UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Trustworthy Conversational Agents for Children

October 2024

Author: Marina Escobar Planas

Directors: D. Emilia Gómez Gutiérrez
D. Carlos David Martínez Hinarejos

Acknowledgements

Como de costumbre dejo los agradecimientos para el final y por poco se me olvida. Aunque siempre creo que es importante agradecer las cosas, el estrés del momento por poco me hace obviar este apartado. Y eso no puede ser.

Por supuesto, caigo en el tópico, empiezo por agradecer a mi familia todo el apoyo recibido durante todos los años de mi doctorado. Han vivido conmigo muchos más, pero estos años tiene especial mérito. Empezando por mis padres que se han apuntado a viajar conmigo a alguna conferencia nacional, pero especialmente a Manu, quien ha estado ahí durante todo el camino, escuchándome e incluso arrimando el hombro con parte del trabajo. Le he tenido a él como a buena parte de mis amigos ayudándome a hacer historias para niños jajaja. Y los amigos, claro, la otra familia, que han sufrido verme menos durante mi tesis y aun así han estado ahí para mí. Mis aljarafeños, mis frikis o mis vacas yeyés son solo algunos de los ejemplos.

Me he encontrado a gente maravillosa, como Söngul, que me animó a empezar este viaje, Emilia que me ha acompañado y guiado en cada paso (con todas sus sonrisas y lágrimas), o Carlos que me ha dado un mágico empujón final cuando ya no quedaba tiempo. Isabelle me ayudó a 'parir mi primera piña' (mi primer paper), David me ha animado y ayudado a identificar e intentar acallar el síndrome del impostor... bueno, el grupo de investigación HUMAINT ha sido donde este trabajo ha nacido, crecido y volado. El Joint Researh Centre de la Comisión Europea me ha permitido estudiar en profundidad un tema que me ha apasionado. Desde Jutta, hasta Sven, pasando por Carlos. Mi amiga Olivia... la maravillosa Lily, y muchos más que quedan sin mención por no querer aburrir demasiado al personal. Todos ellos me han dado la oportunidad, pero también me han ayudado a abrir las alas. Gracias.

El doctorado no ha sido un camino de rosas, e incluso cuando me prometí al principio que no echaría muchas horas extra, me he desvivido y muerto a trabajar jajaja. Sobre todo con el COVID en la fase inicial de la tesis. Y sin querer pintarlo demasiado duro, pero... como he dicho, no ha sido un camino de rosas ni mucho menos. Especialmente el último año, en el que mi beca con el JRC terminó y me vi obligada a tirar del paro. Un último año en el que mi padre ha superado un cáncer, y en el que los servicios sociales han decidido unilateralmente que mientras yo estuviera escribiendo este documento no podía adoptar y ser madre. (...) Ha sido mucha presión, y obviamente estoy feliz de haber terminado con la redacción.

Pero bueno, hemos pasado todo. Ahora también es momento de echar la vista atrás y reflexionar sobre la magia que encierra este documento ¿no? Que también tiene mucha. He sido capaz de trabajar cuatro años en este tema sin aburrirme. Este documento es el resultado de juntar lo asombroso de la inteligencia artificial (robots y otros agentes conversacionales) con niños y sus sorprendentes y tiernas respuestas. Han sido cuatro años extraordinarios, compartiendo experiencias con gente de la UPO, pero también con otros estudiantes en Portugal, expertos en ética en Finlandia, investigadores de Italia, estudiantes y supervisores de Valencia y gente de laboratorios de robótica social en Suecia. Hay mucha magia escondida detrás de este trabajo. Muchos momentos que me hacen sonreír. Las palabras de este manuscrito esconden innumerables vivencias que no creo que tuviera espacio para compartir. Así que bueno, espero que el resultado de dichas vivencias mantenga al menos una pequeña parte de esa magia, y sea del gusto del lector.

Gracias por el tiempo del que haya podido llegar hasta aquí leyendo mis desvaríos, que sinceramente, no creo que seáis muchos. Espero que en vuestra vida también tengáis historias que encierren incontables momentos mágicos que no os quepan en un manuscrito.

Marina Escobar-Planas
Octubre 2024

Abstract

This thesis explores the development of trustworthy conversational agents (CAs) that may interact with children, aligning with the European Commission's human-centred artificial intelligence (AI) initiative. CAs are becoming increasingly integral to children's lives, used in various sectors such as education, entertainment, and healthcare. However, children, as a unique user group, present particular needs and challenges, necessitating the development of guidelines and rigorous evaluations to ensure that these systems are trustworthy.

The research begins by examining the fundamentals of CAs, focusing on their interaction with children. A literature review highlighted both the benefits and risks associated with these technologies, emphasising the need for multidisciplinary approaches to mitigate potential ethical concerns. The concept of trustworthiness is introduced, as well as a significant research gap in CA ethical development.

To address this gap, the thesis integrates knowledge from multiple domains, combining theoretical exploration with empirical studies. This approach leads to the development of a set of guidelines that align with the needs of children and CAs, highlighting the importance of transparency, age-appropriate behaviour, AI awareness, stakeholder involvement, and risk management.

These guidelines were then applied in the design and development of a collaborative storytelling CA. The system's trustworthiness was assessed using the ALTAI framework, demonstrating an improvement in trustworthiness. This evaluation is further supported by an experimental study, examining how children perceive and interact with these systems, with a specific focus on risks such as misinterpreting the CA's non-human nature and children's data sharing behaviour. Key findings underscored the positive impact of transparency, and the need for a more effective approach to clarify the CA's agency.

The contributions of this thesis provide both theoretical insights and practical guidance for developing trustworthy CAs for children, ensuring they are ethically aligned with societal values and children's rights. Future research is encouraged to expand these findings to broader cultural contexts and develop more nuanced evaluation tools for assessing trustworthiness for children.

Resumen

Esta tesis explora el desarrollo de agentes conversacionales (ACs) confiables que puedan interactuar con niños, en consonancia con la iniciativa de la Comisión Europea para que el desarrollo de la inteligencia artificial (IA) se centre en el ser humano. Los ACs están convirtiéndose en una parte cada vez más integral de la vida de los niños, siendo utilizados en diversos sectores como la educación, el entretenimiento y la atención médica. Sin embargo, los niños, como un grupo de usuarios único, presentan necesidades y desafíos particulares, lo que hace necesario el desarrollo de directrices y evaluaciones rigurosas para asegurar que estos sistemas sean confiables.

La investigación comienza examinando los fundamentos de los ACs, enfocándose en su interacción con los niños. Una revisión de la literatura destaca tanto los beneficios como los riesgos asociados con estas tecnologías, subrayando la necesidad de enfoques multidisciplinarios para mitigar posibles preocupaciones éticas. Se introduce el concepto de confiabilidad, así como la falta de investigación existente en el desarrollo ético de los ACs.

Para subsanar esta deficiencia, la tesis integra conocimientos de múltiples dominios, combinando la exploración teórica con estudios empíricos. Este enfoque lleva al desarrollo de un conjunto de directrices que abordan las particularidades de la interacción niño-AC, destacando la importancia de la transparencia, el comportamiento adecuado a la edad, el fomento de la conciencia de qué es una IA, la participación de partes interesadas y la gestión de riesgos.

Estas directrices se aplican luego en el diseño y desarrollo de un AC colaborativo de narración de cuentos. La confiabilidad del sistema se evalúa utilizando ALTAI y demostrando una mejora en la confiabilidad. Además, dicha evaluación se complementa con un estudio experimental que examina cómo los niños perciben e interactúan con estos sistemas, dando un protagonismo específico a los riesgos como la interpretación errónea de la naturaleza no humana del AC y el comportamiento de los niños en la revelación de datos. Los hallazgos resaltan el impacto positivo de la transparencia y la necesidad de un enfoque más eficaz para aclarar la agencia del AC.

Las contribuciones de esta tesis proporcionan tanto ideas teóricas como orientaciones prácticas para el desarrollo de ACs confiables para niños, asegurando que estén alineados éticamente con los valores sociales y los derechos de los niños. Se anima a futuras investigaciones a expandir estos hallazgos a contextos culturales más amplios y a desarrollar herramientas de evaluación más completas para evaluar la confiabilidad en los ACs entorno a los niños.

Resum

Aquesta tesi explora el desenvolupament d'agents conversacionals (ACs) confiables que puguen interactuar amb xiquets, en alineació amb la iniciativa de la Comissió Europea per que el desenvolupament de la intel·ligència artificial (IA) se centre en l'ésser humà. Els ACs s'estan convertint en una part cada vegada més integral de la vida dels xiquets, sent utilitzats en diversos sectors com l'educació, l'entreteniment i l'atenció mèdica. No obstant això, els xiquets, com a grup d'usuaris únic, presenten necessitats i desafiaments particulars, el que fa necessari el desenvolupament de directrius i avaluacions rigoroses per a assegurar que aquests sistemes siguen confiables.

La investigació comença examinant els fonaments dels ACs, centrant-se en la seua interacció amb els xiquets. Una revisió de la literatura destaca tant els beneficis com els riscos associats amb aquestes tecnologies, subratllant la necessitat d'enfocaments multidisciplinaris per a mitigar possibles preocupacions ètiques. S'introdueix el concepte de confiabilitat, així com la manca d'investigació existent en el desenvolupament ètic dels ACs.

Per a abordar aquesta bretxa, la tesi integra coneixements de múltiples dominis, combinant l'exploració teòrica amb estudis empírics. Aquest enfocament porta al desenvolupament d'un conjunt de directrius que aborden les particularitats de la interacció xiquet-AC, destacant la importància de la transparència, el comportament adequat a l'edat, el foment de la consciència de què és una IA, la participació de parts interessades i la gestió de riscos.

Aquestes directrius s'apliquen després en el disseny i desenvolupament d'un AC col·laboratiu de narració de contes. La confiabilitat del sistema s'avalua utilitzant ALTAI i demostrant una millora en la confiabilitat. A més, aquesta avaluació es complementa amb un estudi experimental que examina com els xiquets perceben i interactuen amb aquests sistemes, donant un protagonisme específic als riscos com la interpretació errònia de la naturalesa no humana de l'AC i el comportament dels xiquets en la revelació de dades. Els resultats ressalten l'impacte positiu de la transparència i la necessitat d'un enfocament més eficaç per a aclarir l'agència de l'AC.

Les contribucions d'aquesta tesi proporcionen tant idees teòriques com orientacions pràctiques per al desenvolupament d'ACs confiables per a xiquets, assegurant que estiguen alineats èticament amb els valors socials i els drets dels xiquets. S'encoratja a futures investigacions a expandir aquests resultats a contextos culturals més amplis i a desenvolupar eines d'avaluació més matisades per a valorar la confiabilitat en els ACs involucrats amb xiquets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In recent years, Conversational Agents (CAs) (McTear, 2020), including virtual assistants and large language models (LLMs)(Zhao et al., 2023), have gained widespread popularity, transforming how people interact with technology in various aspects of life. However, this technology has also introduced new ethical risks linked to AI development. For instance, the release of the LLM model GPT-3, revealed significant biases based on gender, race and religion (Brown et al., 2020). While subsequent iterations of GPT have made considerable strides in addressing some ethical considerations, the broader research landscape of human-computer interaction suggests a persistent gap in the ethical design and implementation of CAs (Diederich et al., 2022).

