PUBLISHED BY
JR EDUCATION

Available online at http://ijite.jredu.id

International Journal of Information Technology and Education (IJITE) 5 (1), (December 2025) 139 - 160

International Journal of Information Technology and Education (IJITE)

http://ijite.jredu.id

Managing Anthropomorphism in Student–AI Interaction

Rena Alasgarova[1]*, Jeyhun Rzayev[2]

[1]Baku-Oxford School, Azerbaijan
[2]ADA University, Azerbaijan

*Corresponding author: rena.alasgarova@bakuoxfordschool.net

ARTICLE INFO
Article history:
Received: October 01, 2025; Received in revised form: November 11, 2025; Accepted: December 07, 2025;
Available online: December 07, 2025;

ABSTRACT

Anthropomorphism has become increasingly salient in educational contexts with the rise of large language models (LLMs). While prior research has examined trust and usability in educational AI, fewer studies have explored how students' everyday interactions with AI systems reflect persistent anthropomorphic interpretations and their implications for epistemic vigilance. This qualitative study bases on 15 focus group discussions conducted with 176 students aged 11–17 from nine secondary schools in Azerbaijan. Focus groups were organized by age cohort and examined students' experiences with AI-based educational tools, including translation software, writing assistants, and chatbot-style tutors. Data were analyzed using inductive thematic analysis to identify recurring patterns of linguistic attribution, emotional engagement, and perceived intentionality. Across age groups, students frequently described AI systems using mental-state language (e.g., "knows," "understands," "wants"), even when explicitly acknowledging that AI lacks consciousness or emotions. A recurring pattern was identified in which students simultaneously demonstrated declarative awareness of AI's non-human nature while interacting with it as if it possessed beliefs or intentions, a phenomenon conceptualized here as reverse false belief. Younger students displayed stronger emotional reliance and anthropomorphic framing, while older students paired such framing with explicit disclaimers. The findings highlight a persistent gap between students' conceptual knowledge about AI and their intuitive interactional responses. Anthropomorphic engagement appears to support motivation and emotional comfort while also posing risks for epistemic vigilance. While grounded in an Azerbaijani case study, the analysis offers insights relevant to broader educational debates on AI design, pedagogy, and policy.

Keywords: AI in education, AI literacy, AI trust, anthropomorphism, philosophy of mind, theory of mind.

INTRODUCTION

Managing Anthropomorphism in Student–AI Interaction

The emergence of large language models (LLMs) and related artificial intelligence (AI) systems has introduced a profound shift in how knowledge is accessed, mediated, and produced. Systems such as OpenAI's ChatGPT, Google's Gemini, and Claude can generate fluent, contextually relevant responses across a wide spectrum of domains, from creative writing to technical problem-solving. In educational contexts, these capabilities have been heralded as opportunities for personalized learning, differentiated instruction, and immediate feedback at scale (Holmes et al., 2021). However, alongside these possibilities lies a persistent and underexamined phenomenon: the tendency of both students and educators to interpret AI systems as if they possessed human-like mental states, intentions, and moral agency. This phenomenon called anthropomorphism is not new, but the realism and responsiveness of contemporary AI make it more pervasive, more convincing, and potentially more consequential than ever before.

The risk of anthropomorphizing the non-human has been recognized for millennia. As early as the 6th century BCE, the Greek philosopher Xenophanes criticized the human tendency to project our own characteristics onto the divine. He observed that "Ethiopians say that their gods are snub-nosed and black; Thracians that they are pale and red-haired… if horses or oxen or lions had hands… horses would draw the forms of the gods like horses, and oxen like oxen" (Xenophanes, c. 530 BCE/1983, DK 21B15–16). His remark captures a universal cognitive habit: we interpret the unfamiliar by analogy to ourselves. In the context of modern AI, this same tendency leads users to attribute understanding, intentionality, or moral agency to statistical models that do not and cannot possess them. The philosophical continuity between Xenophanes' critique and contemporary warnings about anthropomorphic AI is striking: both caution against mistaking simulation for reality.

Modern science and philosophy have periodically amplified and problematized this tendency. Descartes' mechanistic view of animals as complex automata in 1641 and 1984 (Thomas, 2020) and La Mettrie's provocative claim in L'homme machine (originally published in 1947) that humans themselves might be conceived as machines framed cognition in ways that invite sometimes misleading comparisons between biological and mechanical processes (de Laguna, 1914). In the 20th century, Turing shifted attention from inner mentality to outward performance with the imitation game, encouraging evaluative focus on conversational behavior rather than subjective experience (Turing, 1950). The public's response to Weizenbaum's ELIZA program in the 1960s offered an early demonstration of what later became known as the ELIZA effect, the tendency to unconsciously attribute understanding, emotions, or intentions to a computer program based solely on its conversational output (Weizenbaum, 1976). Although ELIZA was a simple pattern-matching dialogue system with no memory or reasoning capacity, many users experienced it as a sympathetic interlocutor.

With LLMs, the behavioral realism has increased dramatically; their outputs are richer, faster, and more context-aware than earlier systems, thereby magnifying the cognitive pull toward anthropomorphism.

Anthropomorphism is also deeply rooted in human cognition. Evolutionary accounts suggest that over-attributing agency to our environment may have been adaptive, favoring survival in ambiguous contexts (Epley et al., 2007). In human–computer interaction, the same perceptual and inferential mechanisms that evolved for navigating social life can lead users to perceive a machine as empathetic, fair, or self-aware when it merely simulates such traits through patterned outputs (Nass & Moon, 2000). In education, this matters for both trust and authority. If a student believes that an AI tutor "understands" them in a human-like sense, they may rely on it uncritically, accept its outputs without verification, or form attachments that reshape expectations of teachers and peers. Teachers, too, can slip into anthropomorphic talk, which may tacitly legitimize inaccurate mental models of how AI works. The dynamics of anthropomorphism in educational settings are also shaped by cultural norms, policy frameworks, and technological infrastructures. International examples show that while the tendency to attribute mental states to AI is widespread, its expression is modulated by local pedagogical traditions and broader societal attitudes toward technology (Festerling & Siraj, 2022).

This paper focuses on the educational dimension of anthropomorphism in AI, with particular emphasis on the philosophical, cognitive, and pedagogical mechanisms that sustain it. While there is a growing body of research on AI literacy, which is often defined as the skills and dispositions required to engage critically with AI systems (Long & Magerko, 2020), far less attention has been paid to the specific ways anthropomorphic interpretations shape and sometimes distort those skills in practice. Much of the existing work emphasizes what AI can do and how to use it safely; comparatively little investigates why students and teachers so readily ascribe mental states to AI, even when they can articulate that AI is algorithms. This paper argues that anthropomorphism is not merely a by-product of poor understanding; it is a structural feature of human cognition and human–AI interaction that must be addressed directly through design, pedagogy, and policy.

