

Epistemic trust in AI-mediated social media: An integrative review of young adults' encounters with misinformation

¹Chen Chia-hao*

National Tsing Hua University, College of Education, Taiwan

ABSTRACT

Generative artificial intelligence (AI) is changing how young adults encounter, produce, and assess information online. This integrative review examines how AI-mediated communication intersects with epistemic trust and misinformation, with particular attention to emerging adults who use social platforms for information and social interaction. Rather than presenting new empirical data, the article synthesizes recent peer-reviewed research and selected cross-national evidence published mainly from 2020 to 2026. The literature points to a dual problem. Generative systems can produce fluent but inaccurate material and can lower the cost of creating persuasive synthetic content; at the same time, AI-assisted explanation, correction, and detection may help users evaluate claims. Trust is therefore not simply increased or decreased by AI. It is shaped by source cues, disclosure, perceived expertise, prior beliefs, platform design, and the availability of independent verification. Young adults' familiarity with digital media should not be equated with reliable misinformation detection, and blanket distrust can be as damaging as uncritical acceptance. The review develops a framework connecting system characteristics, social context, individual judgment, and institutional accountability. It argues that responses should combine specific provenance and verification cues, media and AI literacy, accessible independent sources, and transparent platform governance. The evidence base remains uneven across countries, languages, and age groups; accordingly, the article identifies testable questions for comparative research rather than claiming universal effects.

ARTICLE HISTORY

Revised 05 September, 2026
Accepted 23 September 2026

KEYWORDS

Generative artificial intelligence;
Epistemic trust; Misinformation;
Young adults; Social media

Cite as: Chia-hao, C. (2026). Epistemic trust in AI-mediated social media: An integrative review of young adults' encounters with misinformation. *Aureviaj Journal of Social Sciences and Humanities*. 1(1), 53-63 <https://aureviaj.com/ajssh/article/view/23>



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Digital communication has long required people to make judgments about information without direct access to the events, expertise, or evidence behind it. Social media intensifies this problem because information travels through networks of friends, influencers, publishers, anonymous accounts, and automated systems. Generative artificial intelligence (AI) adds another layer: the material encountered online may be written, edited, summarized, translated, or visually represented by a system that produces fluent outputs without necessarily establishing whether their claims are true. For young adults, who often move between social interaction, entertainment, education, and news within the same platforms, the distinction

*CONTACT ¹Chen Chia-hoe c.chiho98@gmail.com National Tsing Hua University, College of Education, Taiwan

between interpersonal communication and information production is increasingly difficult to maintain.

The issue is not simply whether people believe AI-generated content. A more useful social-scientific question concerns epistemic trust: the degree to which a person treats a communication or source as a reliable basis for belief. Trust is necessary for everyday life because individuals cannot independently verify every claim. Yet trust can be misplaced when fluency, familiarity, popularity, or apparent technical authority substitutes for evidence. Conversely, distrust can also be misdirected. If people reject accurate information merely because they suspect it was produced or assisted by AI, they may discount useful communication or unfairly discredit its author. Sahebi and Formosa (2025) describe this as a dilemma in AI-mediated communication: lower trust may be warranted in some cases but unjustified in others, creating risks of both gullibility and epistemic injustice.

Recent research confirms that generative AI and misinformation are now studied as connected rather than separate problems. Park and Nan's (2025) scoping review of 24 empirical studies found that large language models can generate convincing misinformation, but can also assist detection and resistance; mitigation results were mixed. A separate review of research on young people and false information synthesized 151 studies and reported that young people often overestimate their ability to identify false information, while cognitive biases, demographic factors, and peer dynamics influence responses (Kops et al., 2025). These findings suggest that technological capacity, user judgment, and social context need to be examined together.

Survey evidence also shows why the issue has immediate relevance. In six countries, the Reuters Institute's 2025 report found that weekly use of generative AI increased from 18% in 2024 to 34% in 2025; among respondents aged 18–24, 59% reported using an AI system in the previous week. Information seeking became a leading use, while only a minority regularly used AI specifically to obtain news (Simon et al., 2025). The report also found that public comfort declined as news production became more fully automated and that people's trust in AI-generated search answers was conditional, particularly in high-stakes areas such as health and politics. These are cross-national survey findings, not proof that young adults are uniquely vulnerable, but they show that AI-assisted information seeking is becoming part of everyday life.

