From Ethical Guidelines to Practical Guidance to Develop Trustworthy Conversational Agents for Children

Marina Escobar-Planas, Emilia Gómez, Carlos-D. Martínez-Hinarejos

Abstract Conversational agents (CAs), such as chatbots or home assistants, have gained a lot of popularity in the last decade and, despite their adult-centred design, have become part of many children's lives. However, children's voices, behaviours or needs differ from adults, and these differences pose a challenge for CAs. In addition, considering the vulnerability of children, there is a clear need for the development of trustworthy systems that take children's needs into account. In this article we present relevant ethical guidelines on AI and summarize our work on adapting them to the specific case of CAs and children. In particular, we present some concrete recommendations for developers to support the ethical design of CAs for children.

1 Introduction

A conversational agent (CA) refers to a computer program that supports conversational interactions with humans, usually through text or voice [14]. These systems traditionally contain the following modules: Automatic speech recognition (ASR) to transform speech into text; Natural language understanding (NLU) to extract a semantic interpretation from the text; Dialogue manager (DM) to control the actions of the system; Natural language generation (NLG) to generate natural language text; and Text to speech (TTS) to generate a speech from text.

These agents have become widely popular in the last decade, and they are often designed having the average person in mind. This may cause the CA to fail with specific persons, as children, who present unique challenges for this technology [10,

Marina Escobar-Planas and Carlos-D. Martínez-Hinarejos
Universitat Politècnica de València, Camí de Vera, s/n, 46022, València, Spain.

Marina Escobar-Planas and Emilia Gómez
European Commission, Joint Research Center, C. Inca Garcilaso, 3, 41092, Seville, Spain.
e-mail: marescplajob@gmail.com

15]: ASR modules need to understand children's speech [17, 16, 3] and DM modules should consider repair strategies to handle the conversation when the system does not understand the user's input [6, 13] or when usual actions are not convenient for children because of their age[19]. However, children are common users of CAs such as voice assistants [4], and as a vulnerable population, it is important to consider the ethical implications that these systems can bring to the lives of the very young.

There has been extensive research aimed at improving different ethical aspects of child-computer interaction, including the influence of conversational agents on children's choices [2] and the advantages of explaining to children the absence of psychological traits of social robots [18]. Nevertheless, our work provides a comprehensive overview of the field from a broad perspective. Starting from general ethical guidelines, to more practical guidance.

In this article, we summarize existing initiatives dealing with ethical considerations in the design of children-centric artificial intelligence (AI) (Section 2), and summarize the main outcomes of our recent research [9] where we developed ethical guidelines for the development of trustworthy CAs for children (Section 3). Finally, in Section 4, we provide some conclusions and summarize our future work on this area.

2 AI and Ethics

In recent years, there has been an increased attention on the impact of AI systems in people's lives. AI's respect for fundamental rights has been a consistent goal in reports from international institutions. For instance, the European Commission (EC)'s High Level Expert Group on AI released an assessment list [1] to self-assess if an AI system is "trustworthy" by embracing a set of seven requirements: (1) Human agency and oversight; (2) Technical robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination and fairness; (6) Societal and environmental well-being; and (7) Accountability.

Although these initiatives focus on the general population, recent work by UNICEF and the EC's Joint Research Centre have analyzed how to design AI systems that respect children's rights [7, 5]. They emphasize the significance of the 6th requirement (societal and environmental well-being) for children, suggesting the use of AI only for crucial tasks. Other aspects related to children include requirement 3 (Privacy and data governance), and requirement 4 (Transparency).

3 Ethical design of CAs for children

We summarize here the outcomes of our recent study, aimed to adapt ethical guidelines for AI systems to the specific case of CAs and children [9]. A team of four experts in computer science, AI ethics, and children's rights scored and commented

each item of the assessment list on trustworthy AI (ALTAI) in terms of relevance and particular considerations for CAs and children. We followed a Delphi method [12] approach to perform a risk level analysis [11] as follows. The individual ratings were first analyzed to identify critical points and disagreements, which were discussed and resolved at an expert meeting in order to reach a consensus. A thematic analysis was additionally carried out on the annotated comments provided by the experts. The main findings of the study are illustrated in Fig. 1 and summarized below.