While these technologies primarily target adult users, particularly those who make purchasing decisions, their accessibility and utility have also captured the attention of younger demographics, including children under eighteen. Children's innate curiosity and adaptability toward new technologies have made them active users of CAs, engaging in educational tasks, seeking answers to their questions, and participating in playful activities through these digital tools (Garg and Sengupta, 2020; Lovato et al., 2019; Sciuto et al., 2018). The ethical design of CAs for children is especially important, as this particular population presents unique challenges. Children users are in specific developmental stages, which necessitates thoughtful design and careful consideration of how CAs interact with them to safeguard their safety, privacy, and well-being (Charisi et al., 2022; Dignum et al., 2021).

This doctoral research is developed in the context of the collaborative doctoral partnership of the Joint Research Centre (JRC) of the European Commission, which provides scientific support to policy development in Europe. With the European Commission advocating for human-centred AI, regulatory frameworks such as the AI Act (Madiega, 2021) have been introduced to address these concerns and promote trustworthy AI systems that operate transparently, ethically, and responsibly. This thesis aligns with the European Commission's goals of fostering the development of AI that is both innovative and aligned with human rights, particularly when it comes to children.

Furthermore, this research is inherently interdisciplinary, drawing from various fields to address the complexities of designing trustworthy CAs for children (Figure 1.1). The project integrates technical knowledge from AI experts in natural language processing (NLP) (Chowdhary and Chowdhary, 2020) and CAs, societal insights from the field of child-computer interaction (Lehnert et al., 2022), and the regulatory knowledge emerging from the field of ethical AI (Siau and Wang, 2020). By combining these perspectives, the research aims to create a robust framework for designing and evaluating CAs that are both functional and ethically sound, ultimately contributing to safer, more enriching interactions for children.

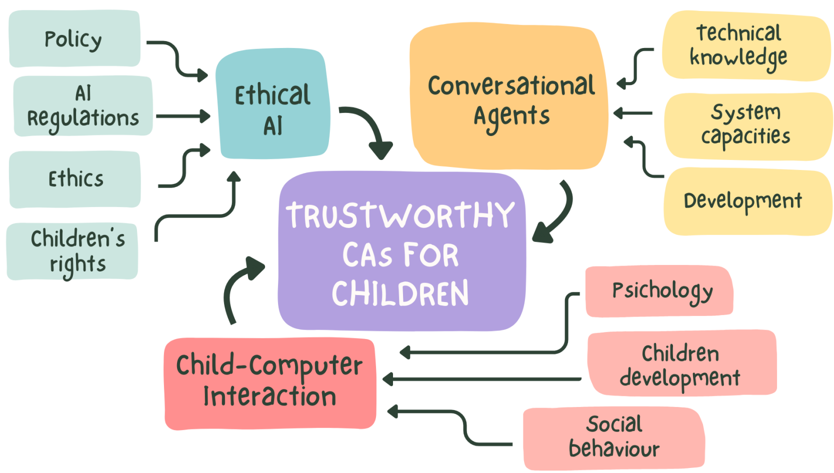


Figure 1.1: Interdisciplinarity of this thesis

1.2 Research goals

By focusing on the unique needs of children, this thesis aims to develop methodologies to enhance the trustworthiness of conversational agents that may interact with them. The focus will be on defining practical guidelines that can be applied to ensure that these technologies operate effectively and ethically in child-inclusive contexts. The research involves rigorous testing of CA interfaces to address the unique characteristics of young users. Through systematic experimentation and user studies, the thesis will identify key factors that influence the efficacy and safety of CAs in child-friendly applications. The research is driven by the following specific goals:

1. Advance the scientific understanding of conversational agents from a children perspective, and how children interact with them, by analysing the state of the art, opportunities, risks, and potential ways to mitigate those risks.
2. Adapt existing ethical guidelines for artificial intelligence systems for the specific case of conversational agents, ensuring that the unique needs and challenges posed by children are addressed.
3. Propose practical recommendations for the development of trustworthy conversational agents for children, translating the guidelines into actionable recommendations and addressing the technical and design challenges.
4. Evaluate children's perceptions and interactions with conversational agents in relevant settings, focusing on how different CA's behaviour affect children.

The anticipated impact of this research is multifaceted. It aims to enrich the field of child-computer interaction by offering novel insights into the design and implementation of child-friendly CAs. By emphasizing ethical considerations and trustworthiness, the findings are expected to support the development of relevant policies in this area. Additionally, the thesis adopts a European perspective, which, while not always explicitly stated, is reflected in the involvement of key stakeholders: children, parents, and policymakers.

1.3 Thesis structure

This thesis is organised into seven chapters, each building upon the last to enhance our understanding of CAs in the context of children and trustworthy AI.

The journey begins with the introductory chapter, setting the scene with motivation, research objectives, and the importance of creating trustworthy CAs for interacting with children. **Chapter 2** lays the scientific groundwork, discussing key concepts such as the nature of CAs, the principles of trustworthy AI, and child-computer interaction dynamics, highlighting the opportunities, challenges, and risks. **Chapter 3** presents a behavioural study with children, where we analyse their perception of an AI system in the context of a child-robot interaction study. This chapter provides some insights on the impact of the system design on children perception and the related ethical considerations. From the lessons learn in the literature review and this first study, **Chapter 4** addresses the development of ethical guidelines for trustworthy AI in the context of conversational agents and children population. These guidelines are later applied in **Chapter 5** to the practical development of a trustworthy conversational agent for children. The developed system is evaluated from a developer perspective in this chapter, followed by a user study with children in **Chapter 6**. **Chapter 7** includes the main contributions and conclusions of the thesis and states the perspectives for future work in this topic.

Each chapter of this thesis is a step in a comprehensive exploration of how CAs can be trustworthy and enriching for children, supporting the broader goals of human-centred AI. While a sequential reading of all seven chapters is recommended for a comprehensive understanding of the work, individual chapters may be focused on according to the reader's particular interest. However, for those already familiar with CAs and ethical AI, the preliminary concepts outlined in Chapter 2 can be briefly skimmed, as long as the concept of trustworthiness is understood not as user perception, but in terms of ethical standards.

Chapter 2

Scientific background

2.1 Introduction

In order to address the development of trustworthy conversational agents (CAs) that may be accessible for children, which is the main focus of this thesis, there is a need to combine scientific knowledge from different domains. This chapter delves into the multifaceted world of CAs, focusing on their interactions with children, and the importance of trustworthiness in their development.

The chapter is structured as follows. We begin with an overview of the basic components and functions of CAs, detailing the key modules that enable these systems to interact with users. Next, we explore the broad range of applications where CAs are currently being utilised. Following this, we focus on children's interactions with CAs by conducting a bibliometric analysis to identify trends and gaps in the research on child-CA interactions, and we present a literature review that discusses both the opportunities and risks associated with these interactions. Finally, we present ethical considerations surrounding the trustworthiness of AI systems, with a particular emphasis on guidelines set by the European Commission's High Level Expert Group on AI (HLEG) and UNICEF, as well as existing evaluation frameworks regarding trustworthiness.

The aims of this chapter are threefold: first, to provide a foundational understanding of CAs and their technological underpinnings; second, to highlight the unique considerations that arise when these systems interact with children, thereby contextualising our work; and third, to establish our definition of trustworthiness, and the importance of developing trustworthy AI systems that align with ethical guidelines and adequately address the needs of young users. Together, this content provides the essential background for understanding the subsequent research presented in this thesis.

2.2 Conversational agents

Conversational agents (CAs), known also as dialogue systems, virtual assistants, or chatbots, are computer programs that facilitate interaction with humans through conversation. A key characteristic of CAs is their reliance on language-based interaction, utilising speech, text, or multimodal inputs and outputs (McTear, 2020). They are typically composed of three essential components: an input module for processing user inputs, a dialogue manager to handle the conversation flow, and an output module to generate responses. Additionally, the dialogue manager may require access to external knowledge sources, such as databases or the internet, depending on the system's functionalities.

Tracing back to the 1966 ELIZA program, which mimicked a psychologist through rule-based interactions (Weizenbaum, 1966), the field of CAs has undergone significant evolution. Today, general-purpose models like those used in ChatGPT (OpenAI, 2022) are revolutionising the world with their ability to handle a broad range of tasks without specific functionalities.

2.2.1 Modules

The typical structure of a conversational agent comprises five main modules (Figure 2.1):

- **Automatic Speech Recognition (ASR):** This module transforms spoken language into written text, allowing the CA to process the user's verbal input. ASR systems have seen significant advancements in recent years, effectively handling diverse languages, accents, and noisy environments with high accuracy. Prominent examples include Whisper by OpenAI (Radford et al., 2023) and Google's Chirp (Y. Zhang et al., 2023), both of which leverage deep learning algorithms to achieve robust performance.

- **Natural Language Understanding (NLU):** This module interprets the text generated by the ASR, extracting the semantic meaning and relevant information from the user's input. NLU modules often perform computational tasks such as named entity recognition (which identifies and classifies key information such as names of people, or locations)(Mohit, 2014), intent classification (which determines the goal or purpose behind the user's input such as asking or commanding)(Schuurmans and Frasincar, 2019), and sentiment analysis (which assesses the emotional tone of the user's input such as positive or negative)(Wankhade et al., 2022). The introduction of Transformers (Vaswani, 2017), particularly the BERT model (Devlin et al., 2018), has significantly advanced the state-of-the-art in NLU by enabling models to focus on the most relevant parts of the text and to understand context.

- **Dialogue Management (DM):** This module manages the flow of conversation, determining the appropriate response based on the information processed by the NLU. Traditionally, DMs have relied on rule-based models (Smith et al., 2011; Z. Wang and Lemon, 2013) and finite-state systems (Fast et al., 2018; John et al., 2017), which follow predefined rules

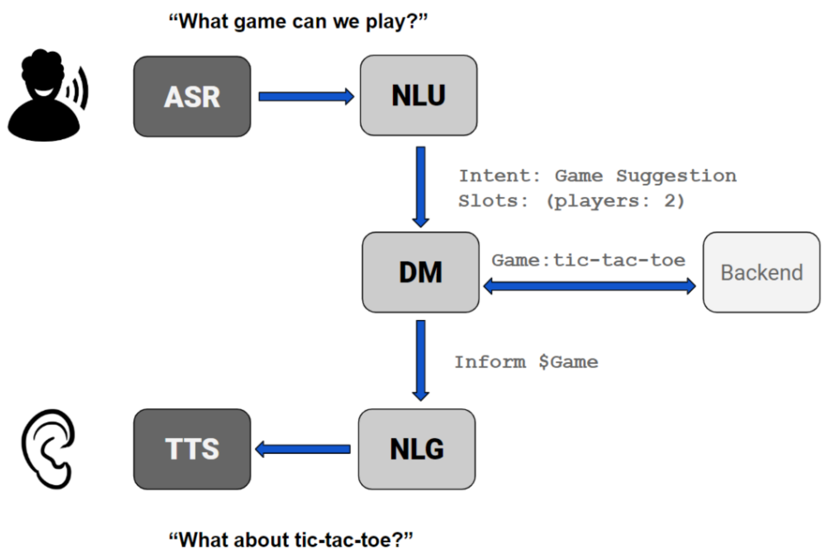


Figure 2.1: Modules typically included in a conversational agent: ASR (Automatic Speech Recognition), NLU (Natural Language Understanding), DM (Dialogue Management), NLG (Natural Language Generation) and TTS (Text-To-Speech).

or state transitions (Brabra et al., 2021). However, these approaches often struggle with flexibility and scalability. More advanced approaches include supervised learning (P.-H. Su et al., 2016; Wu et al., 2019), where the system is trained on a labelled dataset to predict the next action based on the current state, and reinforcement learning (Chen et al., 2019; Cuayáhuitl, 2017), which allows the system to learn dialogue policies by optimising long-term rewards through trial and error.