The argument proceeds in the following parts and makes four contributions. First, building on cognitive psychology and developmental research, we explain anthropomorphism through Theory of Mind (ToM), highlighting how minimal cues such as turn-taking, self-reference, apparent coherence activate the same mentalizing processes used for human interlocutors (Premack & Woodruff, 1978; Wimmer & Perner, 1983; Epley et al., 2007). We introduce the idea of a reverse false-belief phenomenon in human–AI interaction: users may explicitly know that AI has no beliefs, yet behave as though it does, interpreting errors as intentional deception or confusion.

Second, the research connects these cognitive dynamics to philosophy of mind, revisiting behaviorism and functionalism (Skinner, 1953; Putnam, 1975), Searle's critique via the Chinese Room (1980), and Dennett's intentional stance (1987). We argue that while the intentional stance can be a useful predictive strategy, in classrooms it easily ossifies into ontological belief ("the AI understands me") in the absence of explicit literacy framing.

Third, we turn to educational practice, drawing on student focus groups from Azerbaijan alongside international examples. We show how anthropomorphic language emerges in routine interactions, how it correlates with over-reliance on AI outputs, and how simple framing interventions (e.g., "AI doesn't think. It calculates") can shift students toward more critical engagement.

Fourth, we propose a practical framework for AI literacy tailored to the anthropomorphism challenge integrating reflective dialogue, error-analysis tasks, and design-aware pedagogy, while the

next section extends these insights to AI development and policy, arguing for anthropomorphism-aware design patterns and curricular standards. The research concludes by situating the role of education within wider societal debates about AI trust, governance, and democratic resilience.

In this narrative, we move beyond description to consider the implications for society of how today's learners come to conceptualize AI. Students in present-day classrooms are the developers, policymakers, and voters of tomorrow. If they leave school with the tacit belief that AI systems are intentional, empathetic agents, this will shape public discourse and regulatory expectations potentially misallocating responsibility when AI systems fail and over-inflating confidence when they appear to succeed. Conversely, if students acquire robust habits of critical anthropomorphism recognizing the inevitability of human-centered metaphors while keeping them in check, they can harness the advantages of AI without sacrificing epistemic vigilance or civic judgment.

Aims and Research Questions

The present article is conceptually driven and does not seek to test hypotheses or establish causal relationships. Instead, it integrates philosophical analysis with qualitative classroom material to explore how anthropomorphism operates in contemporary educational uses of AI. Qualitative focus group data are used as illustrative and analytical resources to examine how established cognitive and philosophical frameworks manifest in everyday student–AI interaction.

The analysis is guided by the following exploratory questions:

1) How do secondary-school students linguistically and conceptually describe AI systems in educational contexts?
2) In what ways do students attribute mental states, intentions, or agency to AI, even when explicitly acknowledging its non-human nature?
3) How can Theory of Mind and philosophy of mind frameworks help explain these interpretive patterns?
4) What pedagogical and policy implications follow from these observed tendencies for AI literacy and classroom practice?

Methodologically, the paper combines conceptual synthesis with practice-proximal observations. These observations complement international reports from contrasting contexts (e.g., Scandinavia, East Asia, North America), underscoring both the universality of the cognitive tendency and the variability introduced by local culture and policy.

Recent work on generative AI in education distinguishes between demonstrated learning affordances and growing concerns about trust calibration and overreliance. A comprehensive commentary by Giannakos et al. (2025) synthesizes emerging evidence across educational contexts, highlighting that while LLMs can enhance feedback, self-regulation, and engagement, they also introduce risks related to uncritical trust, reduced epistemic vigilance, and poorly understood system authority.

Thus, the central claim of the study is that mitigating harmful anthropomorphism is not a one-time instructional task but an ongoing societal responsibility. It requires coordination among teachers who design learning activities, developers who shape interface cues and system disclosures, and policymakers who set literacy standards and accountability frameworks. The contribution of this paper is to provide an integrated account of how that coordination can be achieved, and why education is the pivotal arena in which it must begin.

LITERATURE REVIEW

Theory of Mind (ToM) and Anthropomorphism in AI

ToM refers to the capacity to attribute mental states such as beliefs, desires, intentions, and emotions to oneself and to others, and to recognize that these states can differ between individuals (Premack & Woodruff, 1978). It is a cornerstone of social cognition, enabling prediction and interpretation of behavior in complex social environments. In developmental psychology, the emergence of ToM is often studied through false-belief tasks, which assess whether a child can understand that another person may hold a belief about the world that is incorrect (Wimmer & Perner, 1983). Passing such tasks is typically taken as evidence that the child can model another's mind as distinct from their own.

The connection between ToM and anthropomorphism arises since the same cognitive machinery that evolved to interpret other humans, and, in some contexts, animals, can be applied to artificial agents. When an AI system produces output that appears contextually appropriate, responsive, or emotionally attuned, users may automatically attribute underlying mental states, even when they know the system lacks consciousness. This is not a purely rational process but an automatic, often unconscious application of social-cognitive heuristics (Epley et al., 2007). The result is a tendency to treat conversational AI as if it were a genuine social partner rather than a predictive language model.

Recent research confirms that theory-of-mind–related attributions can be triggered by surprisingly minimal cues. For example, Cohn et al. (2024) found that even the use of a single pronoun like "I" in responses can lead users to perceive an AI as self-aware, despite knowing it is an automated system. In educational settings with large language models, such cues as turn-taking, personal pronouns, reference to prior conversation, and adaptive tone create abundant affordances for anthropomorphic thinking, reinforcing intuitive engagement even among informed users.

The analogy to false-belief tasks is instructive here. In human development, a child who understands that another person can hold an incorrect belief demonstrates a separation between appearance and reality. In human–AI interaction, however, users may fail a kind of reverse false-belief task: they may understand abstractly that the AI does not have beliefs, yet behave as though it does, interpreting its errors as intentional deception ("the AI lied") or its corrections as signs of learning. This suggests that anthropomorphic responses are not entirely under conscious control, and that metacognitive awareness alone is insufficient to eliminate them.

Cross-disciplinary work in social robotics offers further insights. Studies with humanoid robots (e.g., Breazeal, 2003; Dautenhahn, 2007) demonstrate that embodied cues, such as gaze direction, gesture, and facial expression, can amplify ToM activation. While most LLMs lack physical embodiment, their linguistic embodiment, i.e., the capacity to inhabit social roles through language, can be equally potent. When ChatGPT adopts the persona of a "tutor" or "mentor," users respond as though it occupies that social role, adjusting their behavior and expectations accordingly.