Existing scholarship offers several strands of explanation. Misinformation research examines the creation, circulation, and correction of false claims. Communication research considers how source cues and platform architecture shape credibility judgments. Work on AI-mediated communication addresses how automation changes the relationship between a message and its apparent author. Research on young people considers digital practices, peer influence, and confidence in their own evaluative skills. However, these strands are not always integrated. Reviews of AI and misinformation generally map technological uses and mitigation strategies, while studies of young people's misinformation experiences often predate widespread consumer generative AI. The resulting gap concerns how epistemic trust among young adults may be negotiated when human and machine contributions are mixed, disclosure is inconsistent, and social endorsement competes with independent verification.

This article addresses that gap through an integrative review of recent research and selected institutional evidence. It asks three questions: (1) How does AI-mediated communication alter the cues young adults may use to judge information credibility? (2) What mechanisms connect generative AI, social influence, and misinformation-related trust? (3) Which interventions are supported by the available evidence, and where are the uncertainties greatest? The article's contribution is conceptual rather than empirical: it organizes the literature around a framework connecting system characteristics, message and source cues, social context, and user judgment. It also distinguishes evidence about young adults specifically from broader findings about general populations. This distinction matters because claims about a digitally connected generation can otherwise become assumptions rather than testable propositions

RESEARCH METHODOLOGY

This article uses an integrative literature-review design. An integrative review is appropriate when a question spans different forms of evidence, including empirical studies, conceptual analysis, and high-quality institutional surveys. The purpose here is to compare explanatory approaches and develop a framework, not to estimate a pooled effect or claim an exhaustive systematic review. No new participants were recruited, and no original dataset was collected.

Literature identification focused on publications from 2020 through October 2026, with particular attention to peer-reviewed work published in 2024–2026. Searches were conducted using combinations of terms including “generative AI,” “AI-mediated communication,” “epistemic trust,” “misinformation,” “young adults,” “emerging adults,” “social media,” “AI literacy,” and “disclosure labels.” Search results and reference lists were used to identify work relevant to at least one of the three review questions. Priority was given to peer-reviewed review articles, empirical studies with a clearly described method, and institutional survey reports that disclosed their population and fieldwork context. Foundational work was retained where needed to define concepts or explain established mechanisms.

Sources were included when they addressed one or more of the following: AI-assisted or AI-generated information; credibility, trust, or misinformation judgments; young people's information practices; or interventions intended to improve evaluation of online claims. Material that discussed AI only as a technical tool without a clear connection to communication or social judgment was not central to the synthesis. The review did not treat preprints, opinion pieces, and institutional reports as equivalent to peer-reviewed experiments. Institutional evidence was used to contextualize adoption and public attitudes, while causal claims were drawn only where the underlying study design supported them.

The analysis proceeded thematically. First, concepts were distinguished: misinformation refers to false or misleading content regardless of intent, whereas disinformation ordinarily implies deliberate deception. Epistemic trust refers to reliance on a communication as a source of knowledge, and perceived credibility is the recipient's judgment that a message or source is believable. Second, findings were grouped under four themes: (a) automation and credibility cues; (b) young adults' evaluative practices and social context; (c) trust, distrust, and misinformation interventions; and (d) institutional and design responses. Third, similarities and

tensions across themes were compared to develop a conceptual framework and identify questions for further research.

This method has limitations. Search coverage was not exhaustive, the synthesis did not use duplicate independent screening or formal risk-of-bias scoring, and the included studies differed in population, outcome measures, and research design. It therefore cannot establish the prevalence of any behavior among young adults or support a quantitative ranking of interventions. The findings below should be read as an interpretive synthesis of available evidence, not as the results of a preregistered systematic review.

RESULTS

Generative Fluency and the Problem of Credibility Cues

The reviewed literature identifies a mismatch between the surface quality of AI-generated material and its evidential reliability. Large language models can produce coherent, context-sensitive prose, but coherence does not guarantee factual accuracy. Park and Nan (2025) found evidence that generative systems can produce persuasive misinformation and that exposure may influence trust and decision-making. The social risk is not that every output is false, but that the presentation of a claim can look more certain than the process that produced it warrants.

