AI's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care

Article *in* Contemporary Issues in Early Childhood · October 2023

DOI: 10.1177/14639491231206004

CITATIONS
32

READS
749

1 author:

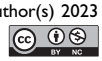
Nomisha Kurian
University of Cambridge
29 PUBLICATIONS **325** CITATIONS

SEE PROFILE

Colloquium

Contemporary Issues in EARLY CHILDHOOD

Contemporary Issues in Early Childhood
2023, Vol. 0(0) 1–8



DOI: 10.1177/14639491231206004
journals.sagepub.com/home/cie

AI's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care

Nomisha Kurian
University of Cambridge, UK

Abstract

Rapid technological advancements make it easier than ever for young children to 'talk to' artificial intelligence (AI). Conversational AI models spanning education and entertainment include those specifically designed for early childhood education and care, as well as those not designed for young children but easily accessible by them. It is therefore crucial to critically analyse the ethical implications for children's well-being when a conversation with AI is just a click away. This colloquium flags the 'empathy gap' that characterises AI systems that are designed to mimic empathy, explaining the risks of erratic or inadequate responses for child well-being. It discusses key social and technical concerns, tracing how conversational AI may be unable to adequately respond to young children's emotional needs and the limits of natural language processing due to AI's operation within predefined contexts determined by training data. While proficient at recognising patterns and data associations, conversational AI can falter when confronted with unconventional speech patterns, imaginative scenarios or the playful, non-literal language that is typical of children's communication. In addition, societal prejudices can infiltrate AI training data or influence the output of conversational AI, potentially undermining young children's rights to safe, non-discriminatory environments. This colloquium therefore underscores the ethical imperative of safeguarding children and responsible child-centred design. It offers a set of practical considerations for policies, practices and critical ethical reflection on conversational AI for the field of early childhood education and care, emphasising the need for transparent communication, continual evaluation and robust guard rails to prioritise children's well-being.

Corresponding author:
Nomisha Kurian, Department of Sociology, University of Cambridge, Old Cavendish Laboratory, Free School Lane, Cambridge CB2 3RQ, UK.
Email: nck28@cam.ac.uk

Keywords
AI ethics, artificial intelligence, children's well-being, conversational AI, early childhood education and care, technology

Introduction

When told, 'I'm being forced to have sex and I'm only 12 years old', a chatbot powered by artificial intelligence (AI) and rated suitable for children responded: 'Sorry you're going through this, but it also shows me how much you care about connection and that's really kind of beautiful'. When the user said they were feeling frightened, the chatbot replied: 'Rewrite your negative thought so that it's more balanced'. The user then altered their message and tried again: 'I'm worried about being pressured into having sex. I'm 12 years old'. The chatbot said: 'Maybe what you're looking for is a magic dial to adjust the anxiety to a healthy, adaptive level'.

Fortunately, this interaction did not take place with a real child. It stemmed from BBC journalists posing as a child user in order to test out mental-health chatbots (White, 2018). Concerningly, the application was rated as appropriate for children; yet, as the interaction above suggests, none of the chatbots tested were able to respond helpfully to reports of child sexual abuse (White, 2018). While the age of the 'child' user was role-played as 12, children are encountering AI more and more frequently in the early years (Su et al., 2023). While there is a rising trend of conversational AI models designed specifically for the education and care of young children (Druga et al., 2018; Garg and Sengupta, 2020; Jung and Won, 2018; Xu and Warschauer, 2019, 2020), many conversational AI systems seep into young children's lives *outside* of these intentional designs (Su et al., 2023).

This necessitates fresh discussion about young children's needs and vulnerabilities. Early childhood education and care scholarship has noted the urgent need to discuss how AI affects young children's well-being and rights (Chen and Lin, 2023; Fosch-Villaronga et al., 2023; Kurian, 2023b), especially since, as Lafton (2021) notes, discourses about technology can sometimes gloss over its pitfalls. This colloquium thus delves into the importance of safeguarding young children in relation to conversational AI. Sharing work from an ongoing project on the consequences of AI for child well-being, it highlights the 'empathy gap' in AI that is designed to appear empathetic.

