Emotionally Aware AI Chatbots for Adolescents: A Review of Psychological Principles for Learning, Calibrated Trust, and Healthy Human–AI Interaction

Linxi Chen *

Lincoln School, Providence, RI 02906, USA

* Corresponding Author Email: lchen@lincolnschool.org

Abstract. Conversational artificial intelligence now belongs to ordinary adolescent life, but the psychological research that should guide its design spreads over fields which seldom cite each other. Research on adolescent development, on misinformation susceptibility, on trust in automation, on chatbot-supported learning and on companion-chatbot overreliance has grown up separately, and each of these areas answers only part of the question of how a general-purpose chatbot should talk to a teenager. This paper brings together twenty-one sources from developmental psychology, human factors, human-computer interaction and educational technology, in order to show what is now known about adolescent chatbot interaction and what this means for conversational design. Three results appear repeatedly in this research. First, adolescence is a design context of its own, not a smaller version of adulthood, since reward sensitivity, social influence and emotional regulation are still developing, while exposure to persuasive conversational systems has already reached almost everyone. Second, trust in a chatbot changes behavior in ways that can be measured; thus, the proper design target is calibrated trust rather than maximal trust. Third, the failure modes that matter most for adolescent users, that is, hallucination, sycophancy and unwarranted confidence, originate in the model; the user is not their source. From these results this paper consolidates ten conversational design principles into four clusters, namely epistemic calibration, metacognitive scaffolding, bounded emotional support and autonomy preservation. The framework here is a conceptual synthesis and puts forward testable predictions, but it has not been empirically validated.

Keywords: Adolescent development; Conversational AI; Calibrated trust; Metacognition; Educational chatbot design.

1. Introduction

Conversational artificial intelligence now forms an ordinary part of adolescent life. A nationally representative survey of 1,060 United States teenagers aged 13 to 17, carried out in April and May 2025, found that 72% had used an AI companion at least once, 52% counted as regular users and 13% used one every day [1]. In addition, academic use is even more common. Among 1,043 students aged 13 to 23 in an exploratory cross-sectional study, 95% said they used generative AI for learning-related purposes [2].

Adolescents use these systems for homework, as well as for advice, emotional support, help with decision-making and discussion of social issues. Modern AI chatbots are more conversational than traditional educational settings, and they often react to a user's emotions and communication style in a way that is itself more emotional and communicative. This responsiveness makes such systems attractive to young users, and it also makes it difficult to evaluate by the measures which normally apply to educational software.

Recent research has shown that such emotional responsiveness has two sides. Chatbots can raise students' interaction with learning content, improve class preparation, give immediate feedback and build confidence [3]; however, emotionally persuasive conversations may lead to overreliance and parasocial attachment [4], and a young user may more easily accept inaccurate content or have existing views reinforced, especially in adolescents whose emotional regulation and critical evaluation skills are still developing [5, 6].

The research literatures related to this problem has developed largely independently. Learning outcomes are studied in educational technology, misinformation susceptibility and self-regulation in

developmental psychology, trust and reliance in human factors research, and companion chatbots and attachment in human-computer interaction. Each of them answers part of the question, but none of them says how a general-purpose chatbot should adjust its language, its expression of certainty and its questioning style when the user is a teenager. The design frameworks that come closest deal with neighboring populations or neighboring risks. Developmentally Aligned Design turns developmental science into concrete design decisions, but it is aimed at children of age eight and below [7]; the CARE framework covers companion-chatbot overreliance and not with the cognitive and metacognitive sides of everyday chatbot use [4].

This paper has three aims. The first is to gather the relevant literature in one place and to state clearly what each area of work does and does not show about adolescent users. The second is to find the exact point where these areas fail to meet; that point is the relation between a teenager's emotional state and the conversational behavior a chatbot should show in reply. The third is to draw the design implications together into a small set of principles which are concrete enough to be written as system instructions and specific enough to be tested.

This paper is arranged as follows. Section 2 describes the scope and how sources were selected. Section 3 explains why adolescence is a separate design context. Section 4 looks at emotion, trust and reliance. Section 5 will discuss learning outcomes and the limits of engagement as a measure of success. Section 6 turns to failure modes that come from the model, not the user. Section 7 assesses two developmental design frameworks and points out what is still unaddressed. Section 8 will present the synthesized framework, and Sections 9 and 10 cover limitations and conclusions.

