or guardians, and schools while keeping platform failures confined to the sphere of platform accountability. Third, it establishes the linkage between each competency and harm-minimization routines that reduce recirculation, preserve evidence, and protect targets from repeated exposure. This focus is in line with advice given by the UK Council for Internet Safety (2024), which states that routines such as these are fundamental to effective safety practice, and with the wider literature on situated learning that highlights the relevance of connecting competence to specific tasks and contexts (New London Group, 1996).

6.2 Competency Cluster A: Synthetic Media Recognition and Verification

This cluster addresses fabrication and impersonation, in which seemingly compelling "proof" is used to harm a peer and platform dynamics can accelerate the spread of that material. From this field, some of the critical measures of learner outcomes could include identifying synthetic media as potential harassment, performing a quick context check, and expressing uncertainty carefully, rather than too quickly asserting certainty. Prominent among them is refraining from forwarding content that targets a peer, including cases where it is reshared "just asking" if it is real. Basic notions of provenance are also critical, in terms of knowing where materials originated from and maintaining a clear chain of custody when materials might be required for reporting (Wittenberg et al., 2025).

For schools, many of the competencies are procedural. Useful practices include confidential intake protocols, secure storage of harmful artifacts, and communication mechanisms that can enable an instructor to separate uncertainty about a piece's authenticity from acknowledgment of harm. Such a methodology helps schools to respond to harm without the need for instant confirmation concerning authenticity but also prevents the temptation of automatically accepting vivid or convincing media (UK Council for Internet Safety, 2024).

6.3 Competency Cluster B: Algorithm and Platform Awareness for Harm Reduction

This cluster accounts for amplification and repeated exposure. Students need to recognize how engagement signals can further spread harmful content and how commenting, reacting, quote-sharing, or remixing can extend or amplify exposure. The intent is to help people find safe alternatives – not to engage humiliating content, not to provoke targets to reclaim harmful visibility via tagging, and using reporting systems where no more content is circulated (Livingstone et al., 2026).

For schools, this cluster reinforces the need to address circulation itself as an aspect of the incident, not a neutral fallout. Advantages include reporting without reposting, supporting individuals with needs behind the scenes as opposed to a loud call to action, and collaborating with families when harmful content spreads across various media (Torgal et al., 2023).

6.4 Competency Cluster C: Privacy, Identity, and Security Practices

This cluster links media literacy with those privacy and security practices that can help decrease the risk of impersonation and facilitate recovery when harm occurs. Learning outcomes in this area include sharing with greater attention to consent, bolstering account security, knowing the basic steps for recovery after an incident, as well as limiting how much identifiable personal information is posted online. This also includes identification of coercive tactics, as well as understanding how impersonation can be used to trigger conflict or falsely implicate a target (Williams et al., 2023).

For schools, the focus is on support, not blame. Reporting norms that minimize shame about account compromise or image misuse, in addition to practical assistance for recovery and reporting, may help to decrease delays in seeking help and attenuate the isolation targets may suffer (Nagar & Talwar, 2023).

6.5 Competency Cluster D: Evidence, Reporting, and Redress

This cluster addresses governance and adjudication gaps, with particular emphasis on recovery. Key to learning these outcomes for individuals is knowing what evidence to capture and how to store it securely, and when the information may be better served by reporting it to a trusted adult through clear reporting channels. Guidance should also articulate common escalation errors including reposting harmful material as "proof," public call-outs that increase visibility, and retaliation as a means of responding to the incident (U.S. Department of Health and Human Services, 2024).

The relevant skill for schools is to have a clear and just process of response. This includes providing defined intake channels, lines of responsibility, standards and protocols documentation, and establishing expectations about confidentiality and realistic response timelines. Where there is uncertainty about how authentic such content actually is, these protocols can sort out the questions between whether it is actually authentic and the questions of harm which this has caused. It is a way to protect targets, sure, but also to protect the due process of accused students. Restorative measures may also be made, if appropriate, and with clear safety protections (Alonso-Rodríguez et al., 2025).