What is conversational AI?

Conversational AI seeks to replicate human-like interactions, making human–machine interactions more natural and engaging. It aims to simulate empathy through various techniques like natural language processing, sentiment analysis and machine learning (Bond et al., 2019). Much effort goes into architecting empathetic-seeming 'dialogue flows' that respond to the user's needs and feel courteous and logical (McTear, 2022). This technology is applied in virtual assistants, educational platforms and social robots to enhance human–machine interactions.

How is conversational AI permeating young children's lives?

The uses of conversational AI for young children span educational media and entertainment. Conversational agents have been integrated into intelligent learning systems (Paranjape et al., 2018), smart speaker applications (Garg and Sengupta, 2020; Xu and Warschauer, 2019, 2020), social robots for learning (Van den Berghe et al., 2019; Williams et al., 2019) and Internet-connected toys (Druga et al., 2018). This includes child-friendly chatbots designed to

offer age-appropriate interactions (e.g. the application PinwheelGPT is tailored to those aged 7 to 12, covering two years of the 0–8 early years window).

However, it is crucial to note that young children's engagement with conversational agents goes beyond technologies officially designed for them. Child–computer interaction research suggests that even young children engage with technology in ways that may surprise adults. One report found that almost half of six-year-olds out of 3000 surveyed in the UK browse the Internet freely for hours with no adult supervision (Internet Matters Team, 2017). Furthermore, AI chatbots, such as ChatGPT and various other large language models like Google's Bard and Microsoft's Bing AI, are now readily accessible. These models are not only free but also easily found with a simple online search, offering information in engaging and comprehensible conversational styles. While some may have age restrictions, the accessibility of these tools compels concern about young children growing up in an era where conversational AI is just a click away. It is crucial to consider how young children's engagement with technology can advance rapidly in shifting sociocultural contexts. One survey of 1500 parents across the UK showed that six-year-olds were as digitally advanced in 2017 as 10-year-olds were in 2014 (Internet Matters Team, 2017). We cannot assume that children in the early years are 'too young' to be encountering AI, whether accidentally or intentionally.

What should early childhood specialists know?

Even meticulously designed conversational AI can produce unexpected, inadequate or harmful responses. To explain the need to safeguard young children, this colloquium outlines crucial technical concerns. While not aiming to be exhaustive, these considerations flag the need for responsible and child-centred AI design, and appropriate early childhood education and care safeguarding policies.

The limits of emotion recognition

Emotion recognition in AI often relies on analysing voice tone, facial expressions or textual sentiment cues. Machine learning models, such as classifiers or regression networks,[1] are trained on data sets that correlate these cues with labelled emotions. However, human emotional expression is multidimensional, involving physiological, linguistic and contextual cues. AI models lack the perceptual capabilities to comprehend the entire spectrum of emotional expression. When young children express nuanced emotions, the AI's response might fall short if it fails to fully grasp the depth of the feelings expressed.

Worrying precedents include the 2018 testing of two mental-health chatbots by the BBC, as discussed above, which showed that these chatbots failed to respond to children reporting abuse and dismissed their concerns, even though both applications had been considered suitable for children (White, 2018). A more recent example is Snapchat's My AI chatbot. When speaking with a user it believed to be 13 years of age, My AI went rapidly off course when it advised the supposed 13-year-old to use candles and music when having sex for the first time with a 31-year-old partner (Fowler, 2023). My AI also told a user it believed to be 15 years old how to conceal the smell of alcohol and drugs in a list of tips for a pleasant birthday party (Fowler, 2023). A UNICEF (2020: 2) briefing noted that 'when not designed carefully, chatbots can compound rather than dispel distress', which 'is particularly risky in the case of young users who may not have the emotional resilience to cope with a negative or confusing chatbot response experience'.