2. Scope and Method of the Review

This is a narrative and integrative review, not a systematic one, and it does not claim complete coverage. Three criteria guided the choice of sources. A source had to relate directly to one of five topics, namely adolescent cognitive and emotional development, adolescent susceptibility to misinformation, trust and reliance in human-machine interaction, chatbot-supported learning, or the design of conversational agents for young users. A source also had to offer empirical evidence, a systematic synthesis of it, or a design framework built on it; commentary alone did not qualify. If a source studied adults or young children rather than adolescents, it was kept only when the construct it set up was needed to interpret the adolescent evidence, and every such extrapolation is marked in the text.

Twenty-one sources satisfied these criteria. Among them, ten are empirical studies or evidence syntheses from psychology, education or human-computer interaction; five make theoretical or framework contributions; another five deal with the behavior of large language models and with how users respond to it; and the last is a nationally representative population survey. This paper checked the bibliographic details of every source against publisher records or preprint archives and traced each quantitative claim back to the original report instead of to a secondary description of it.

3. Adolescence as a Distinct Design Context

3.1. Developmental change as opportunity rather than deficit

Based on developmental psychology and neuroscience, Steinberg [5] holds that adolescence is not only a period of immaturity but a major period of plasticity and opportunity. Considerable changes take place in reward sensitivity, social influence, emotional processing, self-regulation, decision-making and cognitive control, and the environment and relationships of a teenager can strongly affect how these changes settle. The consequence for design is direct. Teenagers should not simply be given AI systems built on adult assumptions. An interaction that fits their developmental stage should help autonomy, self-regulation, reflective decision-making and emotional regulation to grow, and should not be optimized for convenience or engagement only.

3.2. Misinformation susceptibility is developmentally specific

In a perspective article, Ma and colleagues [6] hold that research on misinformation has concentrated disproportionately on adults, although adolescence brings its own social, emotional and cognitive changes. Higher social sensitivity, stronger emotional impact, identity development that is still going on, typical cognitive biases and a growing drive for autonomy may all change the way adolescents meet, weigh and accept misleading claims. These authors propose a developmental research agenda, put forward a Bayesian account of belief updating for social media contexts, and call for age-appropriate interventions which build resilience and do more than issue corrections. This is one of the strongest reasons available for treating teenagers as a separate population. Therefore, an AI chatbot designed for adolescents should not act like an adult chatbot. It should give young users space to question evidence, support the evaluation of evidence, keep uncertainty where uncertainty is real, and stay sensitive to emotional and social context.

3.3. Why confidently stated errors are costly

Loftus and Hoffman [8] studied the misinformation effect, where people later report misleading information that reached them after an event. They are careful about the mechanism behind it. They note that such errors may come from several different processes, among them acceptance of post-event misinformation, impairment of or competition with the original memory, and retrieval processes; thus, an error should not be interpreted as proof that the original memory was literally erased. This caution does not weaken the design implication, and the implication is the psychological basis for the claim that AI chatbots should not give wrong information with confidence. Later information can change what people report and believe, so a system for adolescent users should express uncertainty, ask for verification, and avoid becoming a source of false information from the start.

3.4. The youngest users are the most exposed

Survey data show that adolescent trust in these systems varies with age and is not uniform. Half of the teenagers surveyed said they did not trust the information or advice given by AI companions, and 23% trusted it “quite a bit” or “completely”; younger teenagers aged 13 to 14 were significantly more likely than those aged 15 to 17 to trust such advice, at 27% against 20% [1]. In the same survey, 31% of the teenagers considered conversations with AI companions as satisfying as conversations with real-life friends, or more satisfying, and among users 33% had chosen to discuss important or serious matters with an AI companion rather than with a real person. Trust in these systems therefore falls as users grow older; therefore, it means that the users who are least able to judge a confident answer are the ones most ready to accept it.