6.6 Competency Cluster E: Prosocial Bystander Response and Resilience

This cluster extends to all four threat vectors, with bystanders potentially having a strong influence on how much harm escalates, how isolated targets become, and whether help is sought at all. Uncertainties are commonly high in environments shaped by AI, so competence is understanding how to act safely, despite incomplete information. Therefore, some important learner-centered outcomes are: supporting in private settings, refraining from recirculating harmful content, walking their peers through the reporting process, and reaching for adult intervention when risks escalate (Chen et al., 2024).

This group understands resilience not simply to endure harm, but as a way of receiving meaningful support and practical resources. Intended results include diminished further exposure, increased understanding of where assistance is available, and restored social support following experiences of public humiliation (McVay et al., 2025).

6.7 Stakeholder Alignment and the Competency-to-Intervention Matrix

The main contribution of this study is the competency-to-intervention matrix shown in Table 5. It is helpful not just to list relevant skills, but to make connections from these to observable learner behaviors and attendant institutional actions. In that sense, it is a more implementable model than many existing AI literacy frameworks, which summarize overarching capacities but leave schools to infer how those capacities should play out in incident response. This relates to competence-based designs that stress clear role expectations and actionable statements of capabilities (Carretero et al., 2017; Vuorikari et al., 2022). The full matrix is presented in Table 5.

Competency cluster	Students	Teachers	Parents/ Guardians	Schools	Platforms/ governance
A. Synthetic media recognition and verification	Pause; do not forward; use context checks; state uncertainty	Treat plausible media as non-conclusive; coach safe capture	Support calm checks; avoid public confrontation	Confidential intake; secure storage; harm-first handling in disputes	Clear reporting for impersonation; reduce reuploads; provenance signals
B. Platform awareness for harm reduction	Avoid engagement that boosts reach; report quietly	Teach "visibility amplifies harm"; model private support	Reinforce "do not engage/ share"; support reporting	Treat recirculation as incident behavior; guidance for low-reach reporting	De-rank humiliating content; add friction; improve reporting feedback
C. Privacy, identity, and security	Consent-aware sharing; account security; recovery steps	Teach protection without blame; notice coercion/ compromise	Help with settings and recovery; keep records	Low-shame reporting; practical recovery support	Faster takeover response; anti-impersonation tools
D. Evidence, reporting, and redress	Document safely; seek adult help early	Confidential handling; guide reporting; reduce exposure	Coordinate with school timelines; escalate when needed	Clear workflow, roles, and timelines; due-process aware response	Streamlined redress; clearer outcomes; persistence reduction
E. Bystander response and resilience	Private support; accompany reporting; avoid amplifying	Teach safe bystander scripts; normalize help-seeking	Encourage support and coping access	Peer-support routines; referral pathways; reintegration support	Tools that reduce that reduce coordinated harassment; private reporting support

Note: The matrix distributes responsibilities across stakeholders and emphasizes harm minimization, evidence preservation, and visibility reduction

Table 5: *Competency-to-intervention matrix for AI-augmented cyberbullying (adolescent and school-relevant contexts)*

Source: own processing, 2026

The matrix also notes one point: Education can improve response capacity, but it cannot redesign platform systems. Platform visibility systems and governance rules are structural conditions beyond students and schools' control. Such measures help steer educational work towards feasible school-based responses while pointing out what platform accountability remains critically needed, including more youth-friendly reporting, stronger limits on persistent re-uploads, and clearer provenance signals (Livingstone et al., 2015; Chhabra et al., 2025). The framework should be intended to build procedural preparedness, rather than pledging technical certainty. In many school-based cases, it is not possible to confirm the authenticity of a contested image, clip, or message quickly enough to guide the initial response; the more urgent priority is to limit harm by slowing recirculation, preserving evidence safely, maintaining confidentiality, and facilitating prompt help-seeking. This has the further effect of underlining the necessity to integrate AI-related risks as part of the current media literacy and digital citizenship provision, as opposed to treating them as an extra specialized area. What is required most for schools is not flawless verification, but a clearly defined, proportionate and psychologically protective workflow that can proactively respond while platform-level safeguards are ultimately the responsibility of the platforms themselves.

7 Conclusion

AI appears to be reshaping adolescent cyberbullying by facilitating fabrication, increasing the volatility of visibility, expanding the scalability of harassment, and making response systems more difficult to navigate. This research integrates these changes into four AI-enabled threat vectors and uses them to construct an AI–media literacy competency framework for prevention and response.