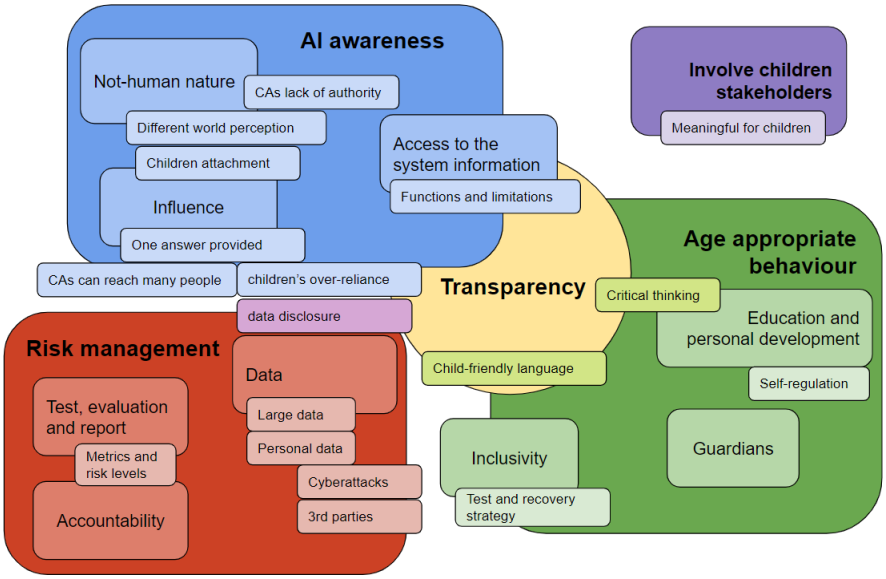

Fig. 1 Thematic analysis of experts' comments on the application of ALTAI to the specific case of CAs and children.

Stakeholders involvement. It is important to involve stakeholders (e.g. children, guardians and teachers) during the whole design process of a CA. This should be done in a meaningful way giving that, for instance, children cannot be considered as a work force when developing a commercial product.

Risk management. Considering children as a vulnerable population, risk management should be highly relevant during the CA development. High privacy and security measures regarding data storage are needed to ensure that personal data is not accessible to third parties. In addition, metrics and risk levels should be defined to track the system performance, facilitating its testing and evaluation as well as external audits. In addition, users' capability to write reports about the system can facilitate the identification of risks and errors. Transparency can also be used to in-

form about privacy concerns and diminish children's data disclosure.

AI Awareness. Due to the different perception of the world that children may have, it is important to highlight the non-human nature of the CA in order to minimize children's attachment to them and CAs influence on the child. Regarding influence, maximising the user's agency will also be beneficial (e.g. in the case of a CA system that offers games to play, providing several options). Transparency can be used to provide constant access to the system's information including nature, functions and limitations.

Age appropriate behaviour. Improving inclusivity is very important for children education and development. It is therefore important to mitigate the technical difficulties that CAs may have when interacting with children and other minorities. A good recovery strategy, to continue interaction after a breakdown, may help. We should also address guardians as responsible for the children, implementing mech-

Table 1 Recommendations to the design of a CA that generates a list of preferred toys/games for children.

General	Specific	Particular measures
Stakeholders involvement	Consider stakeholders in all CA lifecycle	- Define features (e.g. age ranges, max interaction time). - Consult stakeholders throughout design, implementation, and evaluation.
Risk management	Privacy measures	- Minimize the personal data to be stored. - Do not allow additional usages/transfer of stored data.
	Security measures	- Reduce standard black boxes and search engine usage in DM and NLG. - Incorporate a control mechanism for online search. -Define trigger keywords for guardian involvement (e.g. weapons, sex). - Store data in a safe server with cybersecurity measures. - Define metrics for risk management, e.g. time spent, guardian's calls.
	Facilitate reports	-After the interaction, gather feedback from children and guardians. - Offer accessible error reporting and mention it in the welcome message.
AI awareness	Access to the system information	- Include concise relevant CA information to the welcome message and pointers to additional details. - Inform about the system's not-human, not-living and not-feeling nature. - Inform about the system's confidentiality and algorithmic decisions.
	Influence	- Configure the system to display at least 3 suggestions.
Age approp. behavior	Guardians	- Split welcome message into: guardian & child. Consider two consents. - Invoke guardian in security issues (e.g. dangerous requests or persistent breakdowns).
	Education and self-development	- Define toys-classification to benefit children's development. Consider them for suggestions. - Consider gender bias in recommended items. - Control and communicate the time spent on the interaction.
	Inclusivity	- Guess/ask for age information at the beginning of the interaction. - Define functionality as "wish list" if a child is recognized. - Adapt the list of recommended items to age. - Adapt the vocabulary of the interaction to age. - Choose an inclusive ASR module. - Minimize neutral responses in breakdowns.

anisms for double consent, but also approaching them when a problem is encountered. Transparency can be applied using a language adapted to the age of the user. It can also enhance the user's critical thinking and self-regulation.

Transparency. As shown in the paragraphs before, transparency was identified as a crucial tool to fight many of the other previously mentioned critical considerations. Facilitating access to information regarding the nature of the system, privacy and limitations, as well as using age-appropriate language, could improve the trustworthiness of the CA.

