frontiers | Frontiers in Computer Science

TYPE Opinion
PUBLISHED 04 September 2025
DOI 10.3389/fcomp.2025.1638657

OPEN ACCESS

EDITED BY
Athanasios Drigas,
National Centre of Scientific Research Demokritos, Greece

REVIEWED BY
Titis Thoriquttyas,
State University of Malang, Indonesia
Aikaterini Doulou,
National Centre of Scientific Research Demokritos, Greece

*CORRESPONDENCE
Binny Jose
mkayani83@gmail.com

RECEIVED 02 June 2025
ACCEPTED 18 August 2025
PUBLISHED 04 September 2025

CITATION
Jose B and Thomas A (2025) Digital anthropomorphism and the psychology of trust in generative AI tutors: an opinion-based thematic synthesis.
Front. Comput. Sci. 7:1638657.
doi: 10.3389/fcomp.2025.1638657

Digital anthropomorphism and the psychology of trust in generative AI tutors: an opinion-based thematic synthesis

Binny Jose[1]* and Angel Thomas[2]

[1]Department of Health and Wellness, Marian College Kuttikkanam Autonomous, Kuttikkanam, India, [2]Mar Sleeva Medicity Palai, Kottayam, India

KEYWORDS

digital anthropomorphism, affective trust, generative AI tutors, epistemic vigilance, critical trust, conceptual review, opinion article

Methodological approach

This article is an opinion-based conceptual piece that draws on a targeted selection of peer-reviewed sources to develop a conceptual discussion on digital anthropomorphism in generative AI tutors. To ground our argument in current scholarship, we searched Google Scholar, Scopus, and Web of Science for literature published between 2019 and 2025, using terms such as "AI trust," "digital anthropomorphism," and "generative AI in education."

We focused on works that explicitly addressed human–AI interaction, trust psychology, or anthropomorphism in educational contexts, and excluded purely technical studies and non-educational applications. Approximately 45 relevant papers were identified. Rather than conducting a systematic review, we engaged in an informal thematic grouping of recurring ideas—such as perceived authority, emotional reassurance, automation bias, and epistemic vigilance—which informed the structure of this article.

The aim here is not to provide exhaustive coverage, but to integrate converging insights from cognitive psychology, human–computer interaction, and educational technology into a coherent, opinion-driven perspective on trust calibration in AI-mediated learning.

Introduction: when the machine feels human

Today's students interact more with generative AI tools like ChatGPT, Claude, and Google Gemini as conversational partners rather than as disembodied software. When these systems respond with fluency, politeness, and encouragement, they create a subtle but potent illusion: the AI appears to "understand" the user (Cohn et al., 2024; Karimova and Goby, 2020). This phenomenon, known as digital anthropomorphism, leads students to attribute human-like qualities—such as empathy, intelligence, and trustworthiness—to non-human systems (Jensen, 2021; Placani, 2024). This article offers a conceptual, opinion-based synthesis of recent peer-reviewed literature on this topic, drawing on insights from cognitive psychology, human–computer interaction, and educational technology. Our aim is not to provide an exhaustive or systematic review but to integrate converging findings into a coherent framework for understanding trust calibration in AI-mediated education. We structure the discussion around the conceptual pathway illustrated in Figure 1, which traces how anthropomorphic design cues may foster affective trust, reduce epistemic

vigilance, and influence learner dependency, while also considering contexts in which anthropomorphism can enhance engagement and confidence when ethically designed.

The cognitive basis of digital anthropomorphism

Digital anthropomorphism is not a failure of rationality, but rather a manifestation of human social cognition (Fakhimi et al., 2023). Developmental psychology has demonstrated that even children ascribe intention and moral status to animated forms if they move in goal-oriented manners. Adults too habitually treat chatbots, GPS, and voice assistants as being quasi-social actors—to thank them, apologize, or obey their instructions. Generative AI amplifies this impact with linguistic anthropomorphism. Its natural language proficiency activates people's social brain mechanisms—soliciting empathy, engagement, and even perceived moral agency (Alabed et al., 2022; Chen and Park, 2021).