The implications for education are significant. ToM is not merely a background cognitive capacity but is actively engaged in every conversational exchange with AI. If unexamined, this engagement can lead to misplaced trust and reduced critical scrutiny of AI-generated information. Conversely, if harnessed deliberately, it can be used to design interventions that help students

differentiate between simulation and genuine understanding. Such interventions might include meta-dialogues in which the AI explicitly describes its own limitations, or structured activities where students must test and falsify the outputs produced by AI. These approaches acknowledge the inevitability of anthropomorphic cognition while working to channel it into more critical and informed interaction.

Understanding how ToM operates in human–AI interaction lays the groundwork for the philosophical analysis that follows, in which questions of mind, consciousness, and intentionality are brought to bear on the challenge of AI literacy in education. However, while Theory of Mind explains how humans attribute mental states often leading to anthropomorphism in AI use, the philosophy of mind asks a deeper question: what a mind actually is, and whether such a thing could exist in artificial systems at all. In other words, ToM is a psychological account of mental state attribution, whereas philosophy of mind is a conceptual and metaphysical investigation into the nature and possibility of minds.

Philosophy of Mind and the Question of Artificial Understanding

The philosophy of mind provides a rich conceptual framework for examining anthropomorphism in AI, particularly in relation to questions of consciousness, intentionality, and understanding. At its core, the field asks what it means to have a mind and whether such a phenomenon can, in principle, be realized in non-biological systems. This inquiry has direct relevance to education: if educators and students implicitly or explicitly adopt certain philosophical positions, these assumptions can influence how they interact with AI systems and how they interpret the nature of AI-generated knowledge.

Two positions have historically shaped these debates: behaviorism and functionalism. Behaviorism, in its strictest form, defines mental states solely by observable behavior, without making claims about internal subjective experience (Skinner, 1953). From this perspective, if an AI system behaves as though it understands, it can be treated as if it understands. While few philosophers today endorse pure behaviorism, its legacy persists in human–AI interaction: performance is often taken as evidence of understanding, particularly in contexts where outputs are immediate, coherent, and socially appropriate.

Functionalism, by contrast, defines mental states by their causal roles, how they relate to inputs, outputs, and other mental states rather than by their material substrate (Putnam, 1975). This opens the theoretical possibility that a non-biological system could have genuine mental states if it realizes the same functional organization as a human mind. In educational contexts, this functionalist intuition can subtly legitimize anthropomorphism: if an AI appears to process information, form "beliefs," and generate "conclusions," it is tempting to think of it as having a mind in the same sense that humans do.

Critics of this view, most famously John Searle (1980), have argued against the sufficiency of functional equivalence for genuine understanding. In his Chinese Room argument, Searle imagines a person manipulating Chinese symbols according to a rulebook without understanding their meaning. To an outside observer, the system's outputs may appear indistinguishable from those of a fluent Chinese speaker, but no understanding is present, only symbol manipulation. By analogy, Searle contends, even the most sophisticated AI lacks semantic understanding; it merely manipulates symbols according to statistical patterns.

Daniel Dennett (1987) offers a counterpoint through his concept of the intentional stance: a predictive strategy in which we treat an entity as if it had beliefs and desires in order to explain and

anticipate its behavior. From this perspective, anthropomorphism is not necessarily a mistake but a pragmatic tool so long as we remain aware that the stance is instrumental rather than descriptive of an underlying reality. In education, however, this distinction can be difficult to maintain, particularly for younger learners or those without explicit training in critical AI literacy. When the intentional stance hardens into an ontological claim "the AI understands me", misplaced trust can follow.

Recent philosophical debates extend these questions into the realm of synthetic phenomenology, the hypothetical engineering of subjective experience in artificial systems (Reggia, 2013; Gamez, 2018). Some researchers argue that modelling certain features of consciousness could enhance human–AI collaboration, while others warn that doing so risks deepening anthropomorphic biases and creating moral confusion about the rights and responsibilities of machines. Even without actual synthetic phenomenology, the increasingly sophisticated simulations of empathy and self-awareness in AI blur the line between performance and possession of mental states.

For educational practice, the philosophical lesson is clear: whether or not AI systems can possess minds in any robust sense, their perceived mentality is an unavoidable factor in human interaction with them. Teachers and curriculum designers who ignore the philosophical underpinnings of these perceptions risk reinforcing uncritical anthropomorphism. Conversely, incorporating philosophy of mind into AI literacy curricula can equip students to critically interrogate their own cognitive and emotional responses to artificial agents.

METHOD

This study follows a qualitative, interpretive design embedded within a conceptually driven inquiry. Rather than testing predefined hypotheses, the methodology supports theory-informed exploration of how students interpret and describe AI systems in educational settings. Empirical material is used to illuminate and refine conceptual arguments concerning anthropomorphism, Theory of Mind, and epistemic vigilance, not to establish statistical generalizability or causal effects.

Participants

The Azerbaijani case study draws on a qualitative research design centered on focus group discussions with students. Conducted between February and June 2025, the study involved one private school and eight public schools, comprising 15 focus groups and a total of 176 students aged 11 to 17. The AI technologies in use included translation software, writing assistance tools, and early-stage chatbot tutors modelled on commercial LLMs.

Inclusion criteria were: (a) current enrollment in lower or upper secondary education and (b) prior or ongoing use of AI-based tools for educational purposes (e.g., writing assistance, translation, or AI-supported tutoring). Students who did not use AI tools or who did not provide informed consent were not included. No additional exclusion criteria were applied. Participants were grouped by age (11–13 and 15–17) in order to facilitate age-appropriate discussion and to explore potential generational differences in students' interpretations of and engagement with AI systems.

A focus group methodology was selected as it is particularly well suited to exploring shared meanings, interpretive language, and socially mediated perceptions, which are central to the study's aims. As Creswell and Creswell (2022) emphasize, focus groups are appropriate for qualitative inquiry

when the objective is to understand how participants collectively construct meaning around a phenomenon and adapt their responses in dialogue with peers.

In the context of student–AI interaction, anthropomorphic interpretations frequently emerge through conversational normalization (e.g., joking, agreement, disagreement), making focus groups an effective method for capturing not only individual attitudes, but also socially patterned ways of speaking about AI. The group format enabled participants to respond to, elaborate on, or challenge one another's characterizations of AI, revealing how anthropomorphic framings circulate within classroom peer cultures.

Ethical Considerations

Official permission to conduct the study was obtained from the State Agency for Preschool and General Education of the Republic of Azerbaijan. Ethical considerations were strictly observed in accordance with the APA 7 ethical standards for research involving human participants. Participants were informed about the purpose of the study, the voluntary nature of their participation, and their right to withdraw at any time without negative consequences.