The primary contribution of the framework here is to reposition media literacy as a responsive, rather than merely a critical reading, capacity. It highlights five teachable domains that schools can include in both curriculum and incident handling. It also makes clear that educational adaptation cannot substitute for stronger platform governance and more secure platform design.

This study, as a scoping review, identifies an emerging field rather than estimating effect sizes, and the evidence base remains uneven. Certain assertions (specifically with respect to recommendation systems and visibility dynamics) continue to be more predicated on indirect evidence than direct causal evidence. So, it should instead be considered as an evidence-informed and design-oriented, rather than empirically validated, framework. Even so, the central takeaway remains clear: cyberbullying prevention needs to be AI-ready by integrating judgment under uncertainty, visibility-aware harm reduction, and stronger procedural capacity into everyday school practice.

Bibliography

Alexander, K., Alexander, M. D., & Alexander, F. K. (2022). *The law of schools, students and teachers in a nutshell* (7th ed.). West Academic Publishing.

Alonso-Rodríguez, I., Pérez-Jorge, D., Pérez-Pérez, I., & Olmos-Raya, E. (2025). Restorative practices in reducing school violence: A systematic review of positive impacts on emotional wellbeing. *Frontiers in Education*, *10*, 1520137. https://doi.org/10.3389/feduc.2025.1520137

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19-32. https://doi.org/10.1080/1364557032000119616

Biernesser, C., Ohmer, M., Nelson, L., Mann, E., Farzan, R., Schwanke, B., & Radovic, A. (2023). Middle school students' experiences with cyberbullying and perspectives toward prevention and bystander intervention in schools. *Journal of School Violence*, *22*(3), 339-352. https://doi.org/10.1080/15388220.2023.2186417

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347-364. https://doi.org/10.1007/BF00138871

Boulton, M. J., Boulton, L., Down, J., Sanders, J., & Craddock, H. (2017). Perceived barriers that prevent high school students seeking help from teachers for bullying and their effects on disclosure intentions. *Journal of Adolescence*, *56*(1), 40-51. https://doi.org/10.1016/j.adolescence.2016.11.009

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 39-58). Routledge. https://doi.org/10.4324/9780203876527

Cai, J., Chowdhury, S., Zhou, H., & Wohn, D. Y. (2023). Hate raids on Twitch: Understanding real-time human-bot coordinated attacks in live streaming communities. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 342. https://doi.org/10.1145/3610191

Carretero, S., Vuorikari, R., & Punie, Y. (2017). *DigComp 2.1: The digital competence framework for citizens with eight proficiency levels and examples of use*. Office for Official Publications of the European Communities. https://doi.org/10.2760/38842

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*, *34*(3), 828-846. https://doi.org/10.1287/isre.2022.1161

Chen, H., Chen, C., Li, Y., & Fan, C. (2024). Development and validation of the defending behavior scale of cyberbullying for adolescents. *Behavioral Sciences*, *14*(10), 967. https://doi.org/10.3390/bs14100967

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*. https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

Chew, H. E., Soon, C., & Kaur, H. (2025). *Online harms in Singapore: From evidence to action (IPS working paper no. 68)*. Institute of Policy Studies. https://lkyspp.nus.edu.sg/docs/default-source/ips/wp-68-online-harms-in-singapore.pdf?sfvrsn=a932070a_3

Chhabra, J., Pilkington, V., Benakovic, R., Wilson, M., La Sala, L., & Seidler, Z. (2025). Social media and youth mental health: Scoping review of platform and policy recommendations. *Journal of Medical Internet Research*, *27*, e72061. https://doi.org/10.2196/preprints.72061

Chicote-Beato, M., González-Víllora, S., Bodoque-Osma, A. R., & Navarro, R. (2024). Cyberbullying intervention and prevention programmes in primary education (6 to 12 years): A systematic review. *Aggression and Violent Behavior*, *77*, 101938. https://doi.org/10.1016/j.avb.2024.101938

Costello, N., Sutton, R., Jones, M., Almassian, M., Raffoul, A., Ojumu, O., Salvia, M., Santoso, M., Kavanaugh, J. R., & Austin, S. B. (2024). Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media. *American Journal of Law & Medicine*, *49*(2-3), 135-172. https://doi.org/10.1017/amj.2023.25