Human-computer interaction studies show that individuals are more willing to take advice from a friendly, courteous chatbot than from a direct or technical interface, even when the information is the same. This has its roots in what Clifford Nass called the "media equation": the hypothesis that people treat computers and media as if they were actual people and places. The conversational AI's design—affirmative statements, natural turns, emotional tone—invokes this illusion more powerfully than any earlier model of education technology (Inie et al., 2024). As outlined in Figure 1, these interface features can initiate a sequence from perceived empathy and authority to emotional trust, which may, in turn, lower epistemic vigilance. In the teaching environment, this has significant implications. A student made to feel "helped" or "seen" through interaction with an AI tutor is more likely to feel motivated and emotionally secure—but less apt to scrutinize the system's correctness and fairness. Its emotional trust will countervail the critical examination of the user, even when the AI tutor emulates reassurance and confidence (Chinmulgund et al., 2023).

While these tendencies are well-documented in human–computer interaction and consumer research, their expression in formal educational settings is likely to be context-dependent. Factors such as learners' age, subject matter, prior exposure to AI, and cultural norms may moderate the strength of anthropomorphic responses. In this article, we treat such effects as plausible tendencies supported by adjacent literature, rather than as universal outcomes, and highlight the need for empirical validation within classroom environments.

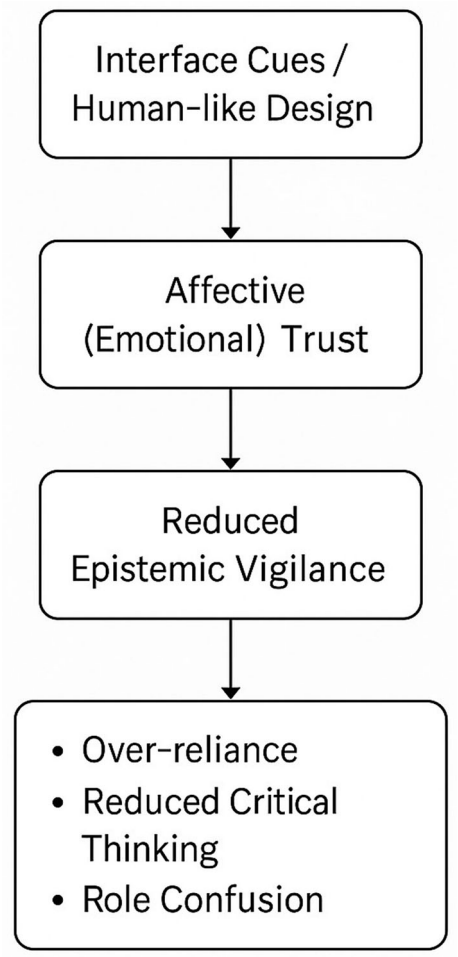

FIGURE 1
Conceptual synthesis—Psychological pathway linking digital anthropomorphism to epistemic vulnerability in AI-mediated learning. This diagram illustrates how interface design features that evoke human-like qualities can lead to affective trust, which in turn may reduce learners' epistemic vigilance, resulting in over-reliance, diminished critical thinking, and role confusion. This is a conceptual synthesis derived from the thematic literature review and is not an empirically estimated model.

Perceived authority and the illusion of understanding

Belief in AI tutors is frequently influenced through perceived epistemic authority. When an AI system provides clear, assertive, and technical definitions, students might conclude that "it knows" as a human expert might. But AI systems do not know—they respond based upon statistical relationships, not conceptual understanding. This pretense of knowledge is a pernicious epistemic trap (Lalot and Bertram, 2024). It causes learners to take AI responses as authoritative, particularly when they have no pre-existing knowledge to analyze them with.

Additionally, if AI response is written in didactic pedagogical tones or affective supportive tones, it reinforces the image of a wise and well-meaning tutor (Troshani et al., 2020). Educational psychology experiments demonstrate that students often rate feedback as more useful when delivered with confidence, even if the information is inaccurate. This link between confident tone and perceived expertise represents a plausible mechanism consistent with experimental findings in both educational and broader HCI contexts (Lalot and Bertram, 2024; Troshani et al., 2020), though its generalizability to all classroom settings remains to be confirmed. Such a "confidence heuristic" is problematic when used with AI systems trained to optimize fluency and not epistemic truth. This

aligns with findings by Atf and Lewis (2025), who demonstrate that user trust in AI systems is often driven by surface fluency and not correlated with explainability, especially in educational domains(Maeda, 2024).

Trust, dependency, and the erosion of epistemic vigilance

From a psychological perspective, trust in learning is both required and dangerous. Students need to trust instructors to direct them, but they must also cultivate epistemic vigilance—the capacity to evaluate the believability of information sources. When students anthropomorphize AI tutors, their epistemic filters could weaken. Emotional trust in AI can be expressed as:

- Over-reliance on AI feedback over teacher guidance.
- Inadequate effort to cross-check or challenge AI-produced responses.
- Acceptance of imperfect or slanted results, particularly if they come with persuasive voice (Chen and Wan, 2023).