Written informed consent was obtained for all participants through appropriate institutional procedures, with particular attention to protecting minors. All data were anonymized during transcription, and no identifying information was retained. Research materials were stored securely and accessed only by the research team (American Psychological Association, 2019).

Research Instrument

Data were generated using a semi-structured focus group guide developed by the research team to elicit students' perceptions, language use, emotional responses, and interpretive framings of AI in educational contexts. The guide was informed by established literature on anthropomorphism, Theory of Mind, epistemic vigilance, and trust in AI-supported educational technologies.

The instrument was structured into four thematic sections: (1) background and frequency of AI use; (2) perceptions of AI's "intelligence" and understanding; (3) emotional and cognitive responses to AI interaction; and (4) the perceived role of AI in learning and educational outcomes. Open-ended questions were designed to prompt reflection, encourage narrative responses, and enable participants to elaborate on their experiences in their own terms. Follow-up prompts were used flexibly to clarify meaning and deepen discussion where appropriate.

In line with qualitative research standards, the focus group guide was not intended as a standardized or psychometrically validated instrument. Instead, validation was established through conceptual, procedural, and iterative researcher validation, as recommended by Creswell and Creswell (2022) and Miles et al. (2020).

First, conceptual validation was achieved by grounding all questions in existing theoretical and empirical literature on anthropomorphism, Theory of Mind, and human-AI interaction, ensuring clear alignment between the research aims, guiding questions, and the instrument content.

Second, expert validation was conducted through collaborative review by the author team. Questions were examined for conceptual clarity, age appropriateness, and relevance to the study's guiding questions, and were refined through iterative discussion prior to data collection.

Third, procedural validation occurred during early focus group sessions, where the researchers monitored whether questions elicited meaningful, comprehensible, and theoretically relevant

responses from participants. Minor adjustments to wording and prompt sequencing were made where necessary to enhance clarity and engagement, while preserving the conceptual structure of the guide.

RESULTS AND DISCUSSION

Data Analysis

Data were analyzed using manual, reflexive thematic analysis. Following Miles at el.'s (2020) approach to qualitative coding, the analysis proceeded through iterative cycles of coding, categorization, and thematic refinement. Initial open coding focused on identifying recurring linguistic patterns, metaphors, evaluative terms, and expressions of intentionality or agency attributed to AI systems.

Coding and theme development were conducted collaboratively by the research team, with ongoing discussion used to compare interpretations, resolve ambiguities, and refine analytic categories. This iterative, dialogic process enhanced analytic credibility by ensuring that emerging themes were grounded in the data and conceptually consistent with the study's theoretical framework. Manual coding enabled close engagement with participants' language and facilitated sensitivity to nuance, context, and age-related variation.

Analytic Orientation

Data analysis focused on identifying recurring linguistic patterns, metaphors, and interpretive framings through which students described AI systems in educational contexts. Particular attention was paid to (a) attribution of understanding and intention, (b) emotional and relational language, and (c) moments of tension between students' explicit knowledge of AI as a non-human system and their intuitive interactional responses. Themes were developed inductively through iterative manual coding across focus group transcripts and refined through constant comparison within and between age groups. Analytic emphasis was placed on clustering recurring forms of expression rather than on quantifying responses.

Theme 1: Linguistic Attribution of Understanding and Intention

Across all focus groups, students consistently described AI behavior using mental-state language typically reserved for human agents. Such formulations appeared spontaneously in peer discussion and were often taken up and reinforced by other participants, becoming a shared conversational shorthand. Students frequently stated that AI "knows" or "understands" their intentions:

"Sometimes it knows what I mean even if I don't explain it well." (Student, age 13)

"When I write something wrong, it still understands what I wanted to say." (Student, age 15)

"It understands the question even if you ask it badly." (Student, age 12)

Other utterances framed AI as having preferences, evaluations, or directional intent:

"It wants me to do it this way, not the other." (Student, age 14)

"It doesn't like that answer." (Student, age 11)

Although many students later qualified these statements by noting that AI is "just a program," such disclaimers did not prevent the continued use of intentional language in practice. Collectively,

these utterances position AI as an entity capable of grasping meaning and intention rather than as a purely computational tool.

Theme 2: Emotional Engagement and Perceived Relationality

Beyond cognitive attribution, students frequently described AI systems in emotionally relational terms. Many participants highlighted qualities such as patience, tolerance, and availability, often contrasting these attributes with those of human teachers or family members.

Several students emphasized emotional safety and accessibility:

"It never judges you and it always answers, even at night." (Student, age 12)

"You can ask the same thing again and again and it doesn't get annoyed." (Student, age 14)

"With AI you're not embarrassed if you're wrong." (Student, age 13)

Some participants explicitly framed AI as occupying a supportive social role:

"It's like a personal coach for studying." (Student, age 15)

"Sometimes it feels like it cares about helping you." (Student, age 11)

In several accounts, this perceived relationality extended beyond academic activities. One student described using a chatbot to manage a personal conflict:

"It helped me find the right words to talk to my friend after a fight." (Student, age 16)

These narratives indicate that AI was experienced not merely as an information source but as a non-judgmental interlocutor integrated into students' emotional regulation and decision-making practices.

Theme 3: Reverse False-Belief Phenomenon in Student–AI Interaction

A recurrent and analytically distinctive pattern emerged in which students articulated correct declarative knowledge about AI's non-human nature while simultaneously responding to it as though it possessed beliefs, intentions, or strategic motives. This pattern is conceptualized here as a reverse false-belief phenomenon.

Students often described disagreement or conflict with AI outputs using intentional framings:

"It gave me a wrong answer, but it was convincing, so I started arguing with it." (Student, age 15)

"Sometimes it lies, or at least it says things like it wants to trick you." (Student, age 13)

"It insists on the answer even when it's wrong." (Student, age 14)

At the same time, these students frequently demonstrated explicit awareness that AI does not think or feel:

"I know it's not actually thinking, but it feels like it is." (Student, age 16)

The coexistence of explicit knowledge ("it's just a program") and intuitive attribution ("it lies," "it insists") illustrates a reversal of classic false-belief logic. Rather than failing to recognize false beliefs in another agent, students struggle to fully suppress belief attribution toward an entity they know lacks mental states.

Theme 4: Age-Related Variation in Anthropomorphic Framing

Clear age-related differences were observed across themes. Younger students (11–13) were more likely to describe AI as if "someone is behind it" and to express emotional reliance in personal terms:

"It feels like there is a real person answering you." (Student, age 12)

Older students (15–17), by contrast, more frequently paired anthropomorphic language with explicit disclaimers:

"I know it's not human, but you still talk to it like one." (Student, age 17)

"You understand it's not a person, but you react like it is." (Student, age 16)

These patterns suggest that early and sustained exposure to AI may normalize anthropomorphic engagement even as conceptual understanding of AI's limitations develops. See table 1.