Cristello, J. V., Strowger, M., Moreno, M. A., & Trucco, E. M. (2024). Navigating the modern landscape of social media: Ethical considerations for research with adolescents and young adults. *Translational Issues in Psychological Science*, *10*(2), 123-134. https://doi.org/10.1037/tps0000408

Dailey, S. F., Roche, R. R., & Sharkey, M. C. (2025). Addressing bullying and cyberbullying in public health: A systematic review of interventions for healthcare and public health professionals. *International Journal of Environmental Research and Public Health*, *22*(11), 1682. https://doi.org/10.3390/ijerph22111682

Dennehy, R., Meaney, S., Cronin, M., & Arensman, E. (2020). The psychosocial impacts of cybervictimisation and barriers to seeking social support: Young people's perspectives. *Children and Youth Services Review*, *111*, 104872. https://doi.org/10.1016/j.childyouth.2020.104872

Diel, A., Lalgi, T., Mellis, F. S., Teufel, A., & Bäuerle, A. (2025). The harm of deepfakes: A scoping review of deepfakes' negative effects on human mind and behavior. *AI & Society*. https://doi.org/10.1007/s00146-025-02774-0

Duma, N. E., Hlongwa, M., Benjamin-Damons, N., & Hlongwana, K. W. (2023). Physiotherapy management of children with cerebral palsy in low-and middle-income countries: A scoping review protocol. *Systematic Reviews*, *12*, 110. https://doi.org/10.1186/s13643-023-02280-8

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*, 13-29. https://doi.org/10.1038/s44159-021-00006-y

eSafety Commissioner. (2024). *Guide to responding to the sharing of explicit material*. https://www.esafety.gov.au/sites/default/files/2022-02/Respond%204%20-%20Guide%20to%20responding%20to%20the%20sharing%20of%20explicit%20material.pdf

eSafety Commissioner. (2025a). *Guide to responding to image-based abuse involving AI deepfakes*. https://www.esafety.gov.au/sites/default/files/2025-06/Respond%203A%20-%20Guide%20to%20responding%20to%20image-based%20abuse%20involving%20AI%20deepfakes.pdf

eSafety Commissioner. (2025b, June 27). *Deepfake damage in schools: How AI-generated abuse is disrupting students, families and school communities*. https://www.esafety.gov.au/newsroom/blogs/deepfake-damage-in-schools-how-ai-generated-abuse-is-disrupting-students-families-and-school-communities

Espino, E., Guarini, A., & Del Rey, R. (2023). Effective coping with cyberbullying in boys and girls: The mediating role of self-awareness, responsible decision-making, and social support. *Current Psychology*, *42*, 32134-32146. https://doi.org/10.1007/s12144-022-04213-5

Fisher, S. A., Howard, J. W., & Kira, B. (2024). Moderating synthetic content: The challenge of generative AI. *Philosophy & Technology*, *37*, 133. https://doi.org/10.1007/s13347-024-00818-9

Freed, D., Consolvo, S., Cosley, D., Kelley, P. G., Ricart, E., Thomas, K., & Bazarova, N. N. (2025). Help-seeking and coping strategies for technology-facilitated abuse experienced by youth. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), CSCW094. https://doi.org/10.1145/3710992

Hadie, S. N. H. (2024). ABC of a scoping review: A simplified JBI scoping review guideline. *Education in Medicine Journal*, *16*(2), 185-197. https://doi.org/10.21315/eimj2024.16.2.14

Helbach, J., Pieper, D., Mathes, T., Rombey, T., Zeeb, H., Allers, K., & Hoffmann, F. (2022). Restrictions and their reporting in systematic reviews of effectiveness: An observational study. *BMC Medical Research Methodology*, *22*, 230. https://doi.org/10.1186/s12874-022-01710-w

Hinduja, S. (2023, May 10). *Generative AI as a vector for harassment and harm.* https://cyberbullying.org/generative-ai-as-a-vector-for-harassment-and-harm.