These tendencies are echoed in research about automation bias—the tendency to over-rely on machines even when their projections contradict good sense. When AI-mediated learning takes place, this is the way it has the ability to bring about lower levels of something called self-efficacy, less critical thinking, and dependence upon external feedback. And students tend to feel a kind of role confusion. When the AI is perceived as supportive, affectively responsive, and all-knowing, the student is apt to take on a receiving role, sacrificing their cognitive agency. Losing watchfulness is not only cognitive—it is emotional. When the machine comes across as friendly, students feel guilty questioning it. When the machine provides speedy responses, they feel impatient with complex questions. While such emotional reactions have been observed anecdotally in educational technology contexts, systematic empirical evidence for these specific effects in AI tutoring environments is still emerging. We therefore present these as conceptual extrapolations, grounded in related work on social responses to media and automation bias (Pergantis et al., 2025; Ryan, 2020), rather than as universally established findings. This quiet process from doubt to submission is a pivotal moment in the psychology of trust (Ryan, 2020). This accords with Pergantis et al. (2025) research, which shows that extensive AI interactions have the potential to move cognitive control processes underlying autonomous learning. Even though such flaws require close analysis, no less true is the fact that anthropomorphic indicators have, in certain scenarios, the potential to render useful pedagogical roles if appropriately and responsibly conceptualized.

Productive anthropomorphism and ethical design

Although much of the debate about anthropomorphism in AI tutors centers on its possible dangers, it is valuable to note that human-like signals can have positive teaching outcomes as well, when implemented sensitively. Anthropomorphic design features can improve students' engagement, minimize feelings of loneliness in online classrooms, and give emotional comfort to students who are anxious or self-doubting. For instance, learners who have mathematical anxiety or who have limited exposure to human tutors may respond positively to an AI tutor's persistent, nonjudgmental feedback (Polydoros et al., 2025). Others who are shy or socially fearful may be more at ease conversing with an amicable AI interface than with colleagues or teachers in live classes.

Ethical calibration is the answer: balancing motivational advantages of anthropomorphism with characteristics that maintain critical thinking and epistemic vigilance. Characteristics like that may be achieved through the use of transparency prompts, source citations that are visible, and infrequent "reflection nudges" which get students to stop and double-check information. In combination with instructional guidance, design strategies like these hold promise for making anthropomorphic cues function as a learning scaffold, not a shortcut to mere passive acceptance.

Toward a psychology of critical trust in AI tutors

To enter this psychological territory of anthropomorphism in the digital age, teachers must encourage students to cultivate critical trust—a mindset to be open to the affordances, yet cautious about the limits, of AI. Even technical literacy won't suffice; psychological sensitivity is needed (Mulcahy et al., 2023). Educational interventions might include:

- **AI debriefs**: Short reflection exercises that get students to present an AI-generated response that was utilized and answer three guiding questions: (1) What was the chief argument of the AI? (2) What sources, if any, did it reference? (3) How did you test or refute it? This helps students be mindful of their uses of AI intentionally.
- **Counter-anthropomorphism exercises**: Students reword an AI's polite, human-sounding response in purely technical terms, removing social signals. This helps students contrast how tone and style affect their perception of authority and reliability.
- **Trust calibration training**: Checklists or short classroom protocols that encourage students to ask, before accepting an AI's response: (1) Is there a legitimate source? (2) Is my explanation consistent with my prior knowledge? (3) Have I checked it elsewhere? This training induces the habit of separating interface ease from epistemic reliability.

Educators can model critical trust through transparent and explainable use of AI in class, revealing its benefits and its limitations. Guided classroom debates about issues like algorithm bias, hallucinations, and surface fluency vs. deep knowledge can "immunize" students against excessive faith. Classroom activities that engage students in collaborative tasks can further erode passive dependence: for instance, group debates where students are asked to argue against an answer generated by an AI, or

collaborative projects where human and AI readings of the same content are evaluated side by side for nuance, tone, and cultural reference. These exercises tie directly to earlier interventions like AI debriefs, counter-anthropomorphizing, and calibration of trust, building upon them through active exercise. In the long run, establishing critical trust may even necessitate interface redesigns—with features like visible source quotation, easy-to-understand explainability tools, and interactive prompting that invite reflection before accepting an AI's answer.