Table 1. Initial Codes, Themes, and Representative Quotations

Initial Codes	Theme	Representative Verbatim Quote
"AI knows", "It understands", "It wants me to…"	Linguistic attribution of understanding and intention	"Sometimes it knows what I mean even if I don't explain it well." (Age 13)
Trust, patience, non-judgment, availability	Emotional engagement and perceived relationality	"It never judges you and it always answers, even at night." (Age 12)
Arguing with AI, persuasive errors, lying / tricking	Reverse false-belief phenomenon	"It gave me a wrong answer, but it was convincing, so I started arguing with it." (Age 15)
Emotional reliance vs. explicit disclaimer	Age-related variation in anthropomorphic framing	"I know it's not human, but you still talk to it like one." (Age 17)

Discussion

While anthropomorphism toward AI can be explained through cognitive and philosophical frameworks, its educational implications emerge most vividly in classroom practice. This section examines these manifestations through a combination of original focus group interviews from Azerbaijani schools and comparative insights from other international contexts. Together, these examples illustrate how anthropomorphism shapes student engagement, trust, and critical thinking and how it can persist even in settings where educators explicitly warn against it.

These interactional patterns, documented in the analysis above, have important pedagogical implications. Many went further, expressing strong trust in AI systems, particularly LLM-based chatbots, and naming specific platforms they used. This trust was often rooted in the perception that AI "understands" them, will not judge them, and can be confided in about personal matters. Several participants contrasted this with their relationships at home, noting that they sometimes share thoughts and experiences with AI that they would not share with their parents either because parents "would not understand," because "there are things you cannot tell them," or because "parents are too busy."

Personal narratives were common. In one example, a student recounted using a chatbot to help reconcile with a friend after a fight, explaining that the AI "helped me find the right words to say." Others described AI as a "personal coach" or "personal trainer," turning to it for advice not only on schoolwork but also on daily life decisions. Many students valued the accessibility of AI available "even at night" and willing to answer "hundreds of times" without complaint.

Clear differences emerged between age groups. Younger students (11–13) tended to form stronger emotional attachments to AI, often speaking as though there were "a real person behind it" who provided the answers. For this group, AI was integrated into their school life from the outset of

middle school, making it a routine part of their learning and personal problem-solving. Older students (15–17), who had begun middle school before AI became widely available in Azerbaijan, also expressed trust in AI but were more reflective about what they shared, acknowledging explicitly that "it's not a real person." Older participants sometimes noted differences between themselves and their younger peers, suggesting that early, immersive exposure to AI may strengthen anthropomorphic tendencies and emotional reliance.

Such framing risks reinforcing inaccurate mental models of AI functionality, making it a key consideration for AI literacy interventions. These patterns reflect the same cognitive mechanisms described earlier in the ToM framework: students can state that the AI lacks memory or feelings, yet behave as if it possesses beliefs, intentions, and emotions. In this sense, their interactions echo a reverse false-belief scenario recognizing the absence of genuine understanding in the abstract while responding in practice as though the AI were a sentient collaborator.

These findings suggest that anthropomorphism in AI use is intertwined with trust, emotional support, and the social roles students assign to technology. They also highlight the role of developmental stage and educational context in shaping how students perceive and interact with AI tools.

International Comparisons

These tendencies are not unique to Azerbaijan. Research from multiple national contexts shows that anthropomorphism in educational AI is a robust, cross-cultural phenomenon, shaped by both cognitive and cultural factors. In the United States, experiments with conversational AI systems show that anthropomorphic cues, even as minimal as the system speaking aloud ("speech + text"), can significantly increase perceived sociability, anthropomorphism, and trust (Cohn et al., 2024). While including the first-person pronoun ("I") did enhance perceived accuracy and lower perceived risk, these effects were limited to specific task domains such as medication advice. Ackermann et al. (2025) found that while combining an on-screen avatar with a physical robot did not increase perceived animacy or agency, it did boost students' initial on-task enjoyment. However, higher perceived sociability was associated with lower task performance highlighting how anthropomorphic cues may enhance engagement but potentially undermine learning.

In Finland, primary schools piloting the NAO-based Elias language-learning robot report that it exhibits "endless patience for repetition," adjusts its responses to students' skill levels, and avoids embarrassing learners, thereby fostering a safe and supportive speaking environment (Reuters, 2018). Teachers observed that Elias created a safe, low-pressure environment where students were more willing to speak and make mistakes (Kantola, 2023). Broader meta-analytic findings indicate that while anthropomorphic robot features can boost social trust and engagement, they may simultaneously undermine competency trust, revealing a tension in how students relate to human-like machines (Stower et al, 2021). Belpaeme et al. (2018) underscore the risk of over-ascription in educational HRI: learners may attribute cognitive or emotional capacities to robots beyond their actual capabilities, highlighting the need for careful framing and transparency to support accurate mental models.

These tendencies are not solely the result of technical novelty. A large-scale cross-national study of 508 K–12 teachers in Brazil, Israel, Japan, Norway, Sweden, and the United States found that trust in AI-based educational technologies is significantly shaped by teachers' AI self-efficacy, understanding of AI, and cultural values (Viberg et al., 2024). Higher uncertainty avoidance and long-term orientation

predicted greater trust, while cultural masculinity and collectivism were linked to greater concerns. These results suggest that anthropomorphism's motivational benefits and its potential to mislead are moderated by deeply ingrained socio-cultural dispositions.

Recent empirical work with school-age learners similarly indicates that interaction with AI tutors can increase engagement, motivation, and perceived learning effectiveness, while offering limited opportunities for learners to critically assess the system's underlying reasoning processes (Kestin et al., 2025). However, much of the existing empirical literature on large language models continues to focus on university or adult populations, whereas the present study contributes practice-proximal evidence from secondary-school classrooms, where anthropomorphic framing appears especially persistent.

Pedagogical Risks and Opportunities

From a pedagogical standpoint, anthropomorphism presents both risks and potential benefits. On the risk side, uncritical anthropomorphism can erode epistemic vigilance, the capacity to evaluate the reliability of information sources (Sperber et al., 2010). Students who believe an AI "understands" them may accept incorrect or biased outputs without verification, reducing their capacity for independent judgment. This is a contemporary manifestation of the ELIZA effect (Weizenbaum, 1976), whereby users unconsciously attribute understanding, emotions, or intentions to a computer program based solely on its conversational output. In the Azerbaijani focus groups, younger students in particular described AI tools as confidants, non-judgmental listeners, and sources of personal advice sometimes preferring them to parents or peers for sharing sensitive matters, illustrating how the ELIZA effect persists even when participants know the system is "just a program."