Further details on the methodology and results of this study can be found in [9]. All these critical points have been taken into account in a subsequent study [8], which has addressed the application of these guidelines to the design of a particular CA that helps to generate a list of the child's preferred toys and games. The application of recommendations is shown in Table 1 and should be implemented in different stages of the development, such as initial design (e.g. involvement of stakeholders), technology (e.g. optimizing algorithms for children), interaction itself (e.g. consideration of guardians) or post-interaction (e.g. system auditing).

4 Conclusions and future work

This article provides an overview of relevant ethical considerations for the design of children-centric CAs. It summarizes existing initiatives on AI and ethics, and provides the main outputs of our recent research on the development of trustworthy CAs for children: involving stakeholders, enhancing the risk management system, raising awareness among children about AI, implementing age-appropriate behavior, and promoting transparency.

While this paper provides a general overview of the problem, we recognize that the recommendations we provide are still very broad. The application of these measures needs to be personalized to every CA system and adapted to the needs and objectives of every project. Therefore, we see the need for further research to develop more concrete guidance and tools to ensure that CAs for children are developed in an ethical and responsible manner.

In the future, we plan to continue our research in this area by conducting empirical studies to validate our recommendations and to identify new challenges and opportunities in the development of trustworthy CAs for children. We believe that this research will contribute to the creation of ethical, safe, and accessible AI systems for children, fostering their development and well-being.

Acknowledgements This work was carried out with the support of the Joint Research Centre of the European Commission in the framework of the Collaborative Doctoral Partnership Agreement No.35500.

References

1. P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, L. Bouarfa, R. Chatila, M. Coeckelbergh, V. Dignum, et al. *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission, 2020.
2. H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. Nudges with conversational agents and social robots: A first experiment with children at a primary school. In *Conversational Dialogue Systems for the Next Decade*, pages 257–270. Springer, 2020.
3. V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9), 2022.
4. C. Biele, A. Jaskulska, W. Kopec, J. Kowalski, K. Skorupska, and A. Zdrodowska. How might voice assistants raise our children? In W. Karwowski and T. Ahram, editors, *Intelligent Human Systems Integration 2019*, pages 162–167, Cham, 2019. Springer International Publishing.
5. V. Charisi, S. Chaudron, R. Di Gioia, R. Vuorikari, M. Escobar-Planas, I. Sanchez, and E. Gomez. *Artificial intelligence and the rights of the child : towards an integrated agenda for research and policy*. Publications Office of the European Union, 2022.
6. Y. Cheng, K. Yen, Y. Chen, S. Chen, and A. Hiniker. Why doesn't it work? voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM conference on interaction design and children*, pages 337–348, 2018.
7. V. Dignum, M. Penagos, K. Pigmans, and S. Vosloo. *Policy guidance on AI for children*. Communications of UNICEF, 2021.
8. M. Escobar Planas, E. Gómez, and C.-D. Martínez-Hinarejos. Enhancing the Design of a Conversational Agent for an Ethical Interaction with Children . In *Proc. IberSPEECH 2022*, pages 171–175, 2022.
9. M. Escobar-Planas, E. Gómez, and C.-D. Martínez-Hinarejos. Guidelines to develop trustworthy conversational agents for children. *arXiv*, 2022.
10. J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 82–90, 2017.
11. N. Kovačević, A. Stojiljković, and M. Kovač. Application of the matrix approach in risk assessment. *Operational Research in Engineering Sciences: Theory and Applications*, 2(3):55–64, 2019.
12. H. A. Linstone, M. Turoff, et al. *The delphi method*. Addison-Wesley Reading, MA, 1975.
13. L. Mavrina, J. Szczuka, C. Strathmann, L. M. Bohnenkamp, N. Krämer, and S. Kopp. "alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant. *Frontiers in Computer Science*, 4:791704, 2022.
14. M. McTear. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251, 2020.
15. J. H. Nilsen and K. Røyneland. "it knows how to not understand us!" a study on what the concept robustness entails in design of conversational agents for preschool children. Master's thesis, 2019.
16. E. Ong, J. B. Alburo, C. R. De Jesus, L. K. Gilig, and D. T. Ong. Challenges posed by voice interface to child-agent collaborative storytelling. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE, 2019.
17. R. Sinha and S. Shahnawazuddin. Assessment of pitch-adaptive front-end signal processing for children's speech recognition. *Computer Speech Language*, 48:103–121, 10 2018.
18. C. L. v. Straten, J. Peter, R. Kühne, and A. Barco. Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2):1–22, 2020.
19. G. Wang, J. Zhao, M. Van Kleek, and N. Shadbolt. Informing age-appropriate ai: Examining principles and practices of ai for children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2022.