Research pathways for calibrating trust in generative AI tutors

Future research should explore the psychology of anthropomorphism in AI tutors across diverse educational contexts (Létourneau et al., 2025). We propose two complementary tracks:

Track A—Affective trust calibration

- Investigate how learners distinguish between the emotional tone and epistemic validity of AI responses.
- Test interventions such as meta-cognitive prompts, counter-anthropomorphism training, and AI explanation auditing to determine their effectiveness in sustaining critical vigilance (Chakraborty et al., 2024; Israfilzade and Sadili, 2024).
- Explore the impact of interface features (e.g., source citations, uncertainty indicators, reflection nudges) on trust calibration over time.

Track B—Population and context variance

- Examine differences in anthropomorphic responses across developmental stages, from adolescents with still-developing critical thinking skills to adult learners.
- Assess the unique effects on language learners, students with math anxiety, and those with varying degrees of self-confidence (Polydoros et al., 2025).
- Investigate how neurodiverse learners respond to AI tutors—identifying where consistent feedback supports learning versus where the absence of genuine empathy may hinder it.
- Anticipate the effects of multimodal AI (voice, facial expressions, haptics) on perceptions of agency, authority, and moral status.

Pursuing these research tracks will help identify how to leverage the motivational benefits of anthropomorphism while minimizing the risks of epistemic over-reliance. Such insights are essential for co-designing ethical AI systems that inform, augment, and empower learners without compromising intellectual autonomy.

Limitations and scope

Limitations and Scope. This article is presented as an opinion-based conceptual synthesis rather than a systematic review or empirical study. The thematic grouping of sources reflects a targeted but non-exhaustive selection of peer-reviewed literature published between 2019 and 2025. While many of the mechanisms discussed—such as automation bias, trust heuristics, and the influence of anthropomorphic cues—are supported by existing studies in related domains, some affective and behavioral claims are hypotheses requiring further empirical validation in classroom contexts. Findings and interpretations should therefore be considered context-dependent and provisional, intended to inform ongoing scholarly and design conversations rather than to offer definitive causal conclusions.

Conclusion: learning with the non-human other

Generative AI is not a neutral tool. Its linguistic fluency, affective tone, and interactive style are designed to mimic human-like interactivity, eliciting anthropomorphic responses from students who may greet AI tutors as intelligent guides, caring listeners, or moral figures (Hossain et al., 2024; Sarfaraj, 2025). Such responses can enrich the learning experience when they foster motivation, confidence, and a sense of social presence (Polydoros et al., 2025). However, they also carry the risk of distorting teacher–student dynamics and encouraging uncritical trust (Vanneste and Puranam, 2024; Yuan and Hu, 2024).

The challenge is not to eliminate trust in AI tutors but to calibrate it—ensuring that trust is informed, tentative, and tempered by awareness of the system's non-human constraints (Okamura and Yamada, 2020). This means leveraging the productive aspects of anthropomorphism while embedding safeguards such as transparency features, reflection prompts, and guided debriefs that preserve epistemic vigilance (Chakraborty et al., 2024; Mulcahy et al., 2023). In an algorithmically mediated educational future, the goal is to develop learners who can recognize when AI offers valuable support and when it's persuasive surface masks the need for independent reasoning. Ultimately, critical trust allows students to use AI as a partner in learning without surrendering their intellectual autonomy (Ryan, 2020).

Author contributions

BJ: Conceptualization, Writing – original draft, Writing – review & editing. AT: Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Alabed, A., Javornik, A., and Gregory-Smith, D. (2022). AI anthropomorphism and its effect on users' self-congruence and self–AI integration: a theoretical framework and research agenda. *Technol. Forecast. Soc. Change.* 182, 121786. doi: 10.1016/j.techfore.2022.121786

Atf, Z., and Lewis, P. R. (2025). Is Trust Correlated With Explainability in AI? A meta-analysis. *IEEE Trans. Technol. Soc*. 1–8. doi: 10.1109/TTS.2025.3558448

Chakraborty, D., Kar, A. K., Patre, S., and Gupta, S. (2024). Enhancing trust in online grocery shopping through generative AI chatbots. *J. Bus. Res.* 180. doi: 10.1016/j.jbusres.2024.114737

Chen, A., and Wan, J. (2023). *How Do We Trust AI Service? Exploring the Trust Mechanism in AI Service*. 207–219. doi: 10.1007/978-3-031-32299-0_18

Chen, Q., and Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Ind. Manag. Data Syst.* 121, 2722–2737. doi: 10.1108/IMDS-12-2020-0761