On the opportunity side, anthropomorphic engagement can be harnessed to sustain motivation, particularly for students who might otherwise disengage from abstract or repetitive tasks. In such cases, the challenge for educators is to design interventions that preserve engagement while systematically deconstructing false mental models. One promising approach observed during the Azerbaijani focus groups involved role-switching exercises, where students alternated between interacting with the AI and explaining its limitations to a peer. These meta-communicative tasks appeared to reduce naïve anthropomorphism over time without diminishing enthusiasm for the technology. Comparable priorities are reflected in the UK, where the Department for Education's early-adopters report describes secondary schools integrating AI tools into lesson design and assessment, while underscoring the importance of reflective practices to maintain accuracy and reliability (Department for Education, 2025).

Design and Policy Considerations

The persistence of anthropomorphism in educational settings suggests that such technical design choices as the use of human-like avatars, conversational small talk, and adaptive politeness strategies should be critically examined. While these features may improve short-term usability, they can also strengthen anthropomorphic framing. Some researchers advocate for anthropomorphism-aware design principles that retain conversational clarity but minimize cues that invite unwarranted mental-state attributions (Złotowski et al., 2015).

At the policy level, integrating anthropomorphism awareness into AI literacy curricula is essential. This does not mean banning anthropomorphic metaphors altogether (these can be

pedagogically useful) but rather making them the subject of explicit reflection. Students should be able to recognize when they are adopting an intentional stance toward the AI and to understand that such stances are interpretive strategies rather than ontological claims about the system.

AI Literacy in Practice

Developing AI literacy in educational settings requires more than teaching technical skills; it involves cultivating an informed, critical orientation toward the capabilities and limitations of AI systems. This section outlines practical approaches for fostering AI literacy that specifically address the challenge of anthropomorphism. Drawing on educational theory, classroom practice, and design considerations, it offers a framework that integrates conceptual understanding with active reflection.

Defining AI Literacy in the Anthropomorphic Context

AI literacy has been broadly defined as the knowledge, skills, and dispositions required to understand, use, and critically evaluate AI technologies (Long & Magerko, 2020). In the context of anthropomorphism, this definition must be extended to include the ability to recognize and manage the cognitive biases that lead to attributing human-like qualities to machines. AI literacy, therefore, is not solely about understanding what AI can do; it is equally about understanding what it cannot do, and why human-like behavior does not imply human-like cognition or consciousness.

This expanded definition positions anthropomorphism awareness as a core competency, on par with understanding technical architecture or ethical implications of AI tools. By making this awareness explicit, educators can help students move beyond a surface-level familiarity with AI toward a more sophisticated, reflective engagement.

Pedagogical Strategies for De-Anthropomorphizing AI

Effective AI literacy education in this domain benefits from embedding anthropomorphism-related content into both formal lessons and informal interactions. One approach involves structured "myth-busting" activities, in which students are presented with common anthropomorphic claims ("The AI learns from me like a human would") and then guided to evaluate these claims against technical explanations of how the system actually operates. These activities are most effective when combined with hands-on demonstrations—for example, deliberately prompting an AI system to reveal its lack of memory or inability to reason beyond its training data.

Another strategy is to integrate reflective dialogue into AI-assisted learning activities. In one Azerbaijani pilot, students using an AI translation tool were periodically asked to explain why the system might have made certain translation errors and whether those errors indicated "confusion," "ignorance," or something else entirely. Such prompts encourage students to critically evaluate the temptation to personalize the system's behavior.

International examples reinforce the value of this approach. In Finland, teacher-led debrief sessions following interactions with a humanoid language-learning robot prompted students to identify which aspects of the robot's behavior were programmed responses and which emerged from adaptive algorithms (Vartiainen et al., 2020). This reflective framing did not reduce student engagement but did increase their ability to articulate the robot's limitations in non-anthropomorphic terms.

Designing for Critical Engagement

Pedagogical interventions are most effective when supported by thoughtful design choices in the AI systems themselves. Design elements such as anthropomorphic avatars, conversational small talk, or emotional tone-matching can increase engagement but also risk reinforcing anthropomorphic interpretations. Developers can mitigate this risk by incorporating transparency cues, such as occasional meta-comments about the system's operations ("I do not have personal experiences, but I can retrieve relevant data") or by offering "explain mode" outputs that clarify how a response was generated.

In practice, the most promising designs balance usability with epistemic clarity. For example, AI tutoring systems could allow users to toggle between "concise answer" and "explain reasoning" modes, with the latter making explicit the statistical or algorithmic basis of responses. Such design features align with educational goals by reinforcing the message that AI output is the product of computation, not cognition.

Long-Term Integration into Curricula

Addressing anthropomorphism should not be a one-off exercise. Instead, it should be integrated into AI literacy curricula as an ongoing thread, revisited across different subjects and grade levels. This integration could take the form of cross-disciplinary projects, e.g., comparing historical anthropomorphism of natural forces with contemporary anthropomorphism of AI, or capstone assignments that require students to design their own AI literacy interventions.

Policy support is also critical. Ministries of Education and school boards can promote consistency by embedding anthropomorphism-awareness competencies into national digital literacy frameworks.

Implications for AI Design and Policy

The policy implications discussed below are based on qualitative findings from Azerbaijani secondary school classrooms and are intended to be illustrative rather than prescriptive. While the observed patterns align with international research on anthropomorphism and trust in educational AI, the present evidence reflects a specific cultural, institutional, and developmental context. Accordingly, the recommendations that follow distinguish between empirically observed tendencies within the Azerbaijani case study and broader considerations that may be relevant for other educational systems adopting AI-supported learning tools.

Design Responsibility

AI developers make countless small choices about how systems present themselves: whether to use first-person pronouns, how to phrase disclaimers, whether to simulate human-like pauses or emotions. These design elements can either reinforce or challenge anthropomorphic interpretations. For instance, conversational agents that frequently use "I think" or "I feel" naturally invite users to interpret the system as having beliefs or emotions, even when this language is metaphorical. Conversely, interfaces that avoid unnecessary self-reference, or that intermittently remind users of the system's non-human nature, can subtly recalibrate expectations without significantly impairing usability.

From a policy perspective, design choices carry epistemic and ethical consequences. This aligns with the principle of value-sensitive design (Friedman et al., 2013), which emphasizes embedding moral and societal values into the technological design process. Where anthropomorphism risks

undermining critical engagement or fostering misplaced trust, developers have a duty to mitigate such risks proactively.

Policy Frameworks: International Examples

Several internationally recognized AI governance frameworks set out principles that, while not explicitly targeting anthropomorphism, offer clear guidance for promoting transparency, accountability, and critical engagement in educational AI. These frameworks can be applied in classroom contexts to ensure that AI tools are integrated in ways that foster motivation and learning while maintaining accurate understandings of their non-human nature.

