CHAILD White Paper II

Children's Autonomy in the Age of AI: Closing the Evidence Gap

Reflections for the 2026 AI Safety agenda and the Indian AI Summit

By Jun Zhao, Carina Prunkl, Manolis Mavrikis, Kaska Porayska-Pomsta, Wayne Holmes, Nigel Shadbolt and the CHAILD team

The 2026 International AI Safety Report [1] provides a comprehensive assessment of advanced AI risks identified to date. These risks span malicious use, system malfunctions, and systemic effects across labour markets and **human autonomy**. While the 2026 report identifies human autonomy as a central concern, it leaves significant evidence gaps regarding how this affects children, a population uniquely shaped by their developmental vulnerability and transition.

As policymakers gather for the Indian AI Summit 2026, this white paper spotlights the need for a greater attention on **children's autonomy and agency in the context of AI**, by presenting emerging evidence and outlining a future policy roadmap.

---

1. Autonomy Risks Identified in the 2026 AI Safety Report

The 2026 report highlights three major autonomy-related risks.

- **Risks to cognitive offloading and erosion of critical thinking:** with AI systems being increasingly used to perform tasks such as "writing, tutoring, problem-solving, and information seeking" [1]. Delegating such tasks may free cognitive resources but it may also weaken long-term critical thinking skills through excessive "cognitive offloading", particularly if tools are adopted without careful consideration of pedagogical soundness and long-term learning effects [2] . For adults, this presents measurable performance concerns [3,4]. However, while evidence for the impact of AI-linked cognitive offloading on adults is beginning to emerge [5,6], similar evidence is largely unavailable for children.
- **Introduction of automation bias:** Over-reliance on AI outputs can lead individuals to prioritise automated judgments over their own reasoning. Evidence is particularly emerging in domains like medical diagnostics and high-stakes decision-making [7,8]. Beyond technical error, automation bias raises a deeper question: What happens to the

authenticity of judgment when systems routinely substitute for human deliberation? For young users, habitual delegation to AI systems could shape epistemic habits before independent reasoning is fully formed [9] and cause lasting effects.

- **Impact on emotional dependence and self-regulation:** Emerging evidence shows increasing emotional attachment to AI chatbots and companions [P92]. High-frequency use may affect self-regulation and, in vulnerable users, mental health. Again, in contrast to adults, who have had a chance to develop personal relations with others, children may face disproportionate exposure [10–12]. Adolescence is a developmental period where identity formation and emotional regulation are particularly sensitive to environmental influences and social pressures.

The 2026 report rightly acknowledges significant gaps, including [P94][1]:

- No consensus definition of autonomy in human–AI interaction;
- Limited transparency and measurement tools;
- Restricted access to real-world chatbot interaction data;
- Minimal longitudinal research on sustained or socially complex AI use;
- Insufficient methods to assess systemic erosion of autonomy over time.

These gaps are particularly acute and developmentally dangerous for children. Without developmental frameworks, we risk extrapolating adult findings onto young populations, overlooking the less visible harm from AI systems to children, particularly in terms of undermining their need for autonomy and agency, which play a critical role in shaping their relationship with others, their identity, and their related socio-cognitive competencies.

---

2. Rethinking Agency: Lessons from the CHAILD Project

The UKRI-funded project *Children's Agency in the Age of AI: Leveraging Interdisciplinarity (CHAILD)* was launched in February 2025 to address a core question: **What does meaningful agency look like for children growing up with AI and how can it be protected?**

Over the past year, we have conducted interdisciplinary reviews of childhood and agency; observed children interacting with AI chatbots for learning and information seeking; studied identity negotiation within social media environments; examined how AI feedback is interpreted and mediated in classroom practice; critically interrogated how 'agency' itself is invoked within AI ethics discourse and built; and built empirical foundations for a

developmental approach to AI safety. Our findings align closely with the autonomy concerns raised in the 2026 AI Safety Report but extend them in two important ways.

Moving Beyond "Choice" as an Agency Metric

Current discourses often equate agency with control or choice. Our research shows that this is insufficient. Across disciplines, agency is understood as a multi-dimensional capacity. It includes: The ability to exercise control over one's actions; The capacity to act on internal motivations; Critical thinking and independent judgement; Problem-solving and adaptive learning; The ability to set goals, monitor progress, and adjust strategies.

This indicates that agency is not simply about whether a child can click a button or access a system; it is about whether they can make meaningful choices that are grounded in understanding, intention, and self-reflection. AI safety dialogue must therefore move beyond the focus of direct harms and **recognise the underlying cumulative harms to children's agency**

Crucially, these capacities are not static. They vary by developmental stage, social context, and individual experience. A teenager's engagement with an AI system differs profoundly from that of a younger child, not only in technical skill, but in self-efficacy, intentionality, and the formation of internal values. For policymakers, this means that age-appropriate design and regulation must **account for developmental nuance**, not just age-based thresholds.

Agency Is Relational — Not Just Individual

A second major insight from our work is that agency cannot be understood solely as an individual trait. Children do not develop autonomy in isolation. Their motivations, decisions, and self-regulation are shaped by parents, peers, teachers, platforms, and increasingly, AI systems. Agency is therefore:

- **Personally exercised**, when acting independently;
- **Supported**, when scaffolded by caregivers or teachers;
- **Delegated**, when others act on their behalf;

- **Collective**, when achieved collaboratively.