- **Natural Language Generation (NLG):** This module generates a response in natural language, based on the information obtained from the NLU, and the response type determined by the DM. Traditional NLG methods that filter output based on hand-crafted rules, such as template-based approaches and statistical models, have been effective but often lacked adaptability (Santhanam and Shaikh, 2019). Recent advancements in deep learning, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) (Rumelhart et al., 1986, Hochreiter and Schmidhuber, 1997), as well as Transformer models (Vaswani, 2017), have significantly enhanced NLG by enabling more coherent and contextually relevant responses.
- **Text-to-Speech (TTS):** This module converts the text generated by the NLG module into spoken language, allowing the CA to speak to the user naturally. Traditional methods include concatenative synthesis (using real speech recordings) and parametric synthesis (fully generated speech), often enhanced by Hidden Markov Models (HMM) (Ning et al., 2019). Modern TTS systems like WaveNet (Van Den Oord et al., 2016), Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2020), and NaturalSpeech (Tan et al., 2024), have achieved near-human quality in speech synthesis.

Some modern applications of deep learning combine the tasks of multiple traditional modules into a single, integrated approach. Large Language Models (LLMs) like GPT 3.5 (Brown et al., 2020), as well as PalM2 (Anil et al., 2023) and LLaMA 2 (Touvron et al., 2023), exemplify this trend by merging NLU, DM, and NLG into unified models capable of handling various natural language processing tasks. These models require high computational power and vast amounts of training data. Moreover, some LLMs have evolved to support multi-modal interactions, processing not just text but also vision and audio, as seen in GPT-4o (OpenAI, 2024a) or Gemini (G. T. Google, 2023). Notably, this end-to-end approach in GPT-4o has even demonstrated superior performance in specific areas, such as surpassing the specialised ASR capabilities of Whisper, another product from the same company (OpenAI, 2024a).

2.2.2 Applications

The versatility and expanding capabilities of CAs are evident in their adoption across diverse sectors.

- In **business operations**, CAs enhance customer-business interactions by simplifying tasks ranging from clothing selections and flight bookings in retail (Sousa et al., 2021; Wei et al., 2018) to maintaining accurate online business information. For instance, after the COVID pandemic, Google's Duplex technology (Leviathan and Matias, 2018) was used to phone businesses and verify and update their information on Google Maps, ensuring timely business data in such uncertain times (Schwartz, 2020 (accessed May 15, 2024)).

- In **healthcare**, CAs are improving patient's lives by facilitating medication management (Mathur et al., 2022; Z. Su et al., 2021), and mental health support (Sabour et al., 2023; Saha et al., 2022). For example, chatbots like Woebot assist users in managing conditions such as stress or burnout through regular, structured conversations (Durden et al., 2023). Additionally, the World Health Organisation utilised a chatbot to effectively reach wide populations during the COVID pandemic, disseminating crucial health information and increasing vaccination rates (Altay et al., 2023).

- In **automotive**, CAs have been enhancing the driving experience for years by facilitating interaction with vehicle control systems (Jaradat et al., 2022; Lin et al., 2018). Recent research focuses on improving the safety of self-driving cars, particularly in developing systems that can reliably hand back control to the driver in critical situations (Totakura et al., 2021; Wong et al., 2019).

- In **finance**, these agents automate routine tasks, enhancing service delivery and expanding operational capabilities. For example, Bank of America's Erica and Wells Fargo's predictive analytics both significantly improve customer service by offering personalised financial advice and anticipating customer needs. These innovations have boosted customer satisfaction and loyalty, contributing to increased sales and revenue for both banks (James et al., 2024).

- In the realm of **entertainment**, CAs create immersive experiences and dynamic storylines as seen with AI Dungeon (Latitude, 2019 (accessed August 24, 2024)), which uses models like GPT to generate engaging narratives based on user interactions. A video game company also explored the use of GPT-3 text generator combined with a powerful voice generator to create an AI character capable of answering spontaneous questions from users in a virtual world (Heaney, 2021), prior to the release of GPT-4o.

- **In education**, CAs support learning by engaging in interactive activities that make knowledge acquisition both engaging and accessible. While some of these tools support Massive Open Online Courses (MOOCs), others utilise the 'learning by teaching' model, where students enhance their understanding by instructing the CA, reinforcing their knowledge in the process (Chhibber and Law, 2019).
- In **home environments**, systems such as Google Assistant (Google, 2016), Siri (Apple, 2010), and Alexa (Amazon, 2014) have become integral to many people's daily routines, streamlining tasks ranging from controlling smart home devices to providing real-time cooking assistance and playing media on demand. Their intuitive interfaces make them accessible to users of all ages, highlighting the need for CAs to accommodate diverse user groups, and demographics, including children.

In conclusion, CAs have become integral to enhancing and simplifying everyday activities across diverse sectors, reflecting their broad utility and adaptability. Their capacity to reach a wide audience underlines their ability to exert a significant impact on society, making it essential to foster positive interactions with these technologies, especially in settings where children are present. Recognising the potential of these interactions allows us better understand and navigate the complexities they introduce. Section 2.3 explores the literature about how children engage with CAs, examining the opportunities and challenges this interaction presents in a digital world.

2.3 Conversational agents and children

Conversational agents (CAs) have permeated various aspects of modern life, becoming commonplace in homes, vehicles and public spaces. While these technologies are originally designed to cater to the general adult population, their accessibility through intuitive interfaces, such as voice commands, has inadvertently made them an everyday part of children's lives. Furthermore, children, with their innate curiosity and adaptability, explore these systems more creatively and intensively than adults, often exploring the limits for which these technologies were initially designed (Garg and Sengupta, 2020; Lovato et al., 2019; Sciuto et al., 2018).

2.3.1 Bibliometric analysis

To gain a deeper understanding of how children's interactions with CAs have been explored in the research community, we conducted a bibliometric analysis. Although methodologies such as PRISMA (Snyder, 2019) are designed for in-depth systematic reviews, the exploratory nature of our consultation led us to proceed by using the *Bibliometrix* tool (Aria and Cuccurullo, 2017). This tool enabled us to analyse the most frequent keywords, identify clusters, and examine co-occurrences of terms within a corpus of research papers. We assembled this corpus from both Web of Science (Clarivate, 1997) and Scopus (Elsevier, 2004) databases using the below search query targeting papers' titles, abstracts, and keywords. The keywords were identified through an informal review of several relevant articles on the topic and deemed suitable for conducting a general exploration of the state of the literature.

(("child" OR "children") AND ("conversational agent" OR "conversational AI" OR "dialogue system" OR "dialogue systems" OR "chatbot" OR "chatbots" OR "virtual assistant" OR "home assistant" OR "voice assistant"))

After removing duplicate articles, we performed an initial screening by reviewing the titles with broad inclusion criteria to ensure a comprehensive overview of the literature. This approach allowed us to analyse a total of 440 papers published between 2000 and 2022. Notably, 54% of these papers were published between 2020 and 2022, indicating a significant recent surge in interest particularly remarkable since 2015 (from 7 papers per year in 2015 to 83 in 2022). This uptick can likely be attributed to the popularity, accessibility and affordability of CAs like Amazon's Alexa and Apple's Siri, along with the emergence of LLMs such as ChatGPT (van Dis et al., 2023).

An interesting expansion occurs when incorporating the terms *"social robot" OR "robot interaction"* into the search query, increasing the paper count from 440 to 2580. This suggests a substantial research focus on the effects of embodiment and non-verbal interaction facets, such as gestures, gaze, and facial expressions. On the verbal communication front, many human-robot interaction studies with children rely on *Wizard-of-Oz* setups (Nasir et al., 2022; Zou et al., 2022), probably due to the challenges that these systems face when interacting with children (exposed in Section 2.3.2). This trend might soon shift towards full automation of voice interactions due to recent advances in ASR and LLMs. However, the maturity, moral capabilities, and trustworthiness of these systems in child interactions remain yet to be explored (Ganguli et al., 2023).

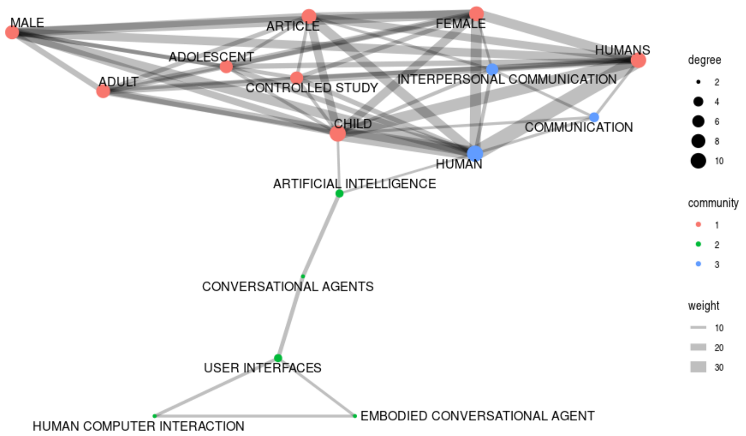


Figure 2.2: Results from our bibliometric analysis of 440 papers focusing on CAs and children: most frequent terms, cluster and co-occurrence network. The thicker the link, the more weight the co-occurrence of words has. The size of the nodes indicates the frequency of the keyword (the larger the radius, the greater the use) and their colour (red, blue or green) denotes which cluster they belong to.

Figure 2.2 illustrates the results from our keyword frequency, co-occurrence, and clustering analysis of the 440 papers related to children and CAs. We identified three primary clusters of terms (nodes). The most prominent cluster, depicted with red nodes, centres around "controlled study" and is strongly linked with "humans" and demographic descriptors like "male", "adult", "adolescent", "female", and "child". This cluster primarily encompasses literature from controlled lab settings, often involving adults who participate as guardians. The second cluster, represented in blue and closely overlapping the first, includes pivotal keywords such as "communication" and "interpersonal communication", highlighting research focused on the communicative aspects of child-CA interactions within controlled environments. The third cluster, distinguished by green nodes and clearly separated from the others, comprises more technical terms like "artificial intelligence", "user interfaces", and "(embodied) conversational agents". The separation between behavioural and technical research indicates a potential research gap that could benefit from enhanced multidisciplinary collaboration among social scientists, AI researchers, and Human-Machine Interaction experts. This is of particular importance in this new era where large language models will pave the way towards fully automated (and less "controlled") human-CA interaction.