This mismatch matters because recipients routinely use cognitive shortcuts. A message that is fluent, specific, emotionally resonant, or delivered in a confident tone may be treated as credible before the recipient checks its sources. The “machine heuristic” discussed in research on media credibility captures the tendency to attribute competence or objectivity to machine-generated communication (Sundar, 2008, as discussed in Park & Nan, 2025). In AI-mediated settings, the heuristic can work in opposite directions: some users may regard automated output as neutral and authoritative, while others may discount it because it is automated. Neither response is a reliable substitute for evaluating evidence.

The source of a message can also become ambiguous. A social-media post may be written by a person, revised by an AI tool, generated from a prompt, or copied from an AI response and reposted without disclosure. This creates a problem of provenance: users may not know who is responsible for the claim, what sources informed it, or whether a human checked it. Sahebi and Formosa (2025) argue that AI-mediated communication can weaken epistemic trust while also creating unjustified suspicion toward some communications. This means that disclosure alone is not enough; the disclosure needs to be informative and paired with evidence that readers can inspect.

A related distinction is between trusting a tool and trusting a particular output. A user may find an AI assistant useful for explaining a concept while recognizing that it can invent a citation or omit context. Conversely, a person may distrust AI in principle but accept a particular AI-generated message because it confirms an existing belief. Trust is therefore task-specific and relational. It depends on what the system is being asked to do, the consequences of error, the user’s prior knowledge, and whether the claim can be independently verified.

Young Adults, Digital Familiarity, and Evaluative Practice

The literature cautions against equating frequent digital-media use with strong misinformation literacy. Kops et al. (2025), reviewing 151 studies on young people and false information, report that young people often overestimate their ability to detect false claims. Their review also identifies cognitive biases and peer dynamics among factors shaping responses. This does not mean that young adults are uniformly poor at evaluating information. It indicates that confidence, exposure, and competence should be measured separately rather than inferred from age or technology use.

Young adults may encounter information in environments where social and informational purposes overlap. A claim may arrive through a friend's story, a creator's video, a group chat, or an AI-generated summary. In these settings, users can rely on relational cues such as who shared the item, how the sender is regarded, and whether other members of a group endorse it. These cues can be useful in ordinary social coordination, but they are imperfect indicators of factual accuracy. A trusted friend may unknowingly forward a false claim, while a less familiar source may provide well-supported information.

Research on emerging adults' trust practices supports treating credibility judgments as socially situated rather than purely individual. Sharon and Encarnación (2024), using in-depth interviews with 19 young adults, describe different ways participants navigated competing truth claims across family, friends, and social media. The small exploratory sample cannot represent young adults generally, but it illustrates how belief formation is embedded in identity, relationships, and prior experiences of acceptance or rejection. AI-mediated information enters these existing social processes rather than replacing them.

Digital familiarity can help users navigate interfaces, compare sources, and recognize common platform conventions. Yet it does not automatically confer knowledge of model limitations, training data, synthetic media, or the difference between generated text and retrieved evidence. AI literacy therefore needs to include practical knowledge about uncertainty and verification, not just the ability to operate a tool. A user who can prompt a model effectively may still be unable to assess the accuracy of its response. Equally, a user with modest technical skill may evaluate a claim well if they have access to credible sources and a clear verification routine.

A further caution concerns generalization. Much published research on AI use and public trust has been conducted in higher-income countries and in languages with comparatively strong digital research infrastructure. Cross-national surveys show substantial variation in adoption and attitudes, and their results cannot simply be transferred to contexts with different media systems, language resources, connectivity, educational opportunities, or political conditions. The term "young adults" also covers people with different levels of education, employment, access, and civic participation. These differences should be treated as explanatory variables, not noise around a single generational profile.

Trust, Distrust, and the Double Role of AI

The evidence does not support a simple conclusion that generative AI necessarily increases or decreases trust. Park and Nan (2025) found a dual role: systems can contribute to the production of misinformation but can also assist detection and mitigation. This duality means that the

social consequences depend on how systems are designed, used, disclosed, and checked. The relevant unit of analysis is often not “AI” in general, but a particular output, task, interface, and information environment.