Hsieh, M.-L., Wang, S.-Y. K., & Lin, Y. (2023). Perceptions of punishment risks among youth: Can cyberbullying be deterred? *Journal of School Violence*, *22*(3), 307-321. https://doi.org/10.1080/15388220.2023.2183865

Jackson, B. A., Diliberti, M. K., & Moore, P. (2025, September 24). *Artificially intelligent bullies: Dealing with deepfakes in K–12 schools*. https://doi.org/10.7249/RRA3930-5

Jaidka, K., Chen, T., Chesterman, S., Hsu, W., Kan, M.-Y., Kankanhalli, M., Lee, M. L., Seres, G., Sim, T., Taeihagh, A., Tung, A., Xiao, X., & Yue, A. (2025). Misinformation, disinformation, and generative AI: Implications for perception and policy. *Digital Government: Research and Practice*, *6*(1), 11. https://doi.org/10.1145/3689372

Kasturiratna, K. T. A. S., Hartanto, A., Chen, C. H. Y., Tong, E. M. W., & Majeed, N. M. (2025). Umbrella review of meta-analyses on the risk factors, protective factors, consequences and interventions of cyberbullying victimization. *Nature Human Behaviour*, *9*, 101-132. https://doi.org/10.1038/s41562-024-02011-6

Kim, Y. (2015). Exploring the effects of source credibility and others' comments on online news evaluation. *Electronic News*, *9*(3), 160-176. https://doi.org/10.1177/1931243115593318

Lan, M., Law, N., & Pan, Q. (2022). Effectiveness of anti-cyberbullying educational programs: A socio-ecologically grounded systematic review and meta-analysis. *Computers in Human Behavior*, *130*, 107200. https://doi.org/10.1016/j.chb.2022.107200

Lechner, V., Crăciun, I. C., & Scheithauer, H. (2023). Barriers, resources, and attitudes towards (cyber-)bullying prevention/intervention in schools from the perspective of school staff: Results from focus group discussions. *Teaching and Teacher Education*, *135*, 104358. https://doi.org/10.1016/j.tate.2023.104358

Lee, S. S., Liang, F., Hahn, L., Lane, D. S., Weeks, B. E., & Kwak, N. (2021). The impact of social endorsement cues and manipulability concerns on perceptions of news credibility. *Cyberpsychology, Behavior, and Social Networking*, *24*(6), 384-389. https://doi.org/10.1089/cyber.2020.0566

Livingstone, S., Carr, J., & Byrne, J. (2015). *One in three: Internet governance and children's rights.* Centre for International Governance Innovation; The Royal Institute of International Affairs. https://www.cigionline.org/static/documents/no22_2.pdf

Livingstone, S., Jessen, R. S., Stoilova, M., Stänicke, L. I., Graham, R., Staksrud, E., & Jensen, T. (2026). Can platform literacy protect vulnerable young people against the risky affordances of social media platforms? *Information, Communication & Society*, *29*(2), 455-472. https://doi.org/10.1080/1369118X.2025.2518254

Lloyd, J. (2020). Abuse through sexual image sharing in schools: Response and responsibility. *Gender and Education*, *32*(6), 784-802. https://doi.org/10.1080/09540253.2018.1513456

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*, 107238. https://doi.org/10.1016/j.chb.2022.107238

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, reviewing, and responding to harassment on Twitter* [Preprint]. arXiv:1505.03359. https://doi.org/10.48550/arXiv.1505.03359

McVay, S. S., Santo, J., & Lydiatt, H. (2025). The impact of cyberbullying victimization on adolescents' school-related distress across nine countries: Examining the mitigating role of teacher support. *Social Sciences*, *14*(9), 559. https://doi.org/10.3390/socsci14090559

MediaNet. (2025). *Media and information literacy in Kazakhstan*. UNESCO; MediaNet: International Center for Journalism https://articles.unesco.org/sites/default/files/medias/fichiers/2025/09/Mapping-MIL-Report-Eng.pdf

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084

Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835

Milosevic, T., Van Royen, K., & Davis, B. (2022). Artificial intelligence to address cyberbullying, harassment and abuse: New directions in the midst of complexity. *International Journal of Bullying Prevention*, *4*, 1-5. https://doi.org/10.1007/s42380-022-00117-x