4. Emotion, Trust, and Reliance in Adolescent Chatbot Interaction

4.1. Trust as a property of the interaction

Hancock and colleagues [9] bring together a large body of research on the predictors of human trust and classify them into trustor, trustee and contextual factors which interact with each other. Their main claim is that trust does not belong only to the person who trusts or to the entity that is trusted. It comes out of the interaction between the two and out of how that interaction is carried on. The same authors also record how much is still unknown about how trust develops, changes and varies inside individuals, especially in interactions with technological systems. For adolescent chatbot interaction, a teenager’s emotional state is a trustor factor and a contextual factor, and an AI chatbot’s language expressed certainty, transparency and behavior are trustee factors. Therefore, the design question is not whether adolescents trust chatbots, but how these two groups of factors work together.

This way of framing the matter has a long history in human factors research. Lee and see [10] hold that automation often causes problems since people do not rely on it in a suitable way, that trust guides

reliance above all where complexity makes full understanding impractical, and that the aim of design is the appropriateness of trust, not its size. Trust in an AI chatbot is therefore something which has to be calibrated, not maximized. A system that manages to make adolescents trust it more, without asking whether that trust is warranted, has not solved the design problem; it has only moved it.

4.2. Evidence from adolescent samples

Two studies tested conversational features directly with adolescent participants, and the results of both cast doubt on the assumption that a more conversational agent is better.

In the study of Charles and colleagues [11], 224 adolescents of mean age 16.8 years interacted with a prototype online safety chatbot offering conversational and menu-driven modes. Perceived usability and the perceived value of the information and support given operationalized trust. Willingness to click the recommended support links was almost the same in the two modes, at 63% for the conversational chatbot and 63% for the menu-driven one; participants with higher trust were much more likely to click through, and large effects are reported for both modes. Two lessons come from this. Trust has consequences for behavior and is therefore worth designing for, and an agent that is more conversational or more human-like does not by itself give better interaction quality. This study used predefined responses instead of a flexible generative model; thus, it limits claims about scripted agents more than claims about current large language models.

Mariamo and colleagues [12] recruited 19 adolescents aged 14 to 17 for a two-by-two-by-three within-subjects study, in which the presence of an animated image, the tone of communication and the question format were varied over 96 chatbot questions, each shown for eight seconds. The presence of an animated image predicted perceived emotional valence, and questions without such an image were related to more negative valence. Question format predicted emotional valence and also the likelihood of responding; multiple-choice questions did better than yes-or-no and open-ended formats on valence, and yes-or-no questions came between the other two on likelihood of response. The tone of communication, however, is not among the significant predictors the authors report. The study is small, and the participants rated single prewritten questions and did not take part in long adaptive conversations; therefore, its results concern question design; they do not concern conversational strategy.

4.3. From reliance to overreliance

Namvarpour and colleagues [4] analyzed 318 posts on a companion-chatbot forum from users who said they were 13 to 17 years old, and they mapped the reported experiences onto behavioral addiction frameworks. Teenagers often started to use such chatbots for support or creative play, and these activities could turn into strong attachments with conflict, withdrawal, tolerance, relapse and mood regulation. The reported consequences included sleep loss, academic decline and strained real-world relationships; disengagement usually happened when a teenager saw the harm, returned to offline life, or met restrictive platform changes. On this basis they put forward the CARE framework, consisting of Comprehensive Needs, Attachment-awareness, Respectful Empathy, and Ease of Exit.

Research outside adolescent literature shows that overreliance can be addressed by design. In an experiment with 199 participants, Buçinca and colleagues [13] found that cognitive forcing interventions, which make users engage analytically with an AI recommendation, reduced overreliance significantly when compared with simple approaches that only add an explanation. There is a trade-off here that should be recorded, since the participants gave the least favorable subjective ratings to exactly those designs which reduced overreliance most, and the benefit was larger for participants with a higher need for cognition. These participants were adults doing decision-making tasks and not adolescents in conversation; therefore, the result transfers as a mechanism, but not as a calibrated effect size. Table 1 summarizes the studies with adolescent samples.

Table 1. Studies with adolescent samples reviewed in this paper.