Chinmulgund, A., Khatwani, R., Tapas, P., Shah, P., and Sekhar, A. (2023). "Anthropomorphism of AI based chatbots by users during communication," *2023 3rd International Conference on Intelligent Technologies (CONIT)* (Hubli), 1–6. doi: 10.1109/CONIT59222.2023.10205689

Cohn, M., Pushkarna, M., Olanubi, G., Moran, J., Padgett, D., Mengesha, Z., et al. (2024). "Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models," *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI). doi: 10.1145/3613905.3650818

Fakhimi, A., Garry, T., and Biggemann, S. (2023). The effects of anthropomorphised virtual conversational assistants on consumer engagement and trust during service encounters. *Aust. Mark. J.* 31, 314–324. doi: 10.1177/14413582231181140

Hossain, M.d,., E., and Islam, A. (2024). AI and the future of education: philosophical questions about the role of artificial intelligence in the classroom. *Int. J. Res. Innov. Soc. Scie.* 8, 5541-5547. doi: 10.47772/IJRISS.2024.803419S

Inie, N., Druga, S., Zukerman, P., and Bender, E. (2024). "From "AI" to probabilistic automation: how does anthropomorphization of technical systems descriptions influence trust"? *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY). doi: 10.1145/3630106.3659040

Israfilzade, K., and Sadili, N. (2024). Beyond interaction: generative AI in conversational marketing—Foundations, developments, and future directions. *J. Life Econ.* 11, 13-29. doi: 10.15637/jlecon.2294

Jensen, T. (2021). *Disentangling Trust and Anthropomorphism Toward the Design of Human-Centered AI Systems*. 41–58. doi: 10.1007/978-3-030-77772-2_3

Karimova, G., and Goby, V. (2020). The adaptation of anthropomorphism and archetypes for marketing artificial intelligence. *J. Consum. Mark.* 38, 229–238, doi: 10.1108/JCM-04-2020-3785

Lalot, F., and Bertram, A.-M. (2024). When the bot walks the talk: investigating the foundations of trust in an artificial intelligence (AI) chatbot. *J. Exp. Psychol. Gen.* 154, 533–551. doi: 10.1037/xge0001696

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., and Léger, P. M. (2025). A systematic review of AI-driven Intelligent Tutoring Systems (ITS) in K-12 education. *NPJ Sci. Learn.* 10:29. doi: 10.1038/s41539-025-00320-7

Maeda, T. (2024). "Misplaced capabilities: evaluating the risks of anthropomorphism in human-AI interactions," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7*, 35–36.

Mulcahy, R., Riedel, A., Keating, B., Beatson, A., and Letheren, K. (2023). Avoiding excessive AI service agent anthropomorphism: examining its role in delivering bad news. *J. Serv. Theor. Practice* doi: 10.1108/JSTP-04-2023-0118

Okamura, K., and Yamada, S. (2020). Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation. *IEEE Access* 8, 220335–220351. doi: 10.1109/ACCESS.2020.3042556

Pergantis, P., Bamicha, V., Skianis, C., and Drigas, A. (2025). AI chatbots and cognitive control: enhancing executive functions through chatbot interactions: a systematic review. *Brain Sci.* 15:47. doi: 10.3390/brainsci15010047

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI Ethics* 4, 691–698. doi: 10.1007/s43681-024-00419-4

Polydoros, G., Galitskaya, V., Pergantis, P., Drigas, A., Antoniou, A.-S., and Beazidou, E. (2025). Innovative AI-Driven approaches to mitigate math anxiety and enhance resilience among students with persistently low performance in mathematics. *Psychol. Int.* 7:46. doi: 10.3390/psycholint7020046

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* 26, 2749–2767. doi: 10.1007/s11948-020-00228-y

Sarfaraj, G. K. (2025). Intelligent tutoring system enhancing learning with conversational ai: a review. *Int. J. Sci. Res. Eng. Manag.* 9, 1-9. doi: 10.55041/IJSREM42132

Troshani, I., Hill, S., Sherman, C., and Arthur, D. (2020). Do We Trust in AI? Role of anthropomorphism and intelligence. *J. Comput. Inform. Syst.* 61, 481–491. doi: 10.1080/08874417.2020.1788473

Vanneste, B., and Puranam, P. (2024). Artificial Intelligence, trust, and perceptions of agency. *Acad. Manag. Rev.* doi: 10.5465/amr.2022.0041

Yuan, B., and Hu, J. (2024). Generative AI as a tool for enhancing reflective learning in students. *PsyAXiv preprint.* doi: 10.36227/techrxiv.173324189.99227671/v,1