These relational forms of agency are especially significant in childhood. Development inherently occurs through interaction with caregivers, institutions, technologies, and now algorithmic systems that mediate information, feedback, and identity formation. For AI governance, this insight carries important implications: AI systems are not neutral tools; they can scaffold, redirect, amplify, or constrain children's motivations. Design choices can either strengthen or weaken children's emerging self-regulation and critical capacities. Safety cannot be reduced to content moderation alone; it must include **safeguarding the developmental and social conditions that support the agency.**

---

3. Empirical Insights: Where Practice Diverges from Intention

Given the lack of evidence on children's lived experiences with emerging AI technologies, two of our recent studies highlighted critical opportunities for fostering children's agency.

Youth PRISM

In the Youth PRISM study, over 100 adolescents engaged with multiple large language model chatbots to complete learning tasks. Our findings show that:

- Adolescents articulate strong internal values around fairness and authenticity;
- Many describe cross-checking AI outputs;
- They prefer structured, context-aware responses.

However, observational data also reveals a gap between stated intention and real behaviour. Fact-checking occurs less frequently than reported. This gap between intention and behaviour has clear implications. It highlights the limits of our support for digital literacy and underscores the need for system-level design interventions that scaffold verification, reflection, and self-regulation in real time.

Algorithmic Mirror

In the Algorithmic Mirror study, adolescents donated social media data to a research platform co-developed with the MIT Media Lab to examine algorithmic categorisation of their identities.

Our key insights include:

- Lived experience matters: When adolescents are given structured opportunities to examine algorithmic classifications of their interests, they are capable of sophisticated critical reflection.
- Identity is central: Young people are deeply motivated by how platforms represent them. Algorithmic miscategorisation can feel personal, and correction mechanisms are highly valued.
- Demand for influence: Adolescents consistently express a desire not just for transparency, but for meaningful mechanisms to influence and correct the systems that shape their digital lives.

For policymakers, this suggests that transparency alone is insufficient. Structural reforms should move toward guaranteeing young users meaningful agency [9] over how they are profiled, categorised, and recommended content.

The Double-Edged Nature of AI in Children's Lives

Together, these studies highlight a dual reality: AI systems can expand access to knowledge, support learning and creativity, and sustain social connection in potentially safer environments. However, they may also undermine critical thinking if over-relied upon, diminish self-regulation if frictionless automation replaces effortful learning and shape identity and opportunity in opaque ways. Effective governance must therefore avoid binary framings of AI as either inherently harmful or inherently beneficial.

The policy challenge is not to categorise AI as harmful or beneficial, but to **amplify developmental benefits while safeguarding the conditions for agency formation.**

---

4. A Child-Centred Agency Assessment Framework

A core priority for our next phase of work is the development of a systematic, child-centred agency assessment framework. Drawing on our empirical findings, we aim to:

- Define measurable indicators of children's sense of agency in AI-mediated environments;
- Identify risk factors where systems may weaken autonomy;
- Establish protective design benchmarks;
- Conduct longitudinal observation of how sustained AI exposure shapes agency development over time.

Longitudinal evidence is particularly critical. Agency is not a fixed trait; it develops. Policymakers need tools to assess not only immediate harms, but also cumulative, developmental impacts.

5. A Policy Agenda for 2026 and Beyond

For policymakers at the Indian AI Summit, the opportunity is substantial. India represents one of the world's youngest and most digitally connected populations, while evidence of the impact of AI on their young population is scarce.

Children are not passive users of AI. They are identity-builders, learners, and future institutional stewards. The central question for 2026 should not only be how to make AI systems safe, but how to ensure they actively support children's agency and their future capacity to question, shape, and govern the systems that will shape their lives.

Closing remarks

CHAILD is funded by the first round of UKRI's new cross research council responsive mode (CRCRM) pilot scheme. The CRCRM scheme has been developed to support emerging ideas from the research community that transcend, combine or significantly span disciplines, to ensure all forms of interdisciplinary research have a home within UKRI. This provides unique opportunities for interdisciplinary research projects like CHAILD.

References

1. Bengio, Yoshua and Clare, Stephen *et al. International AI Safety Report 2026*. https://internationalaisafetyreport.org (2026).
2. Giannakos, M. *et al.* The promise and challenges of generative AI in education. *Behav. Inf. Technol.* **44**, 2518–2544 (2025).
3. Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* **15**, 6 (2025).
4. Fan, Y. *et al.* Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **56**, 489–530 (2025).
5. Macnamara, B. N. *et al.* Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cogn. Res. Princ. Implic.* **9**, 46 (2024).
6. Kosmyna, N. *et al.* Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. Preprint at https://doi.org/10.48550/arXiv.2506.08872 (2025).
7. Dratsch, T. *et al.* Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology* **307**, e222176 (2023).
8. Kücking, F. *et al.* Automation Bias in AI-Decision Support: Results from an Empirical Study. in *Studies in Health Technology and Informatics* (eds Röhrig, R. et al.) (IOS Press, 2024). doi:10.3233/SHTI240871.
9. Fischer*, K. W. & Connell, M. W. Two motivational systems that shape development: Epistemic and self-organizing. in *BJEP Monograph Series II: Part 2 Development and Motivation: Joint Perspectives* (eds Smith, L., Rogers, C. & Tomlinson, P.) (British Psychological Society, 2003). doi:10.53841/bpsmono.2003.cat529.8.
10. Moshman, D. *Adolescent Rationality and Development*. (Psychology Press, 2011). doi:10.4324/9780203835111.
11. Ivey, R., Teubner, J., Fast, N. & Iyer, R. Designing AI to Help Children Flourish. Preprint at https://doi.org/10.2139/ssrn.5179894 (2025).
12. Solyst, J. *et al.* Children's Overtrust and Shifting Perspectives of Generative AI. Preprint at https://doi.org/10.48550/arXiv.2404.14511 (2024).