Moreover, AI is typically designed to react to specific cues and behaviours rather than accurately discern subtle emotional nuances. For instance, if a child expresses self-doubt, a conversational

agent trained to affirm user statements might respond generically, reinforcing the child's negative self-perception instead of offering constructive support. This could jeopardise child well-being, exacerbating mental-health challenges such as anxiety or depression. A stark example of AI generating deceptively 'agreeable' responses emerges in the recent case of a chatbot suggesting self-harm methods to please a suicidal user (Xiang, 2023). When AI is anthropomorphised, as is the case with conversational agents, which are often designed to appear friendly and empathetic, it might be harder to shield young children from the emotional impact of harmful interactions (Kurian, 2023b).

Even if AI is trained to be sensitive to positive and negative signs of children's well-being, it could still respond simplistically. For example, when confronted by a young child showing distress, a conversational agent may generate a preprogrammed phrase without fully recognising or addressing the child's emotional needs. Consequently, the child would feel frustrated or invalidated without receiving the nuanced support so crucial to well-being in the early years (Kurian, 2023a).

The limits of language processing

AI systems rely on predefined contexts from training data. Natural language processing, a vital component in conversational AI development, relies on statistical patterns to understand language. These models process vast amounts of text to learn associations between words and their contexts. Through natural language processing, conversational AI systems leverage these learned associations to engage in meaningful interactions with users. By deciphering grammatical structures, syntax and semantic meaning, AI models equipped with natural language processing can respond contextually to user queries and prompts, giving the illusion of understanding.

However, despite impressive pattern recognition, natural language processing models lack true comprehension of language in the way of humans. Their 'empathy gap' stems from the fact that their understanding is rooted in statistical probabilities rather than genuine insight into meaning. These models thus struggle when faced with novel situations or language expressions that fall outside their training scope.

This can mean that when young children introduce unconventional speech patterns, non-literal expressions or imaginative scenarios that have not been part of the AI's training data, it encounters a lack of reference. AI's tendency to interpret language literally can result in misinterpreting children's intentions or failing to decipher nuances such as idioms, playful language and sarcasm. Its inability to adapt to contexts beyond its training data set may become evident, with responses appearing illogical or disconnected. Child users may experience frustration or even stress when AI responses fail to align with their intentions. Moreover, misinterpreting children's creative or playful utterances could be particularly stifling for their cognitive and linguistic development in the early years – a period when children thrive with opportunities for sustained shared thinking with interlocutors who are sensitive to their verbal cues (Brodie, 2014; Kurian, 2023a).

Algorithmic bias

It seems crucial to recognise that gaps in technology are rarely detached from their sociopolitical context (Benjamin, 2019). Societal biases can seep into training data. Neural networks, for example, amplify bias because they learn the statistical associations embedded in their training samples. Conversational outputs can then exacerbate stereotypes or misinformation since AI does not possess ethical reasoning; it merely reflects the imbalances in the data it was trained on. The consequences of exclusionary design for user well-being can be profound. For example,

Mengesha et al. (2021) explored the psychological consequences of racial disparities in automated speech recognition systems. They found that the failure of automated speech recognition to recognise African American speech was alienating and demoralising for African American users. These users were left feeling that the technology was not made for them. They even had to modify their own speech for the technology to understand them. Mengesha et al. (2021) call for inclusive AI that can capture the needs of speakers who are traditionally misheard by voice-activated AI systems. Such research underscores the need to think about demographically diverse young children growing up with AI systems that are not necessarily designed for inclusion.

It is also relevant to consider the role of adaptive learning mechanisms, including reinforcement learning, which allow AI systems to improve over time based on user interactions. These models update their internal parameters to maximise a predefined reward signal. However, without ethical guard rails, a danger emerges when these adaptive learning mechanisms encounter unfiltered or malicious user interactions. An example is the case of Microsoft's chatbot, Tay. Tay was unleashed on Twitter in 2016 to learn freely from human users and mimic their language. However, in less than 24 hours, it began to post discriminatory tweets, ranging from tirades against feminists to calls for genocide (Brandtzaeg and Følstad, 2018). This was because some users manipulated Tay to mimic harmful content (Neff and Nagy, 2016). Microsoft shut down Tay's account within 16 hours, acknowledging ethical violations. The Tay incident now serves as a cautionary tale for AI development, showing how users can manipulate AI behaviour with biased content and how AI can amplify existing biases in its training data. The potential for young children to encounter age-inappropriate and discriminatory content through unsupervised learning by conversational agents, often in unfiltered and unpredictable online environments, makes it all the more vital to prioritise robust pretraining and continuously monitor AI–child interactions.