2.3.2 Literature review

The findings from the bibliometric analysis highlight a growing interest in examining how CAs impact young users. The presence of these technologies in child-accessible environments calls for a careful examination of their influence. Understanding the dual nature of this impact — where opportunities coexist with challenges, and risks — is essential for developing CAs that safeguard young users while benefiting from their interaction with digital technology. To explore this further, we review the literature on children's interactions with CAs, highlighting both the potential benefits and the critical concerns.

Opportunities

Children's interaction with CAs opens up a range of opportunities that can enhance their developmental and learning experiences. These opportunities are not only transformative in terms of educational support, but also extend to personal growth and social interaction.

- **Education:** CAs engage children in a variety of interactive learning activities. These range from simple information searches (Landoni et al., 2020; Lee et al., 2023) to language acquisition (Nasihati Gilani et al., 2018), reading activities (Liu et al., 2022; Y. Xu and Warschauer, 2020b; Y. Xu et al., 2022) and school material learning (Law et al., 2020; W. Xu et al., 2023; Y. Xu and Warschauer, 2020a). Some families have even benefited from incorporating CAs into their parenting strategies (Beneteau et al., 2020). These systems sustain interest and improve comprehension skills, making learning both engaging and effective.
- **Accessibility:** CAs facilitate communication through voice-activated commands, significantly improving accessibility for all children. This feature is especially beneficial for those with physical, visual, or learning disabilities, as well as very young children who have not yet developed reading and writing skills, enabling them to interact effectively with digital content (Catania et al., 2021; Pradhan et al., 2018).
- **Development:** CAs have been employed to foster social skills and emotional intelligence among children. For instance, they help children develop social behaviours through persuasive game strategies (Fraser et al.,

2018) and support autistic children in enhancing their social skills (Ali et al., 2020; L. Zhang et al., 2020).

- **Safety:** CAs offer support by teaching children about health practices or by providing reminders for medication (Sezgin et al., 2020). They also serve as a friendly interface for teenagers to reduce and control depression and anxiety (Fitzpatrick et al., 2017).

An application of CAs that we consider in this thesis (Chapter 5 and 6), is collaborative storytelling, which leverages children's natural affinity for narratives and engages them in creative story-making. Early collaborative storytelling studies using teleoperated systems like Wizard of Oz setups (M. Sun et al., 2017) and automated CAs have highlighted challenges in accurately interpreting children's voices (D. T. Ong et al., 2018; E. Ong et al., 2019). However, recent advances in natural language processing and speech recognition have reduced these barriers, facilitating further research on CAs as tools to support children's development and well-being. For instance, Lee et al. enhanced children's agency by allowing them choose story paths (Lee et al., 2022), and Elgarf et al. demonstrated how storytelling with CAs can improve children's imaginative skills (Elgarf, 2022; Elgarf et al., 2022), showcasing the potential benefits of CAs for children's growth. Moreover, collaborative storytelling CAs are versatile engagement tools across all age groups (Nichols et al., 2021; Papadopoulos et al., 2013; Pérez et al., 2022), including applications that bring children and parents together for shared activities (Lee et al., 2022; Z. Zhang et al., 2022), strengthening family bonds and enhancing overall well-being.

Challenges and Risks

While CAs offer numerous developmental benefits for children, the complexity of the digital world, combined with the development of children's understanding and behaviour, exposes them to greater risks. These issues require meticulous consideration to ensure that interactions between children and CAs are safe and beneficial. Regarding children particularities, that differ from the general population, some challenges have been identified in the literature affecting the different modules of CAs (Kennedy et al., 2017; Narayanan and Potamianos, 2002; Nilsen, 2019; Sobti et al., 2024). These challenges are detailed in Table 2.1.

Table 2.1: Problems of CAs traditional modules with children.

Module	Children-specific characteristics
ASR	Speech acoustic characteristics, e.g. high pitch range, particular prosody.
NLU	Expressions, vocabulary and grammar.
DM	Information needs, protection, allowed functionalities according to age.
NLG	Need of simpler words or explanations according to age.

The challenges of child-CA interaction have stimulated extensive research aimed at mitigating potential risks:

- **Exclusion:** Differences in children's speech and understanding can lead to exclusion, especially for those with below-average speech skills (Monarca et al., 2020). Research efforts to enhance CA performance for children include developing speech identification for infants (Lavechin et al., 2020) and identifying strategies for when a CA fails to understand a child's input (Røyneland, 2019).

- **Over-trust:** Children often perceive CAs as friends or companions (Druga et al., 2017; Kahn Jr et al., 2012), and often attribute alive properties to these devices (Girouard-Hallam et al., 2021; Y. Xu and Warschauer, 2020c). Transparent information, has demonstrated to reduce the alive perception of the system (Straten et al., 2020).

- **Data privacy:** Children often do not realise that their data can be collected nor understand the potential risks associated with this (G. Wang et al., 2022). Additionally, research indicates that parents do not always deem it necessary to constantly supervise their children's interactions with these technologies (Turner et al., 2022).

- **Gender bias:** Many chatbots are explicitly or implicitly designed with gender-specific cues, predominantly as female (Tolmeijer et al., 2021). In addition, some LLMs as GPT-3 revealed significant biases based on gender, race and religion (Brown et al., 2020). This can transmit biased gender perceptions to children, which has an impact on their personal development and their view of society.

- **Inappropriate activations:** Home assistants can enable children to trigger unwanted actions, such as purchases. These devices also expose children to advertisements and purchasing options without adult supervision, increasing risks such as unwanted transactions (Haas and Keller, 2021). Even though some families manage this risk within their dynamics (Kudina and Coeckelbergh, 2021), the lack of age-appropriate controls underscores the need for better safeguards to protect children from inappropriate activations.

- **Unanticipated risks:** The innovative nature of CAs can lead to unforeseen issues, such as the bullying experienced by girls named "Alexa", a consequence of Amazon's choice of wake word for their devices (Johns, Tim, 2021).

Given these risks, establishing guidelines for the development of trustworthy CAs is crucial. These guidelines must leverage the benefits while carefully evaluating and minimising potential harms, ensuring that CAs serve as supportive tools for children's development without compromising their well-being.

2.4 Trustworthy AI

The analysis of social and ethical implications of algorithmic systems is not new (Friedman and Nissenbaum, 1996). However, as the relevance of AI systems continues to escalate, the ethical implications of how these technologies can affect humans are coming sharply into focus. This has led to the emergence of multidisciplinary efforts to address ethical issues such as privacy, transparency or fairness. Initiatives like the FAccT conference (ACM, 2018) and various global frameworks (*Australian AI Ethics Framework*, 2019; HLEG, 2019; UNESCO, 2021) emphasise the importance of developing AI responsibly, balancing innovation with ethical considerations. In this section, we examine some initiatives and their strategies for promoting trustworthy AI, particularly focusing on the unique considerations for children as users of CA systems. We will also discuss the methods for evaluating the trustworthiness of such systems.

2.4.1 Guidelines

HLEG

The High Level Expert Group (HLEG) on AI, established by the European Commission in 2018, developed the Ethical Guidelines for Trustworthy AI (HLEG, 2019) with the aim of protecting people's fundamental rights across different context of AI applications. These guidelines introduce seven key requirements essential for the trustworthiness of AI systems:

1. **Human agency and oversight.** AI systems should support human agency and human decision-making in order to respect for human autonomy. These systems need to help in creating a society that's equal, successful, and supports everyone's choices. In addition AI systems should always protect basic human rights, with people overseeing and guiding them.

2. **Technical robustness and safety.** AI systems need to be dependable (deliver services that can justifiably be trusted) and resilient (robustness when facing changes). To be technically robust, these systems should be built to prevent risks, work reliably as planned while minimize harm, both by not causing it accidentally and by stopping it if they can. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may challenge or oppose them.

3. **Privacy and data governance.** Closely linked to avoiding harm is the idea of privacy, which is a basic right that AI systems can greatly impact. To protect privacy, it's important to manage data well. This means making sure the data is good and accurate, fits well with where the AI will be used, has clear rules for who can use it, and is handled in a way that keeps people's privacy safe.

4. **Transparency** Making AI trustworthy is really important, and being clear and open about how it works is a big part of that. This includes three key things: being able to follow and understand the AI's decisions (traceability); making sure people can understand why the AI makes certain choices (explainability); and being honest about what the AI can and cannot do (open communication about its limitations).

5. **Diversity, non-discrimination and fairness** To make AI trustworthy, it is important to focus on inclusion and diversity throughout the entire

AI system's life cycle. AI can sometimes have built-in biases from the past, missing information, or suffer from bad governance models. These issues could unfairly affect certain people or groups, increasing bias or exclusion. Harm can also come from intentionally using consumer biases unfairly, like fixing prices in non-transparent ways. We should remove bias early on and design AI to be easy for everyone to use, no matter their age, gender, abilities, or other traits. Making sure it is accessible to people with disabilities, who are part of all groups in society, is especially important.

6. **Societal and environmental well-being.** To ensure fairness and prevent harm, AI must account for societal, environmental and other being's impacts throughout its life cycle. Its widespread use in sectors like work, education and entertainment can influence social dynamics and health. AI should be monitored to prevent adverse effects. It is crucial for AI to support ecological initiatives, like the Sustainable Development Goals (Nations, 2015), and to benefit humanity, including future generations. AI must uphold democratic values and diversity, without compromising the welfare of society.

7. **Accountability.** The principle of accountability requires mechanisms to ensure responsibility for the creation, implementation, and use of AI. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. If AI causes unfair or harmful effects, there should be easy ways to hold those responsible accountable, and ensure those affected have a fair chance for correction and remediation.

Importantly, the requirements outlined in these guidelines are considered foundational in Europe and have played a significant role in shaping the AI Act (Madiega, 2021), the main European Union regulatory initiative aimed at regulating AI applications to ensure they adhere to ethical standards and prioritise human protection. For this reason, these specific requirements have been adopted as the basis for our work.

UNICEF

In response to the growing impact of AI systems on society, UNICEF has crafted a specialised policy guidance aimed at safeguarding children's rights (Dignum et al., 2021). This initiative informs policymakers and businesses about the importance of considering children's unique needs and rights in the context of AI systems. The guidance specifies nine critical requirements centred on the use of AI systems by children:

1. **Support children's development and well-being:**
   Let AI help me develop my full potential.

2. **Ensure inclusion of and for children:**
   Include me and those around me.

3. **Prioritise fairness and non-discrimination for children:**
   AI must be for all children.

4. **Protect children's data and privacy:**
   Ensure my privacy in an AI world.

5. **Ensure safety for children:**
   I need to be safe in the AI world.