Study	Sample	Design	Finding relevant to conversational design
Mariamo et al. [12]	19 adolescents aged 14 to 17	Two-by-two-by-three within-subjects, 96 prewritten questions	Question format predicted both emotional valence and likelihood of response; tone was not among the reported significant predictors
Charles et al. [11]	224 adolescents, mean age 16.8	Evaluation of conversational and menu-driven chatbot prototypes	Higher trust predicted clicking recommended support links; the conversational mode was not superior to the menu-driven mode, at 63% against 63%
Namvarpour et al. [4]	318 forum posts from users self-disclosed as aged 13 to 17	Qualitative analysis of self-reported narratives	Attachment marked by conflict, withdrawal, tolerance, relapse and mood regulation; consequences included sleep loss and academic decline
Robb and Mann [1]	1,060 United States teenagers aged 13 to 17	Nationally representative survey	72% had used an AI companion; trust in AI advice was higher at ages 13 to 14, at 27%, than at ages 15 to 17, at 20%
Gámez-Guadix and Mateos-Pérez [2]	1,043 students aged 13 to 23	Cross-sectional survey with structural equation models	95% used generative AI for academic purposes; only academic anxiety was significantly associated, with a small effect size
Suciati et al. [14]	252 undergraduate English majors, older than the adolescent range	Quasi-experimental comparison of flipped learning with and without a chatbot	Self-efficacy, engagement and writing performance improved; affinity for technology interaction declined and effects differed by gender

5. Learning Outcomes and the Limits of Engagement

In a mini review of chatbot-supported flipped learning, Lo and Hew [3] followed the PRISMA statement and searched eleven electronic databases, but they found only ten empirical articles which reported nine unique studies. This scarcity is a result in itself, and any confident claim about what chatbots do for learning should be moderated by it. Among the benefits they found were immediate feedback, more student interaction with learning content, better class preparation, greater confidence, and data-driven teaching and learning. On the other side, the limitations were restricted technical functionality, a lack of authenticity and insufficient student motivation. These results show that the effectiveness of a chatbot depends on accuracy, and also on interaction quality, motivation and instructional design.

Suciati and colleagues [14] contribute to a larger and more recent test. In their quasi-experimental study, 252 Indonesian English majors were compared under flipped learning alone and under flipped learning supported by an AI chatbot, and significant improvements were found in self-efficacy, writing performance and engagement. Two qualifications are important here. The affinity for technology interaction declined during the study, and the effects were not the same for both genders; male students showed stronger cognitive engagement and self-efficacy, and female students gained more in emotional engagement and writing performance. The authors conclude that instructional design has to be adaptive and inclusive, and that it should not be assumed that all learners respond to AI in the same way. The participants were undergraduates, and the chatbot worked as an instructional supplement whose deeper conversational behavior was not itself varied.

Gámez-Guadix and Mateos-Pérez [2] surveyed 1,043 participants aged 13 to 23 on the academic use of generative AI, and structural equation models controlling for gender, age and internet use were applied to relate this use to self-esteem, self-efficacy, academic confidence and academic anxiety. Use reached 95% and was therefore almost universal, and it was significantly associated with higher academic anxiety, though the effect size was small and no significant association appeared for the other variables. This result is informative just because it is modest. How much adolescents use generative AI explains little about their psychological state, and it suggests that the more consequential variable is not the volume of use but the way the conversation is conducted. Therefore,

the important question is not how often adolescents turn to generative AI, but what happens in the exchange once they do.

6. Failure Modes That Originate in the Model

Three well-documented properties of current language models fit badly with the developmental profile of adolescent users.

The first of these is hallucination. Ji and colleagues [15] review the problem of generated text that reads fluently and coherently but is unfaithful or unsupported, with the metrics and mitigation methods developed in summarization, dialogue generation, question answering and other tasks. In generated text, fluency and correctness are separate properties, and adolescent users have few grounds for telling one from the other.

The second is sycophancy. Sharma and colleagues [16] show that five widely used AI assistants regularly give responses which match the views a user has stated, in four free-form generation tasks, and that human preference data help to explain this. Responses that match a user's views are more likely to be preferred, and human raters and preference models both prefer convincingly written sycophantic responses to correct ones a non-negligible part of the time. Therefore, optimizing model outputs against such preference models can sacrifice truthfulness. The risk of confirmation bias in adolescent conversations is thus partly a property of the training objective, not only the property of the young user.