The 2025 Reuters Institute report offers a useful view of public attitudes. Across six countries, 50% of people who encountered AI-generated search answers said they trusted them, while many remained neutral or cautious; users were more likely to verify information in high-stakes domains such as health and politics. The report also found that only 12% were comfortable with news made entirely by AI, compared with 62% for news made entirely by humans, with intermediate levels of comfort when humans remained involved (Simon et al., 2025). These results concern broad publics in six countries, not young adults alone. They nevertheless suggest that trust is sensitive to the perceived role of human judgment and the stakes of the task.

Trust can fail in at least two directions. Over-trust occurs when users accept a generated claim without sufficient checking, perhaps because it is articulate or agrees with their prior beliefs. Under-trust occurs when users reject a reliable claim solely because AI contributed to its production, or when widespread synthetic content creates generalized suspicion toward online communication. The latter problem can damage the informational value of social platforms and make it harder for users to distinguish well-supported reporting from fabricated material. Sahebi and Formosa (2025) frame this as a normative dilemma: reducing trust may protect users in one case and produce unfair dismissal in another.

Misinformation interventions also have mixed outcomes. Corrections can help users revise false beliefs, but their effects depend on the content, the audience, and the social context. Pre-emptive warnings or inoculation approaches aim to prepare people to recognize manipulation before they encounter it. In a 2025 study, Spearing examined source-focused pre-emptive discreditation, content-focused debunking, and a general disclaimer in relation to AI-generated misinformation. The study adds experimental evidence to a field where concern about harms can outpace knowledge of which countermeasures work under which conditions. Its results should not be interpreted as proof that one message or disclaimer will reliably protect all audiences.

An important implication is that labels should not be treated as complete interventions. A generic “AI-generated” label communicates something about production but not whether the content is true, whether it has been checked, or which sources support it. A label that specifies independent verification status may give users more actionable information, but labels can be ignored, misunderstood, or used strategically by bad actors. Effective design should combine provenance information with accessible source links, correction mechanisms, and explanations of uncertainty. Users should be able to move from a credibility cue to the evidence behind the claim.

A Multi-Level Framework for Epistemic Trust

The synthesis suggests that epistemic trust in AI-mediated social media can be understood through four interacting levels. These levels are analytical categories, not a validated scale, and their relationships require empirical testing.

Table 1. *A multi-level framework for epistemic trust in AI-mediated information*

Level	Mechanisms	Research and practice implications
System and content	Fluency, factual error, synthetic realism, uncertainty cues, traceable sources	Assess factual accuracy separately from readability; provide evidence and provenance.
Message and source	Authorship, disclosure, reputation, expertise, human review	Clarify who is responsible and what checking has occurred.
Social and platform context	Peer endorsement, recommender systems, engagement incentives, group identity	Study circulation and social endorsement, not only individual accuracy judgments.
User and institution	Prior beliefs, AI literacy, confidence, access to independent sources, platform accountability	Build verification skills and ensure that trustworthy alternatives are accessible.

Note. Framework developed from the integrative synthesis; it is a conceptual proposition and has not been empirically validated.

At the system and content level, credibility depends partly on whether the output is accurate, appropriately qualified, and connected to inspectable evidence. At the message and source level, readers need to understand whether the communication was generated, edited, or reviewed by a person and which actor is accountable for errors. At the social and platform level, information is interpreted and circulated within networks shaped by group identity and platform incentives. At the user and institutional level, prior beliefs and AI literacy interact with access to reliable alternatives and the rules governing platform behavior.

The levels are interdependent. A clear disclosure may be ineffective when users have no time or ability to check the source. Strong individual literacy may be insufficient when an interface hides provenance or rewards rapid sharing. A correction may be ignored when it threatens group identity or comes from a distrusted institution. Conversely, a well-designed interface can make verification easier, while social norms that reward correction rather than ridicule can reduce the interpersonal cost of acknowledging an error. The framework therefore discourages interventions that place the entire burden on individual users.