6. **Provide transparency, explainability, and accountability for children:**
   I need to know how AI impacts me. You need to be accountable for that.

7. **Empower governments and businesses with knowledge of AI and children's rights:**
   You must know what my rights are and uphold them.

8. **Prepare children for present and future AI developments:**
   If I am we prepared now, I can contribute to responsible AI for the future.

9. **Create an enabling environment:**
   Make it possible for all to contribute to child centred AI.

While these guidelines are primarily focused on policy, they bring forward crucial child-centred considerations that are essential for the ethical development of AI systems. By emphasising the unique needs and rights of children, these guidelines provide valuable insights that help ensure AI technologies are designed and deployed in ways that ensure the well-being and protection of younger users. For this reason, we have incorporated these child-focused considerations into our research to better align AI development practices with the specific needs of children.

Alignment

To better understand the alignment of these two important guidelines, we conducted a qualitative mapping in our work presented at Ethicomp (Escobar-Planas, Gómez, and Martínez-Hinarejos, 2022). We performed a qualitative mapping between HLEG and UNICEF guidelines, focusing on practical advise for the development of trustworthy CAs.

For this mapping, we analysed each requirement from UNICEF, extracting key concepts and mapping them to the HLEG's requirements. For example, by reading in detail UNICEF's 9th requirement (*"Create an enabling environment"*), we extracted the concept "Include discriminate children on the oversight", which we closely related to "Human Agency and Oversight" from the HLEG's requirements. We also extracted "Be aware of digital divide when developing AI systems, providing access to the majority of people", linked to "Diversity, Non-discrimination and Fairness" and "Open access so other can benefit form your discoverings." which is somewhat linked to "Societal and Environmental Well-being". Additionally, we identified "Government and corporate leaders funding for child-centric ethical AI", which we did not consider a recommendation for developers. More details about the procedure to obtain this matrix are shared at https://github.com/mescpla/CAs4Children-ETHICOMP22.git.

From our results, we observe that HLEG's guidelines have a strong focus on the development of AI devices, while some of the UNICEF's requirements emphasise policy considerations. Nevertheless, as observed in Table 2.2, most requirements from UNICEF align with at least one major, and sometimes additional, requirements of the HLEG. The only exception is the educational aspect of UNICEF's 8th requirement "Prepare children for present and future developments in AI", which focuses on policy and has only minimal overlap with the HLEG's 6th requirement "Societal and Environmental Well-being", which broadly considers work and skills.

Based on this analysis, we consider the HLEG's guidelines as the primary framework for the remainder of our study, with a focus on CAs. These guidelines are supplemented by the UNICEF guidelines on AI for children, which serve as a complementary framework that addresses children's rights, including educational aspects.

Table 2.2: Mapping between UNICEF AI for children requirements and HLEG trustworthy AI requirements: human agency and oversight (H), technical robustness and safety (R), privacy and data governance (P), transparency (T), diversity, non-discrimination and fairness (D), societal and environmental well-being (W), and accountability (A). (*), (**) and (***) indicate low, mid or high correspondence between related requirements.

	HLEG trustworthy AI						
UNICEF AI for children	**H**	**R**	**P**	**T**	**D**	**W**	**A**
Support development and well-being		**	**		**	***	
Ensure inclusion					***		
Prioritize fairness and non-discrimination					***		
Protect data and privacy	**		***				
Ensure safety	**	***				**	
Provide transp., explainab. and account.	***	**		***	**		***
Empower governments with knowledge				*		***	
Prepare children for developments in AI	*			*		*	
Create an enabling environment	**				**	**	

2.4.2 *Trustworthy CAs*

Shifting from the broad considerations of AI to the specific domain of CAs, the concept of trustworthiness often focuses on user perceptions, such as the impact of the agent's embodiment on trust (Lupetti et al., 2023; Robb et al., 2023). However, there is an important gap in research regarding trustworthy design of CAs, particularly in terms of fairness, robustness, and transparency defined in Section 2.4.1 (Diederich et al., 2022). The importance of this gap was underscored by the release of GPT-3 (Brown et al., 2020), which revealed substantial biases based on gender, race, and religion, highlighting the need for further research on the ethical design of CAs. The need for this research is even more urgent in contexts involving children, where dedicated ethical child-CA interaction research remains lacking (Chubb et al., 2022; Seymour et al., 2023).

One of the notable efforts to bridge this gap is the report by the European Commission's Joint Research Centre, which proposes an integrated research and policy agenda to enhance the ethical use of AI technologies while safeguarding children's rights (Charisi et al., 2022). It emphasises the development of conversational agents and other AI applications, aiming to align existing research with robust policy frameworks that protect these young users.

Additionally, the application of UNICEF's policy guidance on AI for children in the development of a CA embedded in a robot, demonstrates the application of these guidelines to real-world scenarios (UNICEF et al., 2021). While these guidelines are not tailored specifically for CAs, their adaptation to include considerations for children's rights demonstrates its broader application in the development of CAs. However, the lack of specific guidelines for CAs that interact with children underscores the need for continued exploration and development in this area.

In summary, while the ethical principles established by organisations such as the HLEG and UNICEF provide a solid foundation for guiding ethical AI, there remains a significant need for comprehensive research focused on developing trustworthy CAs that are accessible to children. This thesis aims to contribute to this effort by promoting the trustworthy design of CAs, aligning with broader societal values and specifically addressing the needs of children.

However, establishing ethical guidelines is only the first step. It is equally important to ensure that these guidelines are effectively implemented through evaluation methods. In Section 2.4.3, we will explore existing evaluation frameworks for CAs and discuss how could we measure trustworthiness on these systems.

2.4.3 Evaluation

Evaluation is essential to measure the applicability and effectiveness of guidelines. However, most evaluations practices in CAs have traditionally focused on metrics related to efficiency and effectiveness. Reviews in the field (Deriu et al., 2021; Dybkjaer et al., 2004; Yeh et al., 2021) primarily reference frameworks such as PARADISE (Walker et al., 1997), which evaluates both task success and dialogue quality by considering metrics like task completion rates and user satisfaction; EAGLES (Brey et al., 2000), which emphasises standardisation in linguistic accuracy and consistency across different systems to ensure comparability of results; and metrics like FED (Mehri and Eskenazi, 2020), an innovative unsupervised evaluation method that leverages pre-trained language models to assess dialogue quality through metrics like coherence, relevance, and informativeness. All these measures emphasise user satisfaction, overlooking important aspects related to trustworthiness. Even while some studies have touched upon ethical considerations for CAs, they often focus narrowly on the system's human-like perception rather than broader trustworthiness issues (Radziwill and Benton, 2017).

Furthermore, this gap in evaluating trustworthiness is not unique to CAs but extends to all AI-based systems. To complement existing research on trustworthy AI presented in Section 2.4.1), the HLEG introduced the Assessment List

Table 2.3: Sample questions from ALTAI, and their corresponding trustworthy requirement.

Question	**Requirement**
Does the AI system simulate social interaction with or between end-users or subjects?	Human agency and oversight
How exposed is the AI system to cyber-attacks?	Technical robustness and safety
Was your AI system developed, by using or processing personal data?	Privacy and data governance
Did you explain the decision(s) of the AI system to the users?	Transparency
Does the AI system corresponds to the variety of preferences and abilities in society?	Diversity, non-discrimination and fairness
Are there potential negative impacts of the AI system on the environment?	Societal and environmental well-being
Did you ensure that the AI system can be audited by independent third parties?	Accountability

for Trustworthy AI (ALTAI) (HLEG, 2020) as a tool to facilitate the evaluation of trustworthiness.

ALTAI

Defined after a piloting process with 350 stakeholders, the Assessment List for Trustworthy AI (ALTAI) (HLEG, 2020) serves as a practical self-assessment tool designed to help developers and organisations evaluate the trustworthiness of their AI systems. ALTAI encompasses 69 self-evaluation questions aligned with the HLEG's requirements presented in Section 2.4.1, as exemplified in Table 2.3. These questions demand a deep understanding of the system's characteristics and are therefore intended for designers and developers to answer.

ALTAI has also been referenced in various research studies, demonstrating its emerging use as a framework for evaluating AI systems' trustworthiness in different applications (Rajamäki et al., 2023; Slosiarová et al., 2023; Stahl and Leach, 2023; Zicari et al., 2021). While this evaluation framework is not specifically tailored for CAs or children, its comprehensive approach to trustworthiness and broad applicability make it a valuable tool for this work.

2.5 Conclusions

This chapter has provided a comprehensive overview of CAs, with a particular focus on their interactions with children, as well as the importance of trustworthiness in these systems. We began by examining the foundational components of CAs, detailing their structure and highlighting a period of rapid advancement that is making significant strides toward more natural communication capabilities. We also explored their expanding role across various sectors, simplifying everyday activities with broad utility and adaptability. Children, in particular, are increasingly exposed to CAs and tend to explore the limits of these technologies in creative and often unanticipated ways.

Through a bibliometric analysis, we identified growing research interest in the interaction of CAs with children, with a significant predominance of embodied CAs. This research faced limitations related to the autonomy and speech recognition of children. However, recent advancements in LLMs and ASR technologies are likely to drive a surge in both embodied and non-embodied CA studies. The bibliometric analysis also revealed a separation between behavioural and technical research, which could be addressed through greater multidisciplinary collaboration.

We then followed with a literature review, providing a deeper understanding of the significant benefits CAs can bring to children, including educational support, social development, and accessibility enhancements. Collaborative storytelling emerged as a standout application, showcasing how CAs can engage children's imaginations and foster creative expression. However, CAs also pose challenges and risks, such as issues related to data privacy, over-trust, and gender bias, particularly when interacting with children.

In terms of ethical AI, existing guidelines from the HLEG and UNICEF provide a strong foundation for developing trustworthy AI systems and are well-aligned in many respects. However, there is a notable gap in specific guidance for CAs that consider ethical design, especially those designed for children. Moreover, there is currently no specific tool designed to evaluate the trustworthiness of CAs, nor is there a tool that considers the unique needs of children. While ALTAI is not specifically designed for CAs or child applications, it offers a comprehensive framework that can be adapted to evaluate trustworthiness in AI systems broadly.

In conclusion, while discussing the landscape of trustworthy AI, it became evident that existing guidelines provide a useful starting point but lack specifications tailored to the nuanced interactions between CAs and children. This thesis aims to bridge this knowledge gap by proposing new approaches to the design and deployment of CAs, ensuring they are both beneficial and suitable for child users.

Chapter 3 transitions from theoretical considerations to an empirical exploration through a user study examining children's interactions with CAs. This study aims to illuminate how children engage with these technologies, providing insights that are critical for refining CAs to better support their needs and experiences.

Chapter 3

Children's interactions

3.1 Introduction

Chapter 2 laid the groundwork by exploring the foundational concepts of Conversational Agents (CAs) and their use by children, highlighting both the opportunities and challenges these technologies present. It emphasised the necessity for ethically designed systems that cater to young users and identified trends and gaps in the existing literature on CA-child interaction. With this understanding in place, we now turn our attention to a more detailed examination of these interactions in a controlled environment.