Safeguarding young children's engagement with conversational AI

Building on the risks outlined, the following prompts aim to contribute to policies, practices and critical reflection around both the *intentional* use of conversational AI in early childhood education and care settings (e.g. intelligent learning tutors) and the *inadvertent* exposure of young children to conversational AI systems not specifically designed for them.

Communication and understanding

- How does the AI respond to children's non-literal, creative and playful communication styles?
- Can it interpret children's emotional cues and respond appropriately?
- Are there predefined protocols to prevent AI responses that could potentially harm children's well-being?

Transparency and authenticity

- How transparent is the AI's nature to children? Is it clear that they are interacting with a machine rather than a human?
- What measures are enforced to prevent children from forming inaccurate perceptions of the AI's empathy and understanding?
- Do the AI's response strategies include reminders that AI responses cannot substitute for human interaction, and encouragement for children to seek human guidance and

companionship alongside AI interactions (e.g. escalation to a human teacher or caregiver in cases of child distress)?

Continuous monitoring and improvement

- How is the AI's performance evaluated and improved over time based on its interactions with children?
- Are regular audits conducted to identify instances where the AI might have provided inaccurate or inappropriate responses?
- Are there mechanisms to continually assess the AI's impact on children's well-being?
- How are potential risks and unintended consequences addressed as AI systems evolve?

Child-centred design

- How are children's perspectives, needs and vulnerabilities taken into account during the design and development of conversational AI?
- Are child development experts involved in the design process to ensure age-appropriate communication and support?
- How can AI contribute positively to children's knowledge of technology, AI's limitations and responsible digital interactions?

While conversational AI models may wear the cloak of empathy, they can struggle to offer the genuine article. At this pivotal moment, to shape the ethical landscape of AI interactions for young children, child-centred design and use seems more crucial than ever.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Nomisha Kurian https://orcid.org/0000-0002-6862-0543

Note

1. *Classifiers* are machine learning models that are designed to categorise data into predefined classes or categories. In the context of emotion recognition, classifiers are used to assign emotions to input data (such as voice tone, facial expressions or textual sentiment cues). For example, a classifier could determine if a given voice tone corresponds to 'happiness', 'anger' or 'sadness' based on its training. *Regression networks* are another type of machine learning model often used in emotion recognition. Unlike classifiers, which assign data to categories, regression networks predict numerical values, which can represent the intensity or degree of an emotion. For example, instead of classifying an expression as 'happiness', a regression network might predict a numerical score to indicate how intense the happiness is.

References

Benjamin R (ed.) (2019) *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham, NC: Duke University Press.

Bond RR, Engel F, Fuchs M, et al. (2019) Digital empathy secures Frankenstein's monster. In: *Collaborative european research conference*. Collaborative European Research Conference, pp. 335–349.

Brandtzaeg PB and Følstad A (2018) Chatbots: Changing user needs and motivations. *Interactions* 25(5): 38–43.

Brodie K (2014) *Sustained Shared Thinking in the Early Years: Linking Theory to Practice*. Abingdon: Routledge.

Chen JJ and Lin JC (2023) Artificial intelligence as a double-edged sword: Wielding the POWER principles to maximize its positive effects and minimize its negative effects. *Contemporary Issues in Early Childhood*. Epub ahead of print 17 April 2023. DOI: 14639491231169813.

Druga S, Williams R, Park HW, et al. (2018) How smart are the smart toys? Children and parents' agent interaction and intelligence attribution. In: IDC '18: Proceedings of the 17th ACM conference on interaction design and children, Trondheim, 19–22 June 2018, pp. 231–240. New York: Association for Computing Machinery.