Munn, Z., Pollock, D., Khalil, H., Alexander, L., Mclnerney, P., Godfrey, C. M., Christina, M., Peters, M., & Tricco, A. C. (2022). What are scoping reviews? Providing a formal definition of scoping reviews as a type of evidence synthesis. *JBI Evidence Synthesis*, *20*(4), 950-952. https://doi.org/10.11124/JBIES-21-00483

Nagar, P. M., & Talwar, V. (2023). The role of teacher support in increasing youths' intentions to disclose cyberbullying experiences to teachers. *Computers & Education*, *207*, 104922. https://doi.org/10.1016/j.compedu.2023.104922

NAMLE. (2024). *Snapshot 2024: State of media literacy*. National Association for Media Literacy Education. https://namle.org/wp-content/uploads/2024/01/Snapshot-2024-State-of-Media-Literacy-FINAL.pdf

National Academies of Sciences, Engineering, and Medicine. (2024). *Social media and adolescent health*. The National Academies Press. https://doi.org/10.17226/27396

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, *66*(1), 60-93. https://doi.org/10.17763/haer.66.1.17370n67v22j160u

NIST. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

OECD. (2023). *Transparency reporting on child sexual exploitation and abuse online*. Organisation for Economic Co-operation and Development. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/09/transparency-reporting-on-child-sexual-exploitation-and-abuse-online_98fc37bb/554ad91f-en.pdf

Ofcom. (2024, May 8). *Tech firms must tame toxic algorithms to protect children online*. https://www.ofcom.org.uk/online-safety/protecting-children/tech-firms-must-tame-toxic-algorithms-to-protect-children-online

Office of the Surgeon General. (2023). *Social media and youth mental health: The U.S. Surgeon General's advisory*. U.S. Department of Health and Human Services. https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf

O'Higgins Norman, J., Viejo Otero, P., Canning, C., Kinehan, A., Heaney, D., & Sargioti, A. (2024). FUSE anti-bullying and online safety programme: Measuring self-efficacy amongst post-primary students. *Irish Educational Studies*, *43*(4), 865-882. https://doi.org/10.1080/03323315.2023.2174573

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Peters, M. D. J., Marnie, C., Colquhoun, H., Garritty, C. M., Hempel, S., Horsley, T., Langlois, E. V., Lillie, E., O'Brien, K. K., Tunçalp, Ö., Wilson, M. G., Zarin, W., & Tricco, A. C. (2021). Scoping reviews: Reinforcing and advancing the methodology and application. *Systematic Reviews*, *10*, 263. https://doi.org/10.1186/s13643-021-01821-3

Pollock, D., Peters, M. D. J., Khalil, H., McInerney, P., Alexander, L., Tricco, A. C., Evans, C., de Moraes, B. É., Godfrey, C. M., Pieper, D., Saran, A., Stern, C., & Munn, Z. (2023). Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBI Evidence Synthesis*, *21*(3), 520-532. https://doi.org/10.11124/JBIES-22-00123

Prem, E., & Krenn, B. (2023). On algorithmic content moderation. In H. Wertjner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to digital humanism: A textbook* (pp. 481-493). Springer. https://doi.org/10.1007/978-3-031-45304-5_30

Raghuvanshi, A., Joshi, S., & Shakya, U. (2024). *Safeguarding futures: Exploring the impacts of generative AI on child online protection in Nepal*. ChildSafeNet. https://www.unicef.org/nepal/media/24156/file/No%20ai%20safeguarding.pdf

Regehr, K., Shaughnessy, C., Zhao, M., & Shaughnessy, N. (2024). *Safer scrolling: How algorithms popularise and gamify online hate and misogyny for young people.* UCL IOE; University of Kent. https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, & A. John (Eds.), *Proceedings of the 2023 AAAI/ACM conference on AI, ethics, and society* (pp. 723-741). ACM. https://doi.org/10.1145/3600211.3604673

Silk, J. S., Sequeira, S. L., James, K. M., Kilic, Z., Grad-Freilich, M. E., Choukas-Bradley, S., & Ladouceur, C. D. (2024). The role of neural sensitivity to social evaluation in understanding "for whom" social media use may impact emotional health during adolescence. *Affective Science*, *5*, 366-376. https://doi.org/10.1007/s42761-024-00252-2