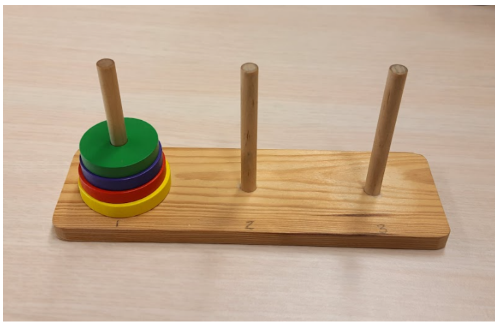
We present a study involving teams of two children interacting with a tabletop social robot —–an embodied CA—– while solving the Tower of Hanoi task. This study aims to explore how the robot's speech and behaviour affect the children's responses and their perceptions of the CA. Our decision to use a social robot was informed by our previous bibliometric analysis (Section 2.3.1), which identified it as a popular area of research in child-CA interaction studies. Building on this insight, we chose the social robot as a means to explore emerging ethical issues relevant to CAs, situating this research within the specific realm of social robots (as embodied CAs). However, while the physical presence of the robot enhances interaction, our research primarily aims to understand how verbal and behavioural elements can be optimised to support children's developmental needs, minimising risks while maximising opportunities. Our investigation provides a valuable framework for exploring the impact of a CA's speech and behaviour on children's responses and perceptions.

The chapter is structured as follows. First, we provide a detailed explanation of the methodology, including the design of the robot and the experimental setup. Next, we present a comprehensive analysis of the data collected, combining quantitative measures from the task with qualitative insights from semi-structured interviews conducted after the interaction. This dual approach enables us to capture both the observable behaviours and the underlying perceptions of the children.

This approach aligns with our broader aim to enhance trustworthiness of CAs for children. By applying established methodologies of child-robot interactions and examining the specific influences of conversational behaviour, our study contributes foundational insights into the effective and ethically aligned design of CAs for children.

3.2 Methodology

3.2.1 Robot and behaviour design



(a) Tower of Hanoi. The logic game played by the children with the robot. The game involves moving all disks from one peg to another, maintaining ascending order by size and moving one disk at a time.

(b) Haru robot, as presented at the 2018 ACM/IEEE International Conference on Human-Robot Interaction (Gomez et al., 2018)

Figure 3.1: (a) Tower of Hanoi game and (b) Haru robot used in our experiment.

We employed the tabletop robot Haru (Figure 3.1b), in an experimental setting to play an interactive game with children. Haru is equipped with a behaviour tree that controls its interactions, including announcing the start of the game, encouraging teamwork (if any child is far from the robot), offering assistance (if any child does not move in a long period of time), suggesting movements during gameplay (during the robot's turn or if children ask for help), and congratulating the children upon completion. The robot's responses are triggered by automatic tracking of children's skeletons and the disks' positions during the game, supplemented by a joystick operated by the experimenter to monitor game turns and requests for help. This setup facilitates both verbal and non-verbal communication between the children and the robot.

The experimental study is designed as a 2x2 factorial experiment, manipulating two key variables of the conversational agent, cognitive reliability and expressivity, which are integrated into the Dialogue Management (DM), Natural Language Generation (NLG), and Text-to-Speech (TTS) modules in the following manner:

- **Cognitive reliability.** This factor is focused on the way the DM performs in the Hanoi Tower game (Figure 3.1a). We designed an optimal robot, that always suggests the optimal movement to solve the task, and a sub-optimal robot, that always suggest a non-optimal movement to solve the task. Note that the robot does not perform any movement itself, it just provides suggestions leaving children the final decision about what movement to perform.

- **Expressivity.** This factor is focused on the social signals of the robot. The NLG was configured to create two different behaviours: an expressive version of the system, which employed more emotive and engaging phrasings (like *"What do you think super-team? Do you feel like playing again?"*), and a neutral version (that used phrasings like *"Would you like to repeat the game?"*). In addition, these verbal expressions were coupled with different configurations of the TTS, as outlined in Table 3.1, to control the expressivity of the robot's sentences. While in the neutral condition the CA used all the time the *n*eutral configuration, the expressive condition changed between *original, calm* and *happy* configurations. We selected two different features (pitch and speed), and empirically set those parameters during the design phase.

Combining the two factors we formulated the four experimental conditions: optimal expressive (OE), sub-optimal expressive (SE), optimal neutral (ON) and sub-optimal neutral (SN).

Table 3.1: TTS configurations for expressive and neutral conditions.

Expressions	**Mood**	**Speed**	**Pitch**
Neutral	Serious	80%	-3
Expressive	Original	100%	0
	Calm	85%	0
	Happy	120%	+7

Table 3.2: Distribution of participants by condition, sex, and age.

Condition	Female	Male	Younger	Older	Children	Teams
Optimal Expressive (OE)	8	12	8	12	20	10
Sub-optimal Expressive (SE)	11	10	9	12	21	11
Optimal Neutral (ON)	9	11	6	14	20	10
Sub-optimal Neutral (SN)	7	13	8	12	20	10
Total	35	46	31	50	81	41

3.2.2 Participants

For the experimental study, we recruited 84 children from a local school in Spain. To ensure a degree of homogeneity in the study groups, the children were paired based on biological sex (Female and Male) and age group (Younger, 5-6 years old, and Older, 7-8 years old). Each pair was randomly assigned to one of the experimental conditions (OE, SE, ON, SN), as shown in Table 3.2. Due to technical issues during the data collection process, the analysis was ultimately conducted on 41 teams (encompassing 81 children).

The study received approval from the Ethics Board Committee of the Joint Research Centre of the European Commission. Informed consent was obtained from all participants' parents or legal guardians, and assent was obtained from the children at the beginning of the experiment, ensuring their voluntary participation in the study.

3.2.3 Interaction and experimental design

This study consisted of three distinct sessions: a preliminary session, a robot intervention session, and an interview session, as shown in Figure 3.2. All sessions were video-recorded.

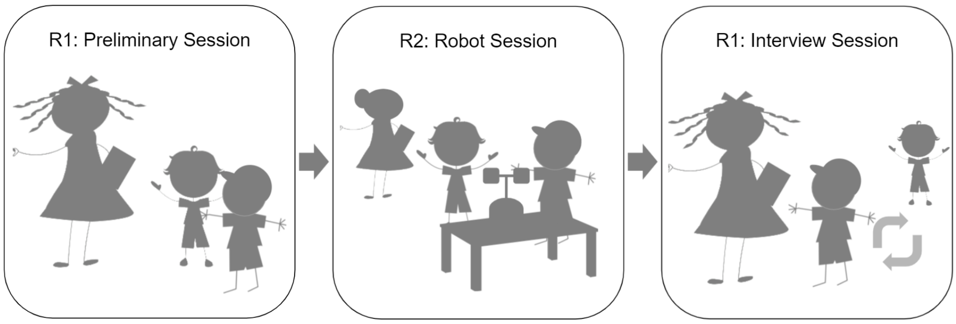


Figure 3.2: Experimental sessions were conducted in rooms R1 and R2. Children commenced the Preliminary Session in Room R1 with the interviewer, transitioned to Room R2 for the Robot Session with Haru, and concluded with the Interview Session back in Room R1 where each child was interviewed individually.

Preliminary Session (Approximately 25 minutes).

In the preliminary session, the interviewer briefed the children about the experiment's structure and informed them that all collected data would be anonymised for research purposes only. Children's assent was obtained to ensure their willing participation. Additionally, the experimenter explained the rules of the Tower of Hanoi game.

Robot Intervention Session (Approximately 30 minutes).

Before the children entered the room, the robot, Haru, was activated to ensure readiness for interaction. This session was autonomously conducted by Haru but closely monitored by developers from a hidden location. Upon entry, the experimenter introduced Haru, explaining its purpose and capabilities in learning various languages and games, which was important for setting the context of the interaction. The session followed with the pre-intervention manipulation check, the collaborative problem-solving task, and the post-intervention manipulation check, as follows:

- Pre-intervention Manipulation Check. This check was designed to assess the clarity of Haru's expressive behaviours. To accommodate the young age of the participants, the experimenter used child-friendly language, employing adjectives such as "serious", "vivid", or "sympathetic", and supported these descriptions with body-language demonstrations to ensure that children could grasp the nuances between different expressions. The children were then exposed to a series of Haru's interactions, alternating between expressive and neutral behaviours (e.g., *"<child's name>, I really like your name!"* versus *"Greetings, <child's name>"*). After experiencing these behaviours, the children completed a questionnaire to identify which behaviours they perceived as more expressive. This process was not only intended to validate Haru's expressiveness but also to familiarise the children with the robot, thereby mitigating the novelty effect. It is important to note that Haru's cognitive reliability, being contingent on task-specific interactions, was assessed only during the post-intervention manipulation check.

- Problem-solving Task. The main activity involved solving the Tower of Hanoi, a structured task designed to facilitate meaningful interaction with Haru. The task was divided into three phases to adapt to the escalating complexity of the game and varied social dynamics among the participants. The *Baseline Phase* had Haru observing without intervening, allowing an assessment of the children's initial problem-solving strategies. During the *Intervention Phase*, Haru actively engaged by suggesting possible moves, directly influencing the problem-solving process. In the final *Evaluation Phase*, Haru's involvement was minimised, allowing the children to apply their learned strategies independently, facilitating children's autonomous problem-solving.

- Post-intervention Manipulation Check. Following their session with Haru, the children completed a questionnaire designed to assess their perceptions of the robot's reliability and expressivity. The questionnaire used child-friendly language to ensure that the children could easily understand and accurately respond to the questions.

Table 3.3: Interview questions categorized by topic: (I) familiarization, (II) focus on the robot, and (III) social capabilities of the robot.

Topic	Question
I	1) *Did you enjoy it?* 2) *Can you tell me something about the robot?*
II	3) *How would you describe the robot to your friends?* 4) *Could the robot be your friend? Why?* 5) *Do you think that the robot would like to be your friend? Why?* 6) *How would you describe the robot at home?* 7) *Could the robot see you? Why?* 8) *Could the robot hear you? Why?*
III	9) *Did the robot help you? Why?* 10) *Did you help the robot? Why?* 11) *Would you like to have a robot? Why?* 12) *Do you think that the robot would like to have a lot of friends? Why?* 13) *Could the robot be able to harm anybody? Why?* 14) *Is the robot able to help you with your homework?* 15) *Does the robot know everything?*

Interview Session (Approximately 25 minutes)

The purpose of this session was to gather in-depth insights into the children's perceptions of their interaction with the robot. Each child was interviewed individually to ensure responses were not influenced by their peers, but the collective data was also analysed to understand broader patterns. The interview was structured into two distinct tasks:

- Semi-structured interview: The semi-structured interview was organised around a predefined set of questions (Table 3.3) that were previously validated and used (Charisi et al., 2017; Davison et al., 2021). The initial questions (Topic I, Questions 1-2) served to familiarise children with the interview process and recall the context of the activity. Subsequent questions (Topic II, Questions 3-8) focused the children's attention on their experiences with the robot. The final set of questions (Topic III, Questions 9-15) delved into specific aspects of the robot's social capabilities. A mix of close-ended and open-ended questions was used to ensure thorough exploration of topics and to identify any inconsistencies in responses. Adequate time was given for children to express additional thoughts or details beyond the structured questions.