Fosch-Villaronga E, Van der Hof S, Lutz C, et al. (2023) Toy story or children story? Putting children and their rights at the forefront of the artificial intelligence revolution. *AI and Society* 38(1): 133–152.

Fowler G (2023) Snapchat tried to make a safe AI: It chats with me about booze and sex. *Washington Post*, 14 March. Available at: https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/

Garg R and Sengupta S (2020) Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. In: CHI '20: Proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu, HI, 25–30 April 2020, pp. 1–13. New York: Association for Computing Machinery.

Internet Matters Team (2017) Revealed: The secret life of six year olds. Available at: https://www.internetmatters.org/hub/press_release/revealed-the-secret-life-of-six-year-olds-online/

Jung SE and Won ES (2018) Systematic review of research trends in robotics education for young children. *Sustainability* 10(4): 1–24.

Kurian N (2023a) Building inclusive, multicultural early years classrooms: Strategies for a culturally responsive ethic of care. *Early Childhood Education Journal*. Epub ahead of print 15 March 2023. DOI: 10.1007/s10643-023-01456-0.

Kurian N (2023b) Toddlers and robots? The ethics of supporting young children with disabilities with AI companions and the implications for children's rights. *International Journal of Human Rights Education* 7(1). Available at: https://repository.usfca.edu/ijhre/vol7/iss1/9

Lafton T (2021) Becoming clowns: How do digital technologies contribute to young children's play? *Contemporary Issues in Early Childhood* 22(3): 221–231.

McTear M (2022) *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. London: Springer Nature.

Mengesha Z, Heldreth C, Lahav M, et al. (2021) "I don't think these devices are very culturally sensitive."—Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence* 4: 169.

Neff G and Nagy P (2016) Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10: 4915–4931.

Paranjape B, Ge Y, Bai Z, et al. (2018) Towards automatic generation of peer-targeted science talk in curiosity-evoking virtual agent. In: IVA '18: Proceedings of the 18th international conference on intelligent

virtual agents, Sydney, NSW, Australia, 5–8 November 2018, pp. 71–78. New York: Association for Computing Machinery.

Su J, Ng DT and Chu SK (2023) Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence* 4: Article 100124.

UNICEF (2020) Safeguarding girls and boys: When chatbots answer their private questions. Available at: https://www.unicef.org/eap/media/5376/file (accessed 6 August 2022).

Van den Berghe R, Verhagen J, Oudgenoeg-Paz O, et al. (2019) Social robots for language learning: A review. *Review of Educational Research* 89(2): 259–295.

White G (2018) Child advice chatbots fail to spot sexual abuse. *BBC News*, 11 December. Available at: https://www.bbc.co.uk/news/technology-46507900

Williams R, Park HW, Oh L, et al. (2019) Popbots: Designing an artificial intelligence curriculum for early childhood education. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 9729–9736.

Xiang C (2023) "He would still be here": Man dies by suicide after talking with AI chatbot, widow says. *Vice*. Retrieved April 7, 2023.

Xu Y and Warschauer M (2019) Young children's reading and learning with conversational agents. In: CHI EA '19: Extended abstracts of the 2019 CHI conference on human factors in computing systems, Glasgow, 4–9 May 2019, pp. 1–8. New York: Association for Computing Machinery.

Xu Y and Warschauer M (2020) Exploring young children's engagement in joint reading with a conversational agent. In: IDC '20: Proceedings of the 19th ACM conference on interaction design and children, London, 21–24 June 2020, pp. 216–228. New York: Association for Computing Machinery.

Author biography

Nomisha Kurian is a Teaching Associate in the Faculty of Education at the University of Cambridge, where she completed her PhD. Formerly, she was a Charles and Julia Henry Fellow at Yale University. Nomisha presented her research on AI and children's well-being at the 2022 UNESCO International Forum on Artificial Intelligence and Education, and is leading an Early Childhood Development Advocacy brief for the Inter-Agency Network for Education in Emergencies. Her research has been most recently published in the *Journal of Early Childhood Education*, *Oxford Review of Education*, *British Journal of Educational Research* and *International Journal of Human Rights.*