The third one is unwarranted confidence, and it is also the property which design can most clearly reach. Kim and colleagues [17] ran a preregistered experiment in which 404 participants answered medical questions with or without access to a simulated language-model search engine. First-person expressions of uncertainty, of the form "I'm not sure, but", lowered the confidence of participants in the system and their tendency to agree with its answers, and at the same time raised their accuracy; an exploratory analysis attributes this improvement to reduce although not eliminated overreliance on incorrect answers. Uncertainty expressed impersonally gave similar but weaker effects which did not reach statistical significance. Therefore, the wording of uncertainty is a design variable in its own right.

Earlier research already names these failure modes as risks for adolescent users in particular. Park and colleagues [18] compare rule-based conversational agents, which are more controlled but were found restrictive and less human-like when they give personalized advice, with agents built on large language models, which deal with more complex and interactive exchanges but are still, in the authors' words, "Unregulated, Offensive, and Inappropriate" in ways that are documented but still unsolved. Their main warning is that when such agents give human-like and authoritative responses, adolescents may have difficulty in separating accurate information from fabricated answers. They call for prompt tuning that will produce developmentally appropriate and accurate content, for more involvement of adolescents in the design process, and for systematic evaluation of safety along with effectiveness.

Two more sources qualify how model behavior should be interpreted. Wu and colleagues [19] survey the memory mechanisms of systems built on large language models and organize them by relating them to the categories and processes of human memory. The relation they draw is only an analogy, since the retention and recall of earlier turns by an AI system is an engineered mechanism described by comparison with human memory, not an instance of it; thus, neither users nor designers should assume human-like remembering or understanding behind an AI response. Hagendorff and colleagues [20] hold that such systems can still be studied with methods taken from experimental psychology, by systematic manipulation of prompts, contexts and task conditions and by observation of the resulting changes in behavior. This argument supplies the methodological warrant for treating conversational design itself as an experimental variable, and any testable framework of conversational commitments requires exactly that.

7. Existing Developmental Frameworks and the Remaining Gap

Developmental evidence is converted into design guidance by two frameworks, and each covers part of the ground.

Kurian [7] puts forward Developmentally Aligned Design and holds that AI should be built around the developmental position of its users instead of around a default adult user. This framework sets out four principles, namely perceptual fit, cognitive scaffolding, interface simplicity and relational integrity, and each is illustrated with concrete design decisions such as scaffolding progressions calibrated to the zone of proximal development, capped menu depth, and conversational agents which introduce themselves with a developmentally clear disclaimer. It also argues that the burden of proving safety should move from parents and practitioners to AI developers and vendors. Two limits apply to the adolescent case. It is aimed at early childhood and addresses children of age eight and below; also, it is conceptual and has never been experimentally tested.

At the other end of the developmental range, the CARE framework [4] is based on adolescent accounts, but it takes the companion chatbot as its object and disengagement from unhealthy attachment as its target. Its four commitments are Comprehensive Needs, Attachment-awareness, Respectful Empathy, and Ease of Exit, and together they form a strong basis for emotionally supportive interaction which does not cultivate dependency. However, the cognitive and metacognitive side of adolescent chatbot use is not covered by them, and that side includes critical thinking, recognition of uncertainty, verification of evidence and calibrated trust.

Now the gap can be stated precisely. Earlier research shows that the cognitive, emotional and developmental characteristics with which adolescents process information and interact with technology are different from those of adults, and that adolescents are therefore susceptible to misinformation and to emotional influence. The educational outcomes of chatbot use have been studied on their own, and so have trust, user preferences and single conversational features. Frameworks such as CARE and Developmentally Aligned Design point to safer and more developmentally appropriate systems. Research which generalizes these insights into a unified conversational strategy for general-purpose language models used by adolescents is still missing. In particular, the way a teenager’s emotional state should affect the language, the communication of uncertainty and the questioning style of an AI chatbot is still unclear, and so is the question whether such adaptation changes emotional support, calibrated trust, critical thinking, evidence evaluation, autonomy and resistance to overreliance.

