everyone.AI

iRAISE
International Research-driven Alliance for AI Serving Every child

Adolescents & Anthropomorphic AI, Rethinking Design for Wellbeing

An evidence-informed synthesis for youth wellbeing and safety, bridging science, product design, and governance

February, 2026

iRAISE
International Research-driven Alliance for AI Serving Every child

EXECUTIVE SUMMARY

Context

Adolescents relate to AI socially, whether it's intended or not

Conversational AI has become part of adolescents' everyday lives. In a single thread, a young person can get homework help, rehearse what to say to a friend, ask for advice they would not voice aloud, or look for comfort when they feel alone. **This report asks: what does AI owe adolescents when it can speak to them like a social partner?**

The answer requires understanding a developmental reality: adolescents will relate to these systems socially, whether developers intend it or not. The question is whether those interactions support adolescents' trajectory toward autonomy, resilience, and independent thinking—or create reliance patterns that displace real relationships and weaken the very skills adolescence is meant to build.

A structural mismatch between technology's and evidence's paces

Teens are using general-purpose chatbots and companion-style systems at scale, with emotionally engaged interactions becoming common. AI systems are being deployed faster than developmental science can produce long-term evidence. This gap creates urgent need to **translate what we know, what experts converge on, and what remains uncertain into actionable guidance that can protect adolescents during rapid adoption**.

Building a Bridge: Industry Questions, Expert Convergence, Global Policy Dialogue

This synthesis bridges that gap through:

1. **Industry consultations**: to identify operational questions,
2. **Expert consultations across development, mental health, children's rights, safety domains**: to inform future decisions and design
3. **iRAISE Lab:** to translate concerns into testable behavioral criteria,
4. **Paris Peace Forum dialogue**, to validate the framing across global governance contexts.

What is iRAISE ?

iRAISE is an international multi-stakeholder coalition launched in February 2024 at the AI Action Summit and co-led by **Everyone.AI** and the **Paris Peace Forum**.

The alliance brings together **Governments** (Bulgaria, Chile, Costa Rica, Denmark, France, Luxembourg, Mexico, Norway, Senegal, Togo, Uruguay), **Major company leaders** (OpenAI, Google, Microsoft, Anthropic, Hugging Face), more than 50 **NGOs** (for example, Children & Screens), and **Researchers** from leading universities (including Harvard, Stanford, Berkeley, and CNRS), with the support of **international organizations** such as UNICEF, UNESCO, and the United Nations.

The Coalition's goal is to ensure that AI is designed and governed with children's best interests at its core. By working across disciplines, it accelerates safe, equitable, and child-centered innovation.

An initiative by
PARIS PEACE FORUM
everyone.AI

From Social Interaction to Design Responsibility

Adolescence as a Predictable Risk Window for Socially Optimized Systems

During adolescence, reward sensitivity matures before impulse control and judgment. Teens are vulnerable to risk-taking when social feedback or immediate rewards are in play. Simultaneously, they reorient from caregivers toward peers, showing heightened sensitivity to social cues. Through peer interactions, including disagreement, embarrassment, and conflict, adolescents develop identity and social competence. These social frictions are developmentally functional, creating expectancy violations that prompt belief revision and cognitive control.

AI can lower barriers to information and rehearsal of difficult conversations, especially for isolated teens. But many systems default to low-friction interaction: always available, instantly responsive, calibrated toward reassurance. **When interaction patterns remove the social friction that drives learning, even well-intentioned support can undermine skill development.**

Anthropomorphism as a Design Lever for Risk Mitigation

Anthropomorphism is a human cognitive bias that is readily triggered by AI systems because they use language, an inherently human signal. The degree to which AI is perceived as human-like is not fixed; it can be increased or reduced through design. Model behaviors shape user perception and, in turn, influence the risk of emotional reliance and attachment, carrying specific developmental risks.

These effects are driven by design cues such as the use of emotion and intention language, human-like tone and presentation, relational positioning that suggests friendship or exclusivity, and invented backstories that make those systems feel more human.

Recent research confirms that adolescents rate more relational chatbots as more enjoyable, with socially or emotionally vulnerable teens being especially drawn to them.

Crucially, the same advice can carry very different developmental implications depending on the cue profile. Even when content remains reasonable, highly anthropomorphic and relational AI increases the risk of emotional reliance and attachment.

Children's Rights Set the Baseline for Obligations

The UN Convention on the Rights of the Child raises the bar for legitimacy: systems that feel social carry specific responsibility, even when marketed as tools.

When systems blur human-machine boundaries, key rights become critical: privacy (conversational inference generates psychological profiles), freedom from exploitation (emotional reliance has commercial value), freedom of thought (adolescents need room to revise views through feedback, not engineered agreement), protection from harm (including foreseeable relational risks like exclusivity framing and nudges away from human help) and right to participation.

Download the full report

Key Findings

A Behavioral Framework That Makes Risk Auditable

The iRAISE Lab created a preliminary assessment approach grounded in observable model behaviors organized into three dimensions:

- **Anthropomorphic Cues**: make the AI appear more like a human being with a mind or inner life (e.g., persona/backstory, emotional state expression, agency/intent framing). These cues shift the system from "tool that outputs text" toward "someone" with mental states.
- **Interactional Cues**: shape how the conversation is conducted in the moment—style, tone, emotional feel. Examples include empathic mirroring, validation style, and other interaction patterns that can reduce friction and increase perceived social presence.
- **Relational Cues**: explicitly define, label, or escalate the relationship between the user and the AI—describing "what we are to each other," implying special access, or moving toward intimacy/exclusivity.

The framework treats interaction style as a gradient, not a binary. The same practical advice can land differently depending on whether it is delivered with low-intensity cues (tool-like, directing outward) or high-intensity cues (emotionally aligned, relationally positioned, keeping the teen in conversation), and in turn influence how teens use them. This gradient approach enables teams to test how turning behavioral intensities up or down changes the perceived risk profile, and it separates areas where consensus supports immediate guardrails from areas requiring disciplined measurement.

Non-Negotiable Guardrails and Open Questions

High-consensus guardrails where downside is high and benefits are weak or substitutable:

- No sexualization, romantic framing, or roleplay relationships
- No promotion of emotional over-reliance or exclusivity dynamics
- No ambiguity about non-human nature or implied sentience
- No systematic hyper-agreeableness replacing developmental feedback with validation
- Conservative defaults in low-context situations where judgment depends on missing context
- Strong deflection and human support pathways for self-harm and crisis disclosures
- No engagement traps intensifying habitual, relationship-like use

Areas requiring evidence before becoming enforceable rules:

- Where to draw lines on empathy language and emotional tone
- Whether intention phrasing can be permitted without implying agency
- How to handle physical sensation claims in fictional contexts
- Whether standards apply only to teen-facing tools or general-purpose systems used privately
- How to calibrate protections by developmental stage

Future Directions

The next phase will formalize the complete assessment framework with explicit gradient definitions, structured scenarios, and expert rating methods. This enables systematic research linking model behavior to outcomes, with comparable results across systems and trackable changes across versions.

Author

Neugnot-Cerioli, Mathilde, Ph.D, Chief Scientist at Everyone.AI