Singapore MOE. (2025, April 4). *How do MOE and schools manage bullying and hurtful behaviours?* https://www.moe.gov.sg/news/edtalks/how-do-moe-and-schools-manage-bullying-and-hurtful-behaviours

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. https://doi.org/10.1111/j.1469-7610.2007.01846.x

Sorrentino, A., Sulla, F., Santamato, M., di Furia, M., Toto, G. A., & Monacis, L. (2023). Has the COVID-19 pandemic affected cyberbullying and cybervictimization prevalence among children and adolescents? A systematic review. *International Journal of Environmental Research and Public Health*, *20*(10), 5825. https://doi.org/10.3390/ijerph20105825

Suri, H. (2019). Ethical considerations of conducting systematic reviews in educational research. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic reviews in educational research: Methodology, perspectives and application* (pp. 41-54). Springer. https://doi.org/10.1007/978-3-658-27602-7_3

Tayie, S. S. (2025). Fostering algorithmic literacy in education: Navigating news ecosystems for critical media understanding. *Comunicar*, *33*(82), 127-137. https://doi.org/10.5281/zenodo.16388403

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277-287. https://doi.org/10.1016/j.chb.2009.11.014

Torgal, C., Espelage, D. L., Polanin, J. R., Ingram, K. M., Robinson, L. E., El Sheikh, A. J., & Valido, A. (2023). A meta-analysis of school-based cyberbullying prevention programs' impact on cyber-bystander behavior. *School Psychology Review*, *52*(2), 95-109. https://doi.org/10.1080/2372966X.2021.1913037

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467-473. https://doi.org/10.7326/M18-0850

UK Council for Internet Safety. (2024). *Sharing nudes and semi-nudes: Advice for education settings working with children and young people*. UK Council for Internet Safety. https://assets.publishing.service.gov.uk/media/65d62b02188d770011038855/UKCIS_sharing_nudes_and_semi_nudes_advice_for_education_settings__Web_accessible.pdf

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In F. Floyd Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI conference on human factors in computing systems* (article 779). ACM. https://doi.org/10.1145/3613904.3642382

U.S. Department of Health and Human Services. (2024, December 11). *Report cyberbullying*. https://www.stopbullying.gov/cyberbullying/how-to-report

Vissenberg, J., d'Haenens, L., & Livingstone, S. (2022). Digital literacy and online resilience as facilitators of young people's well-being? A systematic review. *European Psychologist*, *27*(2), 76-85. https://doi.org/10.1027/1016-9040/a000478

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The digital competence framework for citizens with new examples of knowledge, skills and attitudes*. Office for Official Publications of the European Communities. https://doi.org/10.2760/115376

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, *47*(2), 155-177. https://doi.org/10.1177/0093650219854600

Wang, S., Chu, T. H., & Huang, G. (2023). Do bandwagon cues affect credibility perceptions? A meta-analysis of the experimental evidence. *Communication Research*, *50*(6), 720-744. https://doi.org/10.1177/00936502221124395

Wei, M., Consolvo, S., Kelley, P. G., Kohno, T., Roesner, F., & Thomas, K. (2023). "There's so much responsibility on users right now:" Expert advice for staying safer from hate and harassment. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI conference on human factors in computing systems* (article 190). ACM. https://doi.org/10.1145/3544548.3581229

Wei, M., Yeung, C., Roesner, F., & Kohno, T. (2025). "We're utterly ill-prepared to deal with something like this": Teachers' perspectives on student generation of synthetic nonconsensual explicit imagery. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty,

& P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI conference on human factors in computing systems* (article 174). ACM. https://doi.org/10.1145/3706598.3713226

Williams, O., Choong, Y.-Y., & Buchanan, K. (2023). Youth understandings of online privacy and security: A dyadic study of children and their parents. In P. Gage Kelley, & A. Kapadia (Eds.), *Proceedings of the nineteenth symposium on usable privacy and security* (pp. 399-416). USENIX Association.