- Picture Task: As an alternative mode of expression, the picture task involved a series of images representing relevant role-models (Figure 3.3).

The selection of the pictures was based on previous research on perceptions about robots. For example, the teacher represents someone who instructs (Belpaeme et al., 2018), the teddy bear represents a toy (Peca et al., 2014); the notebook represents a non-technological tool; the dog represents a pet (Bartlett et al., 2004); the car represents a technological tool that can be used by humans; the laptop represents a digital tool and the children represent friends (Westlund et al., 2015). The interviewer presented these pictures one at a time, explaining each and ensuring children understood what they represented. Children were then asked to identify which pictures they felt were most similar or dissimilar to the robot and to explain their reasoning. They were allowed to select as many or as few pictures as they felt appropriate.

At the end of the session, the children were thanked for their participation, and each received a diploma acknowledging their contribution to the study.

Figure 3.3: Illustrations representing the images shown during the picture task to assess children's perception of the robot. From left to right: car, teddy bear, notebook, dog, computer, teacher, and friends.

3.3 Data Analysis

Our aim is to evaluate the impact of the system's performance on children's behaviours and perceptions.

3.3.1 Children's behaviours

We considered children performance with the Tower of Hanoi task, children communication with the system and children communication with each other. The analysis of these behaviours were obtained through both automatic and manual annotations from video recordings of the sessions.

- Task Performance, K, was measured using the formula $K = \frac{L-Op}{Op}$, where L is the number of movements performed by the team to solve the task, and Op is the optimal number of movements required for that task. This metric evaluates the efficiency of children's problem-solving in the presence of the robot and is used to classify children into groups based on their performance levels (Low K and High K), with a lower K value indicating better performance. The values L and Op were automatically recorded and calculated by a camera system.

- Help-Seeking Behaviour was assessed by two metrics. HS represents the frequency of help requests made to the robot during the task, defined as $HS = \frac{n_H}{L}$, where n_H is the number of times help is requested, normalised by the number of movements. We categorised teams based on their request for robot's assistance (Low HS and High HS). The time taken to ask for help for the first time, t_H, in seconds, is also recorded. The instances of help-seeking, n_H and t_H, were automatically logged by the system whenever assistance was requested.

- Social Interaction, S, was quantified as the number of task-related interactions between the children $S = \frac{S_1+S_2}{L}$, where S_1 and S_2 represent the number of interactions initiated by each child, including task-related pointing and verbal interactions. These S_1 and S_2 interactions were manually annotated by a Spanish native speaker reviewing the video recordings of the interaction sessions. We also measured the planning disparity, D, as the absolute difference in the number of interactions initiated by each child, $D = |S_1 - S_2|$. A lower D value indicates a more balanced interaction between the children.

3.3.2 Children's perceptions

In this section we summarise the methodology for data analysis based on four data sources: pre-manipulation check data, manipulation check data, recorded interviews data and picture task data.

Pre-manipulation check data

Prior to the main interaction with the robot, a pre-manipulation check assesses children's ability to distinguish between its expressive and neutral behaviours. Using the questionnaire from Appendix A.1, we counted the number of times children correctly recognised the expressive behaviour of the robot when comparing it with the neutral behaviour.

Manipulation check data

After interacting with the robot, the post-manipulation check analyses the consistency and clarity of the robot's behaviour as perceived by the children. Using the questionnaire from Appendix A.2, we map the children's perceived cognitive reliability, $Cognit_p$, and perceived expressivity, $Express_p$, to $\{0,1\}$ values. Later we use a confusion matrix to compare their perception with the actual robot behaviour (the displayed one). The results on cognitive reliability ($Cognit_p$) and expressivity ($Express_p$) are used to split children into groups for further analysis.

Recorded interviews data

From the video recordings, the answers of the interviews are first translated from Spanish to English. Later, these annotations are analysed in the following way:

- *Yes/No questions.* Yes/No answers to questions are translated to 0,1 for statistical analysis. To analyse general tendencies in children's answers, we study the percentage of "yes" answers to each question. For those answers where there is not a high agreement among children (lower than 90% rate), we further study the influence of certain independent variables. In particular, we look for statistical differences in answers when splitting children by age (younger: 5-6 y.o, older: 7-8 y.o.), biological sex (female, male), task performance (high K, low K), help-seeking behaviour (high HS, low HS), manipulation condition (OE,SE,ON,SN), robot cognitive reliability (optimal: OE+ON, sub-optimal: SE+SN), robot expressivity (expressive: OE+SE, neutral: ON+SN) and manipulation check data (perceived cognitive reliability $Cognit_p$ and perceived expressivity $Express_p$). The distribution of these independent variables can be seen at Table 3.5. For the statistical tests, we have in consideration the normality of the data and the outliers.

  Additionally, we study the similarity of individual responses from children within the same group. We define the *answer vector* of a child as the vector formed by their Yes/No answers. We then measure the Euclidean distance between individual *answer vectors*, with a lower Euclidean distance indicating a higher similarity in their responses.

- *Open questions.* In order to have a deep understanding from children's answers, we perform a thematic content analysis considering all the open questions from the questionnaire. We also perform a data mining analysis, focusing on specific questions that demonstrated statistical significance, to further support and refine our insights.
  - For the qualitative analysis, we follow Braun and Clarke (Braun and Clarke, 2006, 2019) method for thematic analysis, to report the perceptions of the participants after their experience of interaction within the hybrid team of two children and a robot. For the development of the annotation scheme, during the experiment the research team meets regularly to discuss patterns observed in children's perceptions. Once all the interviews are transcribed, the research team annotates the data in a deductive way based on the concepts that are predefined and explored by the set of questions, based on the taxonomy for social robots presented by Breazel et al (Breazeal et al., 2016).
  - For the data mining analysis, the annotations from every question are processed to find the most frequent words. First we preprocess all the annotated answer: we remove punctuation marks and stop words (including articles, general stop words and specific stop words such as "robot") and we perform lemmatization and stemming using the *nltk* library in Python (NLTK, 2001) to standardise the representation of similar words (e.g. "coloured" and "color" are both represented as "color"). Finally, for each question, we group together all the preprocessed answers and compute the frequency of words.

Picture task data.

As mentioned in Section 3.2.3, seven pictures were used to represent different roles of the robot: car, dog, computer, notebook, teacher, friends and teddy bear (Figure 3.3). Children pointed at the pictures they found more and less similar to the robot. They could pick as many pictures as they wanted, but for data standardisation, we just considered the first two picks from each child. With the purpose of studying the influence of the robot behaviour on the perception of the robot role, we counted the times a picture was selected. We studied the percentage of children that picked each picture (as a whole and by conditions: OE, SE, ON, SN). We also performed statistical analysis in order to see the different answers in different groups split by age, biological sex, K, HS, condition, robot cognitive reliability, robot expressivity, $Cognit_p$

and $Express_p$. For the statistical tests, we have in consideration the normality of the data and the outliers.

These metrics collectively provide a comprehensive framework for understanding various dimensions of children's interactions with CAs. They allow for a nuanced analysis to understand both the influence of the different behaviours and children perception of the system.

3.4 Results

3.4.1 Children's behaviours

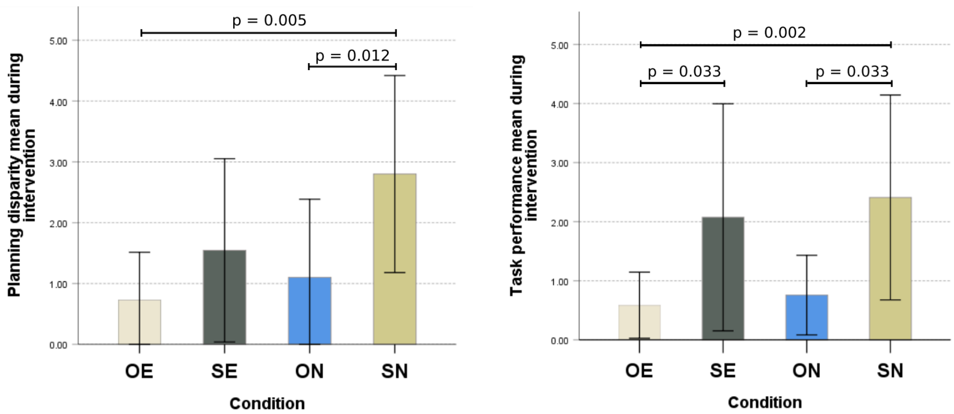
- Task performance. We first examined the task performance K in the different phases of the problem-solving task: BL (Baseline phase), INT (Intervention phase) and EV (Evaluation phase). We ran a Shapiro-Wilk test to evaluate data normality. Based on the results, we ran a non-parametric with three dependent groups (BL, INT, EV) Friedman Test, which indicated a significant difference between these sessions ($\chi^2(2)$=26.09, p <.001). Dunn-Bonferroni post-hoc tests showed a significant difference between the BL (Mean=0.31, SD=0.17) and the INT (Mean=1.32, SD=1.32), and between the BL and the EV (Mean=1.52, SD=1.35) with p <.001 after Bonferroni adjustments in both cases. There was no significance between the INT and the EV. This indicates that there is overall learning throughout the entire problem-solving task.

  In the intervention phase (INT), we ran a non-parametric with two independent groups Mann-Whitney U test which showed a statistically significant difference (U=25, p <.001) on K between the optimal robot behaviour — OE and ON — (Mean=0.41, SD=0.28) compared to the sub-optimal robot behaviour — SE and SN — (Mean=2.23, SD=2.80) (Figure 3.4b). This indicates that the sub-optimal behaviour negatively impacted the team's performance.

  Finally, for the evaluation phase (EV) we examined K distribution for all conditions (OE, ON, SE, SN), but we did not observe any significant differences among conditions. This indicates that neither the expressivity nor the cognitive reliability of the system enhanced children's learning.

- Help-Seeking Behaviour. We tested the *HS* distribution for all conditions (OE, ON, SE, SN) during the evaluation phase, but we did not observe any significant differences among conditions. This indicates that neither the expressivity nor the cognitive reliability of the system enhanced children's requests of help.

However, regarding t_H, a Mann-Whitney U text showed that the teams in the sub-optimal conditions (SE and SN) asked faster for help at the evaluation phase (Mean=47.53, SD=33.51) compared to the teams in the optimal conditions (OE and ON) (Mean=98.44, SD=70.21), with a significant difference of U=45.5, p=.009. This suggests that sub-optimal CAs lead to quicker requests for help from children.



(a) Task Performance, K, by condition.

(b) Planning disparity, D, by condition.

Figure 3.4: Mean plots by the four experimental conditions: optimal expressive (OE), sub-optimal expressive (SE), optimal neutral (ON) and sub-optimal neutral (SN).