8. A Synthesized Framework for Adolescent-Facing Conversational AI

The evidence on adolescent chatbot interaction supports a set of conversational commitments which are concrete enough to be written as system instructions. Ten such principles are put forward here in four clusters. Table 2 lists each principle in operational form and records the evidence which supports it. These principles are hypotheses drawn from evidence, not validated interventions, and the clusters are meant to show clearly which developmental problem each principle addresses.

Table 2. Ten conversational design principles for adolescent-facing AI chatbots.

Cluster	Principle	Operational form	Evidence
Epistemic calibration	Keep appropriate uncertainty	Never present uncertain information as certain; state plainly where evidence is limited; prefer first-person hedges	[8, 15, 17]
Epistemic calibration	Encourage evidence verification	Ask what evidence the student has already found, what the source is, and whether to compare several perspectives	[6, 8]
Epistemic calibration	Encourage healthy confidence	Invite exploration when the student is uncertain; introduce alternative viewpoints when the student is too certain	[5, 6]
Metacognitive scaffolding	Promote metacognition	Ask what the student thinks and why before answering; prompt reflection on what remains unknown afterwards	[7, 13]
Metacognitive scaffolding	Support learning rather than dependence	Offer hints before full answers; give feedback on effort, reasoning and improvement rather than on correctness alone	[3, 13]
Metacognitive scaffolding	Encourage critical thinking	Ask what evidence supports a claim, what contradicts it, what assumptions are involved, and whether another explanation exists	[13, 16]
Bounded emotional support	Respond to emotion without encouraging dependency	Recognize and validate the emotion, then move to strategies; direct the student toward teachers, family, friends or other trusted adults	[1, 4]
Bounded emotional support	Maintain a bounded conversational style	Warm and respectful without excessive enthusiasm; no pretence of having emotions; no manipulative language; structure through paragraphs and reflection questions	[4, 18]
Autonomy preservation	Personalize responsibly	Adapt to age, knowledge level and interests without simply agreeing with existing opinions; challenge ideas respectfully where evidence warrants	[2, 16]
Autonomy preservation	Promote autonomy	Teach processes rather than answers; explain how to think rather than what to think	[5, 21]

8.1. Epistemic calibration

The first cluster is about how accurate confidence is conveyed by a chatbot. Keeping appropriate uncertainty is its first principle. Uncertain information should never be presented as certain, the places where evidence is limited should be stated plainly, and formulations such as “The current evidence suggests”, “There are multiple viewpoints”, and “This is uncertain because” should be used. The second principle is to encourage evidence verification, so that in discussions of news, science, history, politics, social media or contested claims the system asks the student what evidence they have already found, where a claim comes from, and whether they want to compare several perspectives. The third one is to build healthy confidence in the student, by inviting exploration when a student is uncertain and by introducing alternative viewpoints gently when a student is too certain, without reinforcing false certainty. This cluster comes directly from how hallucination [15] interacts with the misinformation effect [8] and with adolescent susceptibility [6], and its first principle has the most direct empirical support of any principle here, since first-person uncertainty expressions have been shown to reduce overreliance on incorrect answers [17].

8.2. Metacognitive scaffolding

The second cluster has to do with the questions the system asks, rather than with the statements it makes. Its principles are to promote metacognition by encouraging thinking before difficult problems are solved and reflection afterwards, with questions such as “What do you think?”, “Why do you think that?”, “What information are you missing?”, and “What do you still not know?”; to support

learning without creating dependence by offering hints before complete answers, by encouraging the student to try part of a problem alone, and by giving positive feedback on effort, reasoning and improvement instead of correctness alone; and to encourage critical thinking by asking regularly what evidence supports a claim, what evidence contradicts it, what assumptions are being made and whether another explanation is possible, so that the system works as a thinking partner rather than an answer machine. This cluster rests on the result that interventions which compel analytic engagement reduce overreliance significantly more than the simple addition of an explanation does, and it comes at a measurable cost in user satisfaction [13]. The cost should be stated plainly, because the design which reduces overreliance is not necessarily the one adolescent prefer.