This framework also clarifies why trust should not be measured with a single general question such as “Do you trust AI?” A more informative design would distinguish trust in a tool, trust in an individual output, perceived accuracy, willingness to share, confidence in source attribution, and actual verification behavior. These outcomes may diverge. Someone can believe a message is probably accurate but decline to share it; another person can doubt a claim

yet circulate it as a question or joke. Studies that collapse these outcomes risk drawing misleading conclusions about the relationship between trust and behavior.

Evidence Gaps Identified by the Review

First, the evidence base is uneven across age groups and national contexts. Studies often report results for broad adult samples, while work focused on young people may examine misinformation in social media without isolating generative AI. Researchers should avoid describing general-population findings as youth-specific. Studies in Afghanistan and other under-researched settings would be particularly useful because language availability, platform use, media access, and trust in institutions may differ from conditions in the countries most often studied.

Second, outcomes are not standardized. Credibility ratings, factual discernment, sharing intention, actual sharing, trust in AI, and willingness to verify are related but distinct. Experimental findings about a single synthetic image or short text cannot automatically predict behavior in a real feed, where users encounter repeated messages, competing sources, and social feedback. Future studies should report outcome definitions clearly and, where possible, measure behavior as well as stated intention.

Third, research often tests interventions in isolation. Labels, corrections, prebunking, source links, and AI literacy training may interact. A label may be more useful when a verification link is visible; a correction may work differently when delivered by a trusted source; and literacy training may have limited impact if platform design makes source checking difficult. Factorial experiments and longitudinal studies could assess these interactions without assuming that a single intervention will solve the problem.

Fourth, evidence on multilingual and culturally situated AI-mediated communication remains limited. Generative systems may vary in quality across languages, and users may have unequal access to independent fact-checking resources. Research should test whether credibility cues and interventions work similarly across languages and political contexts rather than assuming that an intervention validated in one country will transfer unchanged. Such comparisons are necessary to identify which mechanisms are general and which depend on local institutions or media practices.

DISCUSSION

The central conclusion of this review is that epistemic trust in AI-mediated social media is conditional, socially situated, and distributed across several actors. The available evidence does not justify treating young adults as either naturally skilled digital evaluators or uniquely gullible users. Rather, their judgments arise from the interaction of system behavior, communication cues, peer relationships, prior beliefs, and institutional conditions. This interpretation is consistent with the review by Kops et al. (2025), which identifies cognitive and social factors in young people's responses to false information, and with Park and Nan's (2025) account of generative AI's dual role in misinformation generation and mitigation.

The first implication concerns the difference between persuasive presentation and reliable knowledge. AI systems can generate polished text quickly, but the social meaning of fluency

depends on the recipient's assumptions about machines and expertise. If users infer that a machine-generated response is objective, they may under-check it. If they infer that all AI involvement makes a message untrustworthy, they may dismiss accurate material. Sahebi and Formosa's (2025) account of epistemic trust helps explain why neither blanket confidence nor blanket suspicion is an adequate response. The practical goal should be calibrated trust: confidence proportionate to the quality of evidence, the transparency of the process, and the stakes of the decision.

The second implication is that digital competence needs to be defined more carefully. The ability to navigate a platform, create content, or prompt an AI assistant is not equivalent to the ability to judge the truth of a claim. AI literacy should include understanding that generated language may contain factual errors, knowing how to trace a claim to independent sources, distinguishing source attribution from source reliability, and recognizing when a question requires specialist or local knowledge. These skills are best taught through repeated practice with realistic examples rather than through abstract warnings alone.

Third, interventions should be designed around the information environment rather than the individual alone. Platforms influence what users see, how often they encounter similar claims, and whether a correction is visible. News organizations and public institutions influence whether reliable alternatives are accessible and whether AI use is disclosed. Schools and universities can teach verification, but cannot compensate fully for opaque interfaces or inaccessible evidence. A shared-responsibility approach would assign clear obligations to developers, platforms, publishers, educators, and users, while avoiding the assumption that every misinformation problem can be solved through a fact-checking label.