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, *4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Zhang, W., Huang, S., Lam, L., Evans, R., & Zhu, C. (2022). Cyberbullying definitions and measurements in children and adolescents: Summarizing 20 years of global efforts. *Frontiers in Public Health*, *10*, 1000504. https://doi.org/10.3389/fpubh.2022.1000504

Appendix

This scoping review was conducted through a systematized search process to extract evidence from multidisciplinary research on adolescent cyberbullying and related youth online harassment using AI-related features. The search was structured to encompass four main focus mechanisms, namely those involving – (1) synthetic media and generative fabrication or impersonation, (2) algorithmic ranking and recommendation, (3) automation at scale, and (4) automated detection, moderation, reporting, and redress.

The review included records published from 1 January 2018 to 31 December 2025. All searches were conducted on 26 January 2026. Only records in English were included.

Searches of the primary databases were performed in ERIC, PsycINFO, Scopus, and Web of Science Core Collection. Google Scholar and selected organizational sources were utilized for additional searching. Targeted organizational sources were significant intergovernmental and policy or research organizations on education, youth online safety, and AI-related harms (e.g., UNESCO, UNICEF, OECD, Council of Europe, RAND, and selected reputable youth online safety organizations).

Searches combined three concept blocks:

(1) population terms: adolescen* OR teen* OR youth OR “young people” OR student* OR “school-age” OR “secondary school” OR “K-12”;

(2) cyberbullying terms: cyberbull* OR “online bullying” OR “online harassment” OR “peer harassment” OR “digital harassment” OR “online abuse”;

(3) AI mechanism terms: deepfake* OR “synthetic media” OR “AI-generated” OR “generative AI” OR “voice clon*” OR “text-to-image” OR “image generation” OR “recommender system*” OR recommend* OR “ranking algorithm*” OR “algorithmic amplification” OR “content moderation” OR “automated moderation” OR bot* OR “automated account*”.

Example database search string (Scopus; TITLE-ABS-KEY).

TITLE-ABS-KEY

(adolescen* OR teen* OR youth OR “young people” OR student* OR “school-age” OR “secondary school” OR “K-12”)

AND

(cyberbull* OR “online bullying” OR “online harassment” OR “peer harassment” OR “digital harassment” OR “online abuse”)

AND

(deepfake* OR “synthetic media” OR “AI-generated” OR “generative AI” OR “voice clon*” OR “recommender system*” OR recommend* OR “ranking algorithm*” OR “algorithmic amplification” OR “content moderation” OR “automated moderation” OR bot* OR “automated account*”)

Database syntax was adapted as needed for each platform. Filters were applied consistently where available: publication years 2018–2025, English language, and article/review or peer-reviewed limits where supported by the database.

Database retrieval counts (before deduplication).

- Scopus: 1,485
- Web of Science Core Collection: 1,020
- ERIC: 2
- PsycINFO: 708

Together, the database searches identified 3,215 records before supplementary searching.

Google Scholar and targeted organizational sources were used to identify additional potentially relevant records, including high-quality public reports and items that may not have been fully captured in the database searches. Because search-engine and website result totals are unstable and not consistently reproducible, raw hit counts from these supplementary sources were not used as the headline identification total. Instead, only records that were actually exported and screened from supplementary searching were counted in the review workflow. Any additional records identified in this way were incorporated into the screening process and are reflected in the PRISMA-style flow diagram as records identified through supplementary or other methods.

Author

Zhu Luo, MA.
University of New South Wales
Faculty of Arts, Design & Architecture (ADA)
High St,
Kensington NSW 2052, Australia
Z5545584@zmail.unsw.edu.au
ORCID-ID: 0009-0009-8922-2067

Zhu Luo is a researcher in Journalism and Communication at UNSW Sydney. His research sits at the intersection of journalism studies and AI, focusing on how generative AI is framed in public discourse and how AI assisted content reshapes credibility and trust online. Through a mixed methods and quantitative qualitative content analysis methodological approach, he analyses rhetorical patterns in opinion journalism and audience responses to AI-mediated citizen journalism in major Chinese platforms including Douyin, Weibo, Bilibili, and Xiaohongshu. He concentrates on transparency signals – creator disclosures, provenance features, and platform affordances – and their impact on perceptions of authenticity, responsibility, and journalistic legitimacy. He aims to create tangible, evidence-based recommendations that will enhance media literacy and support clearer provenance and accountability standards in digital media.