The OECD AI Principles, adopted in 2019 by 46 countries, set out high-level guidelines for trustworthy AI (Organisation for Economic Co-operation and Development [OECD], 2019). These include commitments to transparency, accountability, safety, and respect for human rights and democratic values. Although not focused on anthropomorphism, the transparency provisions require that users understand when and how they are interacting with AI systems, supporting educational practices that clarify AI's non-human nature and foster critical engagement.

The European Union AI Act, set for phased implementation, introduces a tiered risk framework for AI systems, including requirements for transparency and user awareness (European Union, 2024). Systems interacting with humans must disclose their non-human nature "in an intelligible manner" (Art. 50). In educational contexts, compliance could mean that AI tutoring systems must introduce themselves as software and periodically remind students of their operational basis. However, the Act stops short of prescribing the form these disclosures should take, leaving open the possibility of compliance in letter but not in spirit. This creates an opportunity for educational AI design to go beyond minimum legal requirements, embedding recurring transparency cues that actively support AI literacy.

The UNESCO Recommendation on the Ethics of Artificial Intelligence, adopted in 2021 by 193 Member States, emphasizes transparency, accountability, and cultural diversity (UNESCO, 2021). It notes the potential for AI to "affect human cognitive frameworks" and calls for public awareness campaigns and educational interventions to ensure AI's role is understood accurately. While it does not single out anthropomorphism, its broad framing of cognitive influence clearly encompasses it, making it a valuable reference point for designing AI literacy strategies that address anthropomorphic misconceptions.

Together, these three frameworks provide complementary levers: the OECD AI Principles articulate high-level values, the EU AI Act sets enforceable legal standards, and the UNESCO Recommendation embeds ethical and cultural considerations. Applied in synergy, they can help ensure that educational AI fosters motivation and engagement without reinforcing false mental models.

Intersections Between Policy and Classroom Practice

The convergence of classroom pedagogy and policy frameworks offers an opportunity to address anthropomorphism systematically. For example, The OECD AI Principles emphasize transparency, accountability, and the need for users to understand when and how they are interacting with AI systems. In an educational setting, these principles support practices that make AI's non-human nature explicit and encourage critical engagement. Similarly, the EU AI Act transparency requirements can institutionalize non-anthropomorphic reminders within educational AI systems, thereby reinforcing

classroom lessons. UNESCO global guidelines could serve as a normative reference, encouraging countries to embed anthropomorphism-awareness into national curricula.

However, these opportunities depend on implementation. As long as the transparency is reduced to a one-time disclosure at first use, the corrective effect on anthropomorphism will be minimal. Effective policy must encourage repeated, context-sensitive reminders, integrated with pedagogical strategies that invite reflection rather than mere compliance (see Table 2).

Table 2. Risk–Benefit Table: Anthropomorphic Cues in Educational AI

Potential Benefit	Associated Risk	Policy / Design Mitigation
Increased student engagement through relatable interaction style	Overestimation of AI understanding or moral agency	Pair anthropomorphic elements with explicit literacy instruction and reflective debriefs
Reduced learner anxiety, especially for marginalized students	Preference for AI feedback over human correction	Ensure AI feedback is supplemented with and framed by teacher input
Facilitation of rapport in long-term tutoring contexts	Formation of emotional attachment leading to reduced critical evaluation	Rotate interaction styles; periodically emphasize system's non-human nature
Encouragement of persistence in problem-solving tasks	Misplaced trust in incorrect AI outputs	Include designed-in uncertainty indicators and prompt verification behaviors

This table 2 is not exhaustive but illustrates that anthropomorphic cues are not inherently harmful. Their effects depend on context, user background, and the presence or absence of counterbalancing literacy efforts. As such, policy should avoid simplistic bans or endorsements of anthropomorphism, focusing instead on guiding its constructive use.

Coordinated Action

Mitigating harmful anthropomorphism requires coordination between education systems, AI developers, and regulators. Educators need training and resources to address anthropomorphism as it arises in practice; developers must embed anthropomorphism-aware design principles; and policymakers should set minimum standards for transparency and user awareness. The synergy between these actors can ensure that students engage with AI as powerful tools rather than imagined peers.

Building AI literacy therefore requires both technical knowledge and reflective awareness of one's own cognitive responses to these systems. As discussed in the philosophy of mind framework, this means learning to shift between stances: the intentional stance (Dennett, 1987), which can be productive for engagement and motivation, and the design stance, which focuses on the system's actual architecture and limitations. Cultivating the ability to move deliberately between these perspectives can help students enjoy the benefits of anthropomorphic engagement without mistaking performance for genuine understanding, a lesson equally relevant in the classroom and beyond. This stance-shifting also answers Searle's worry from the Chinese Room: it reminds learners that fluent outputs need not entail semantic understanding.

CONCLUSION

Closing Remarks

The evidence presented in this paper demonstrates that anthropomorphism in AI use is neither a marginal curiosity nor an incidental side effect. It is a recurring cognitive pattern that manifests even among informed users, across diverse cultural and educational settings.

The persistence of anthropomorphic framing, even in the presence of correct declarative knowledge about AI, signals that critical AI literacy must go beyond factual explanation. It must equip learners with strategies to monitor and recalibrate their interpretations in real time, during actual interaction with AI systems. The reflective exercises described in this paper, such as structured debriefs, error-analysis tasks, and role-reversal activities, are examples of such embedded strategies. When applied consistently, they can help students maintain a functional understanding of capabilities and limits of AI tools, without eroding the relational ease that sometimes supports learning motivation.

This pedagogical imperative intersects directly with design and policy responsibilities. Developers influence anthropomorphic interpretation through seemingly small interface choices: pronoun use, tone of response, and even pacing. Policy frameworks, such as the OECD AI Principles, the EU AI Act's transparency provisions, and UNESCO's ethical guidelines on the cognitive effects of AI, already provide levers for systemic influence. Nevertheless, their impact will depend on robust implementation moving beyond tokenistic disclosures toward sustained, context-aware reminders of a non-human nature of AI, ideally in synergy with classroom practice.

Anthropomorphic cues in AI present both risks and opportunities. They can lower learner anxiety, encourage persistence, and make technology feel approachable. At the same time, they can foster misplaced trust, distort perceptions of epistemic authority of AI, and displace human-to-human feedback. The balance between these outcomes is context-dependent, and striking it requires coordination between educators, designers, and policymakers. In education specifically, the goal is not to strip AI of all human-like features, but to scaffold those features with critical literacy so they function as bridges to engagement, not traps for misconception.