The Reuters Institute's 2025 findings are relevant here because they indicate both growing AI use and continuing reservations about its role in news. The report found a substantial difference in comfort between entirely human-made news and news produced entirely by AI, and it reported that users' trust in AI search answers varied by context (Simon et al., 2025). These survey findings do not establish how young adults respond in every setting, but they suggest that human oversight and visible accountability may matter to public acceptance. Further research should test whether the perception of human review improves accuracy judgments, sharing decisions, or only stated comfort.

Disclosure is likely to be most useful when it answers practical questions. A label should, where possible, indicate the nature of AI involvement, identify the responsible publisher or author, and distinguish between content that has merely been generated and content that has been independently checked. However, disclosure can also become a ritual that transfers responsibility to users without improving the evidence available to them. It should therefore be evaluated by whether it helps recipients make more accurate judgments, not simply by whether a label is present.

There are important limits to this synthesis. The review was integrative rather than systematic, and the selection and interpretation of studies may have omitted relevant research. Studies differed in their definitions of trust, age categories, platforms, and outcomes. Some findings derive from conceptual analysis, others from experiments or surveys, and they cannot be treated as interchangeable. The evidence reviewed also develops quickly as products and platform

features change. These limitations mean that the framework should be considered a set of propositions for testing rather than a settled causal model.

Future studies should compare young adults across languages and social contexts, distinguish AI-generated from AI-assisted communication, and measure both belief accuracy and behavior. Longitudinal designs could examine whether repeated exposure changes trust calibration over time. Experiments could compare generic AI labels with disclosures that specify human review and provide accessible sources. Field studies could test whether users actually follow verification links when those links are made more visible. Researchers should also investigate whether interventions have unequal effects across educational backgrounds, language groups, and levels of institutional trust. Such work would move the field beyond general claims about AI's dangers or benefits toward more precise accounts of when, for whom, and under what conditions trust is warranted.

CONCLUSION

Generative AI has complicated the social process through which people decide what to believe online. Its outputs can be fluent without being accurate, and its ability to produce convincing material can support misinformation. Yet AI can also assist explanation, detection, and correction. The available literature therefore points to a problem of calibration rather than a simple choice between trust and distrust. Young adults' digital familiarity should not be assumed to guarantee effective evaluation, and evidence from broad adult samples should not be presented as youth-specific without justification. This review proposes a multi-level framework linking system and content characteristics, source cues, social-platform context, and user and institutional conditions. Practical responses should combine transparent provenance, accessible independent evidence, AI and media literacy, and accountable platform design. Because existing studies vary in populations, methods, and outcome measures, the framework remains provisional. Comparative, multilingual, and behavior-focused research is needed to establish which interventions improve accuracy and verification in real information environments.

REFERENCES

- Altay, S., Hoes, E., & Wojcieszak, M. (2025). Following news on social media boosts knowledge, belief accuracy and trust. *Nature Human Behaviour*, 9, 1833–1842. <https://doi.org/10.1038/s41562-025-02205-6>
- Kops, M., Schittenhelm, C., & Wachs, S. (2025). Young people and false information: A scoping review of responses, influential factors, consequences, and prevention programs. *Computers in Human Behavior*, 108650. <https://doi.org/10.1016/j.chb.2025.108650>
- López-Borrull, A., & Lopezosa, C. (2025). Mapping the impact of generative AI on disinformation: Insights from a scoping review. *Publications*, 13(3), Article 33. <https://doi.org/10.3390/publications13030033>

- Park, S., & Nan, X. (2025). Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY. Advance online publication*.
<https://doi.org/10.1007/s00146-025-02620-3>
- Sahebi, S., & Formosa, P. (2025). The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence. *Synthese*, 205, Article 128. <https://doi.org/10.1007/s11229-025-04963-2>
- Sharon, T., & Encarnación, M. (2024). Social media & declining trust: An epistemic challenge for emerging adults? *Emerging Adulthood*, 12(3).
<https://doi.org/10.1177/21676968241234091>
- Simon, F. M., Nielsen, R. K., & Fletcher, R. (2025). *Generative AI and news report 2025: How people think about AI's role in journalism and society*. Reuters Institute for the Study of Journalism. <https://doi.org/10.60625/risj-5bjv-yt69>
- Spearing, E. R. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, 12(6), Article 242148.
<https://doi.org/10.1098/rsos.242148>