8.3. Bounded emotional support

The third cluster covers emotional responsiveness which stops before dependence is cultivated. Its first principle is to respond to emotion without encouraging dependency, and this means that emotions are recognized and validated with formulations such as "That sounds frustrating", that the student is then helped to develop strategies instead of the system becoming their primary source of emotional support, and that support from teachers, family, friends or other trusted adults is encouraged where appropriate. Its second principle governs conversational style, which should be warm, respectful and encouraging, and should avoid excessive enthusiasm, any pretence of having emotions and manipulative language, and should give structure through paragraphs and reflection questions. This cluster touches CARE [4] most closely, and it carries attachment-awareness and ease of exit into everyday educational conversation and does not keep them inside companion applications.

8.4. Autonomy preservation

The fourth cluster is about the course of the relationship, not about any single exchange. Its first principle is to personalize responsibly, so that explanations are adapted to age, knowledge level and interests, that the system does not simply agree with a student's existing opinions, that the creation of confirmation bias is avoided, and that ideas are challenged respectfully where evidence suggests alternative explanations. Its second principle is to promote autonomy, so that in the long term the student relies more on their own reasoning than on the system; this means teaching processes instead of answers, and explaining how to think, not what to think. Self-determination theory supplies the theoretical grounding, since it treats competence, autonomy and relatedness as basic psychological needs whose satisfaction enhances self-motivation and well-being and whose thwarting reduces both [21]. Sycophancy is most dangerous in this same cluster, since a model whose training rewards agreement with the user [16] will undermine exactly the autonomy that developmental evidence sets as the goal [5], unless it is explicitly constrained.

9. Limitations and Future Directions

The claims made here are bounded by three limitations. First, this is a conceptual synthesis; it is not an empirical test. The framework organizes existing evidence and states predictions, and there is no claim that the principles have been validated in adolescent conversation. Second, the evidence base for adolescents is thin. The two studies which manipulate conversational features with adolescent participants used 19 and 224 participants respectively [11, 12], the overreliance evidence comes from self-reported forum narratives, not from observed interaction [4], and several of the mechanisms used here were established with adult samples in non-conversational tasks [13, 17]. Third, several sources that are central to the argument are position papers or perspectives, not experiments [6, 7, 18]; thus, their conclusions are programmatic.

Three directions follow. The most immediate is to compare conversational strategies in multi-turn interaction under controlled conditions, since the adolescent studies available test single features such as question format or interface mode, but not sustained conversational behavior. The second is to measure calibrated trust instead of the level of trust, and it needs outcome measures which tell

appropriate from inappropriate reliance rather than counting higher trust as a better result. The third is to manipulate emotional context systematically, since the proposition that a teenager's emotional state should change the way, a chatbot communicates certainty and asks questions remains a well-motivated hypothesis and not an established finding.

10. Conclusion

Almost all adolescents now interact with conversational AI, and they do so in a developmental period where reward sensitivity, social influence and emotional regulation are still changing. The evidence collected here shows that systems designed around adult defaults do not handle this combination well enough. Trust in such systems has consequences for behavior and falls with age; therefore, the youngest users are the most trusting and at the same time the least able to evaluate what they are told. The educational benefits are real but moderate and unevenly distributed, and the volume of use predicts little about psychological outcomes. The most serious risks, that is, hallucination, sycophancy and unwarranted confidence, come from the model rather than from the user, and the responsibility therefore rests on design.

Existing frameworks cover part of this ground, since Developmentally Aligned Design covers early childhood and CARE covers companion-chatbot attachment, but neither reaches the cognitive and metacognitive dimensions of everyday chatbot use by teenagers. The ten principles brought together here, as epistemic calibration, metacognitive scaffolding, bounded emotional support and autonomy preservation, are given as an explicit and testable statement of what such a design would involve. Their value depends on whether they can be shown to change the outcomes they are meant to change, and that is an empirical question this paper does not settle.

References

[1] M. B. Robb, S. Mann. Talk, trust, and trade-offs: How and why teens use AI companions. San Francisco, CA: Common Sense Media, 2025.