The stakes reach beyond the classroom. As AI systems are embedded in public services, healthcare, and governance, anthropomorphic interpretations will influence not only individual decisions but also public trust, civic discourse, and the ethical climate in which AI operates. The interpretive habits formed in education, whether they foster uncritical trust or reflective engagement, will echo in adult interactions with AI across domains. If schools and universities can cultivate citizens who see AI as powerful tools rather than imagined peers, they will contribute to a more resilient, informed, and critically engaged society.

Philosophically, this challenge aligns with a lineage of thought stretching back to Xenophanes, who warned that humans depict gods in their own image. Today, the "gods" are algorithms, and the projections are no less revealing of ourselves. Recognizing and managing this tendency is not about diminishing the potential of AI tools but about situating them correctly acknowledging their capabilities without ascribing it human agency or moral standing. In doing so, we safeguard both the integrity of our learning environments and the clarity of our collective reasoning.

The task, then, is ongoing. Anthropomorphism will not disappear; it will adapt to each new generation of systems. Our response must be equally adaptive, embedding critical literacy into education, embedding awareness into design, and embedding transparency into policy. If we can align

these three domains, we will not only meet the practical demands of AI integration but also uphold a deeper cultural and intellectual responsibility: to understand the non-human on its own terms, and in so doing, better understand ourselves.

Limitations and future directions

This paper has combined conceptual synthesis with qualitative focus groups from a limited set of Azerbaijani schools. These cases are not intended to be statistically representative, but rather illustrative of broader cognitive and pedagogical tendencies. While comparative examples from other countries suggest cross-cultural parallels, systematic empirical studies across a wider range of cultural, linguistic, and technological contexts are needed to refine and test the framework proposed here. Likewise, the policy analysis has focused on prominent international and national initiatives; local governance structures and informal educational practices may raise distinct challenges and opportunities that merit separate examination. Future research could also explore longitudinal impacts of anthropomorphism-awareness interventions, assessing how early educational framing influences adult engagement with AI in civic and professional life.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used QuillBot Premium in order to proofread and edit the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Ackermann, H., Henke, A., Chevalère, J., Yun, H. S., Hafner, V. V., Pinkwart, N., & Lazarides, R. (2025). Physical embodiment and anthropomorphism of AI tutors and their role in student enjoyment and performance. Npj Science of Learning, 10(1). https://doi.org/10.1038/s41539-024-00293-z

American Psychological Association 7th Edition. (2019). Ethical principles of psychologists and code of conduct. American Psychological Association. https://www.apa.org/ethics/code

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social Robots for Education: A Review. Science Robotics, 3(21). https://doi.org/10.1126/scirobotics.aat5954

Breazeal, C. (2003). Toward sociable robots. Robotics and Autonomous Systems, 42(3–4), 167–175. https://doi.org/10.1016/S0921-8890(02)00373-1

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., & Heldreth, C. (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, 1–15. https://doi.org/10.1145/3613905.3650818

Creswell, J. W., & Creswell, J. D. (2022). Research design: Qualitative, quantitative and mixed methods approaches (6th ed.). SAGE.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

de Laguna, G. A. (1914). [Review of Man a Machine, by J. O. De La Mettrie]. The Philosophical Review, 23(3), 359–360. https://doi.org/10.2307/2178631

Dennett, D. C. (1987). The intentional stance. MIT Press.

Department for Education. (2025, June 27). The biggest risk is doing nothing: Insights from early adopters of artificial intelligence in schools and further education colleges. UK Government. https://www.gov.uk/government/publications/ai-in-schools-and-further-education-findings-from-early-adopters/the-biggest-risk-is-doing-nothing-insights-from-early-adopters-of-artificial-intelligence-in-schools-and-further-education-colleges

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864–886. https://doi.org/10.1037/0033-295x.114.4.864

European Union. (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689 of the European Parliament and of the Council). Official Journal of the European Union.

Festerling, J., & Siraj, I. (2022). Anthropomorphizing technology: A conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. Integrative Psychological & Behavioral Science, 56(3), 709–738. https://doi.org/10.1007/s12124-021-09668-y

Friedman, B., Hendry, D. G., & Borning, A. (2013). A survey of value sensitive design methods. Foundations and Trends in Human–Computer Interaction, 7(3–4), 55–106. https://doi.org/10.1561/1100000015

Gamez, D. (2018). Human and machine consciousness. Open Book Publishers. https://doi.org/10.11647/OBP.0107

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2024). The promise and challenges of Generative AI in Education. Behaviour & Information Technology, 44(11), 2518–2544. https://doi.org/10.1080/0144929x.2024.2394886

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Kantola, J. (2023). Elias robot's effects on students' willingness to communicate in a second language from the teacher perspective: A qualitative study [Master's thesis, University of Jyväskylä]. JYX Digital Repository. https://urn.fi/URN:NBN:fi-fe20231213153808

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). Ai tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-97652-6

Long, D., & Magerko, B. (2020). What Is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-16). Association for Computing Machinery.
https://doi.org/10.1145/3313831.3376727

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). Qualitative Data Analysis: A methods sourcebook. SAGE.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Putnam, H. (1975). Mind, language, and reality: Philosophical papers (Vol. 2). Cambridge University Press.

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with Computational Models. Neural Networks, 44, 112–131. https://doi.org/10.1016/j.neunet.2013.03.011

Reuters. (2018, March 27). Techno teachers: Finnish school tests robot educators. VOA News. https://www.voanews.com/a/finland-school-robot-educators/4319807.html

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Skinner, B. F. (1953). Science and human behavior. Macmillan.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stower, R., Calvo-Barajas, N., Castellano, G., & Kappas, A. (2021). A meta-analysis on children's trust in Social Robots. International Journal of Social Robotics, 13(8), 1979–2001. https://doi.org/10.1007/s12369-020-00736-8

Thomas, E. (2020). Descartes on the Animal Within, and the Animals Without. Canadian Journal of Philosophy, 50(8), 999–1014. https://doi.org/10.1017/can.2020.44

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

UNESCO. (2021, November). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization. Retrieved from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Vartiainen, H., Tedre, M., & Valtonen, T. (2020). Learning machine learning with very young children: Who is teaching whom? International Journal of Child-Computer Interaction, 25, 100182. https://doi.org/10.1016/j.ijcci.2020.100182

Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., Coelho, R., & Kizilcec, R. F. (2024). What explains teachers' trust in AI in education across six countries? International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-024-00433-x

Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation. W. H. Freeman & Co.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Xenophanes. (1983). The fragments of the pre-Socratics (G. S. Kirk & J. E. Raven, Trans.). Cambridge University Press. (Original work c. 530 BCE)

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. International Journal of Social Robotics, 7(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6