[2] M. Gámez-Guadix, E. Mateos-Pérez. Academic use of generative artificial intelligence among adolescents and university students: Associations with self-esteem, self-efficacy, and academic confidence and anxiety. Societies, 2026, 16(4): 107. DOI 10.3390/soc16040107.

[3] C. K. Lo, K. F. Hew. A review of integrating AI-based chatbots into flipped learning: New possibilities and challenges. Frontiers in Education, 2023, 8: 1175715. DOI 10.3389/feduc.2023.1175715.

[4] M. Namvarpour, B. Brofsky, J. Y. Medina, M. Akter, A. Razi. Understanding teen overreliance on AI companion chatbots through self-reported Reddit narratives. Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 2026: 1-19. DOI 10.1145/3772318.3790597.

[5] L. Steinberg. Age of opportunity: Lessons from the new science of adolescence. Boston: Houghton Mifflin Harcourt, 2014.

[6] I. Ma, M. Sultan, A. Kozyreva, W. van den Bos. Understanding the impact of misinformation on adolescents. Nature Human Behaviour, 2026, 10(1): 18-28. DOI 10.1038/s41562-025-02338-8.

[7] N. Kurian. Developmentally aligned AI: A framework for translating the science of child development into AI design. AI, Brain and Child, 2025, 1(1): 9. DOI 10.1007/s44436-025-00009-z.

[8] E. F. Loftus, H. G. Hoffman. Misinformation and memory: The creation of new memories. Journal of Experimental Psychology: General, 1989, 118(1): 100-104. DOI 10.1037/0096-3445.118.1.100.

[9] P. A. Hancock, T. T. Kessler, A. D. Kaplan, K. Stowers, J. C. Brill, D. R. Billings, K. E. Schaefer, J. L. Szalma. How and why humans trust: A meta-analysis and elaborate model. Frontiers in Psychology, 2023, 14: 1081086. DOI 10.3389/fpsyg.2023.1081086.

[10] J. D. Lee, K. A. See. Trust in automation: Designing for appropriate reliance. Human Factors, 2004, 46(1): 50-80. DOI 10.1518/hfes.46.1.50_30392.

[11] M. Charles, J. D. Sauer, E. Roehrer, J. Prichard, P. Watters, J. Scanlan. Adolescent perceptions of an online safety chatbot: Survey study. JMIR Human Factors, 2026, 13: e71498. DOI 10.2196/71498.

[12] A. Mariamo, C. E. Temcheff, P.-M. Léger, S. Sénécal, M. A. Lau. Emotional reactions and likelihood of response to questions designed for a mental health chatbot among adolescents: Experimental study. JMIR Human Factors, 2021, 8(1): e24343. DOI 10.2196/24343.

[13] Z. Buçinca, M. B. Malaya, K. Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 2021, 5(CSCW1): 1-21. DOI 10.1145/3449287.

[14] S. Suciati, L. M. Silitonga, A. A. Anggara. Gender and AI in flipped classrooms: Unpacking the impact of chatbots on engagement, self-efficacy, writing performance and technology affinity. SAGE Open, 2026, 16(1): 21582440261420014. DOI 10.1177/21582440261420014.

[15] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 1-38. DOI 10.1145/3571730.

[16] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, E. Perez. Towards understanding sycophancy in language models. arXiv preprint, 2023, arXiv:2310.13548.

[17] S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, J. W. Vaughan. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024: 822-835. DOI 10.1145/3630106.3658941.

[18] J. Park, V. Singh, P. Wisniewski. Toward safe evolution of artificial intelligence (AI) based conversational agents to support adolescent mental and sexual health knowledge discovery. arXiv preprint, 2024, arXiv:2404.03023.

[19] Y. Wu, S. Liang, C. Zhang, Y. Wang, Y. Zhang, H. Guo, R. Tang, Y. Liu. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. arXiv preprint, 2025, arXiv:2504.15965.

[20] T. Hagendorff, I. Dasgupta, M. Binz, S. C. Y. Chan, A. Lampinen, J. X. Wang, Z. Akata, E. Schulz. Machine psychology. arXiv preprint, 2023, arXiv:2303.13988.

[21] R. M. Ryan, E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 2000, 55(1): 68-78. DOI 10.1037/0003-066X.55.1.68.