- <u>Social Interaction.</u> We examined the effect of robot behaviour manipulation on S during the three phases of the problem-solving task: BL, INT and EV. Overall, a Mann-Whitney U test revealed that in all sessions older children had a higher rate in S (Mean=0.15, SD=0.12) than the young children (Mean=0.09, SD=0.09), U=2018, p=.016, indicating that young children interact with each other less often.

  Regarding the intervention phase (INT), there was no significant difference across our four independent groups (OE, ON, SE, SN). However, we found some differences in the balance of the teams as for the team members' verbal planning behaviour D. A non parametric Kruskal-Wallis Test revealed a significant difference between the mean values of at least one pair of groups ($\chi^2(3)$=13.76, p=.003). Posthoc Dunn's pairwise tests showed that the teams in SN were more imbalanced in terms of planning

(Mean=2.33, SD=0.71) than the teams in OE (Mean=0.73, SD=0.79) and ON (Mean=0.78, SD=0.83) with p=.005 and p=.012 adjusted with the Bonferrroni correction respectively (Figure 3.4a). This indicates that sub-optimal cognitive reliability of the embodied CA may lead to imbalanced interactions between children.

In addition, a Mann-Whitney U test showed that there was a significant difference in K (U=297, p <.001) between teams with a balanced planning performing better (Mean=0.51, SD=0.40) compared to groups with an imbalanced planning behaviour (Mean=1.61, SD=0.98). This suggests that low task performance may lead to imbalanced communication between children.

3.4.2 Children's perceptions

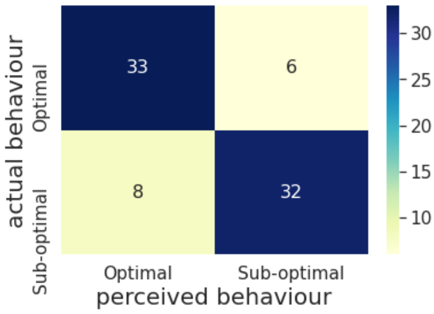
Pre-Manipulation Check data: understanding of the robot expressivity

As mentioned in Section 3.3, the pre-manipulation check evaluates children's understanding of the expressive robot behaviour. In 84% of the examples, children correctly identified the expressive behaviour in comparison with the neutral one.

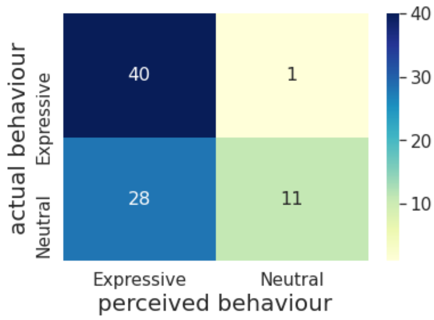
Manipulation Check: perceived robot behaviour

This post-interaction check evaluates how accurately children perceive the robot's behaviours. Figure 3.5c shows a confusion matrix built from manipulation check data and represents the robot perceived vs actual behaviour. We observe how expressive behaviours are mostly correctly perceived (higher diagonal values), while neutral ones are frequently confused (lower diagonal values). We observe an overall perception accuracy of 57%, that we further study according to the two main behaviour factors as follows.

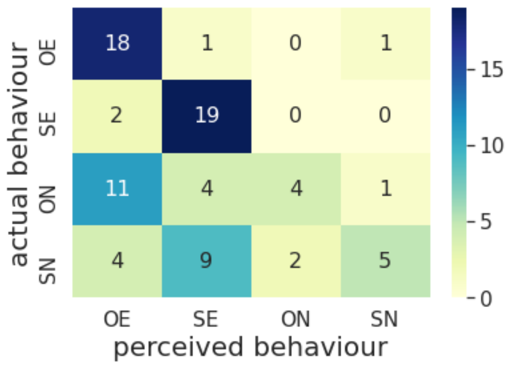
- Perceived cognitive reliability. Figure 3.5a shows a confusion matrix representing children's perception of the robot cognitive reliability. We observe a perception accuracy of 82%, higher than the overall perceived behaviour accuracy mentioned above, which reflects a more accurate perception of the cognitive factor. This finding is complemented by a Mann-Whitney U test (U=1284, p <.001) that showed that children who experienced the optimal behaviour perceived a higher cognitive reliability, $Cognit_p$, (Mean=0.84, SD=0.06) than those that experienced the sub-



(a) Cognitive reliability. The diagonal elements of this matrix are clearly defined.



(b) Expressivity. While the expressive robot is mostly correctly perceived, the neutral robot tends to be perceived as expressive.



(c) Robot behaviour. OE and SE are correctly perceived, while ON and SN tend to be confused.

Figure 3.5: Factorial confusion matrices built from manipulation check data. Rows represent the actual behaviour of the robot, while columns represent the perceived behaviour.

optimal one (Mean=0.20, SD=0.07). The high accuracy of this perception suggests that children can correctly identify some sub-optimal movements or errors by the robot. We also analysed statistical differences of perceived cognitive reliability in terms of task performance K as introduced in Section 3.3. A Mann-Whitney U test (U=801.5, p =.030) indicated that $Cognit_p$ was significantly higher for children that performed better with the Tower of Hanoi (Mean=0.60, SD=0.07) than from children with a lower task performance (Mean=0.32, SD=0.10).

- Perceived expressivity. The confusion matrix built on the expressivity factor of the perceived behaviour (Fig.3.5b) shows a perception accuracy of 63% , which is lower than that of cognitive reliability. To further study these results, we performed a Mann-Whitney U test (U=1005.5, p =.001) showing that even when there is a general tendency to perceive the robot as expressive, children that experienced the expressive robot perceived it more often as expressive (Mean=0.93, SD=0.02) than children that experienced the neutral one (Mean=0.72, SD=0.07).

Interviews: yes/no questions

This section presents the analysis of binary interview questions that could be answered with yes/no. Table 3.4 presents the percentage of positive 'Yes' answers for each question. We observe a high agreement (above 90%) in questions 1, 4, 5, 7, 8, 11, 12 and 13, suggesting that children mostly enjoyed the experiment, think that the robot can and wants to be their friend, consider that the robot can see and hear them, would like to have a robot, think that the robot would like to have a lot of friends and believe that the robot is not able to harm anybody.

Table 3.4: Interview questions and corresponding percentage of positive answers. High agreement (above 90% agreement) are marked with *.

Question	Yes%
1) *Did you enjoy it?*	**100***
2) *Can you tell me something about the robot?*	-
3) *How would you describe the robot to your friends?*	-
4) *Could the robot be your friend? Why?*	**94***
5) *Do you think that the robot would like to be your friend? Why?*	**98***
6) *How would you describe the robot at home?*	-
7) *Could the robot see you? Why?*	**97***
8) *Could the robot hear you? Why?*	**100***
9) *Did the robot help you? Why?*	88
10) *Did you help the robot? Why?*	78
11) *Would you like to have a robot? Why?*	**90***
12) *Do you think that the robot would like to have a lot of friends? Why?*	**99***
13) *Could the robot be able to harm anybody? Why?*	**9***
14) *Is the robot able to help you with your homework?*	81
15) *Does the robot know everything?*	49

As mentioned in Section 3.3, we carried out statistical analysis on the interview questions with disparity answers (9, 10, 14, 15) looking for independent variables that may influence children's answers. Shapiro-Wilk tests showed that our data are not normally distributed and, after considering outliers, we performed Mann-Whitney U tests (Table 3.5) for our independent variables: age, biological sex, task performance, help-seeking behaviour, robot cognitive reliability, robot expressivity, perception of the robot's cognitive reliability and perception of robot's expressivity. For the independent variable "robot condition", as it splits children in four different groups (OE,SE,ON,SN), a Kruskal-Wallis test was required. We obtained significant differences among independent variables in questions 9, 10 and 14 marked with asterisks in Table 3.6 that we comment below. We did not find statistical differences in Question 15 "Does the robot know everything?", neither in the rest of the questions (1, 4, 5, 7, 8, 11, 12 or 13).

Table 3.5: Description of our independent variables: age, biological sex (Sex), task performance (K), help-seeking behaviour (HS), robot cognitive reliability, robot expressivity, children's perceived cognitive reliability ($Cognit_p$), children's perceived expressivity ($Express_p$) and robot condition. Note that not all children answered to questions regarding $Cognit_p$ and $Express_p$ in the manipulation check.

Independent variable	Groups and N				Total
Age	younger	29	older	52	81
Sex	female	35	male	46	81
K	high	24	low	57	81
HS	high	58	low	23	81
Cogn.Rel	optimal	40	sub-optimal	41	81
Expressivity	expressive	41	neutral	40	81
$Cognit_p$	optimal	41	sub-optimal	38	79
$Express_p$	expressive	68	neutral	12	80
Condition (4 groups)	OE (20)	SE (21)	ON (20)	SN (20)	81

Table 3.6: p values of Interview questions regarding our independent variables: Age, Sex, K, HS, Cognitive Reliability (Cognit), Expressivity (Express), $Cognit_p$, $Express_p$ and Condition (Cond) showed in Table 3.5. Statistically significant differences ($p \leq .05$) are marked with *.

Question	Age	Sex	K	HS	Cognit	Express	$Cognit_p$	$Express_p$	Cond
9) Help you	.228	.444	.789	**.040***	**.038***	.725	.368	.774	.352
10) Your help	.925	.603	.403	.477	.421	.903	**.001***	.875	.814
14) Homework	**.018***	.901	**.001***	.746	.490	**.001***	.528	.443	.298
15) Know everyth.	.295	.707	.845	.399	.566	.909	.530	.499	.950

- Question 9: *"Did the robot help you?"* (88%) Even if the tendency is to perceive the robot as a helper, we found a significant difference (U=740, p =.038) indicating that children that experienced the optimal robot behaviour recognised more the robot's help (Mean=0.97, SD=0.03) compared to those children that experienced the sub-optimal robot behaviour

(Mean=0.91, SD=0.05). This suggest that the optimal behaviour of the robot improves children perception of that help. We also found significant differences in *HS* (U=736, p =.040), suggesting that children that asked for help more often, recognised more the robot's help (Mean=1.00, SD=0.00) compared to children that asked for help less times (Mean=0.84, SD=0.06). This result suggest, as expected, that children that ask for help more often, perceive more the robot's help.

- Question 10: *"Did you help the robot?"* (78%) Even if children have the tendency to think that they help the robot, we found some significant differences (U=432, p =.001) showing that children that perceived the robot as sub-optimal reported more often their help to the robot (Mean=0.94, SD=0.04) than children that perceived the robot as optimal (Mean=0.61, SD=0.08).

- Question 14: *"Is the robot able to help you with your homework?"* (81%) As shown in Table 3.6, we found significant differences in three Mann-Whitney U test that we run on this question. The first one (U=693, p =.001) indicates that the expressive robot is more often perceived as homework helper (Mean=1.00, SD=0.00) than the neutral robot (Mean= 0.79, SD=0.08). The second test (U=455, p =.018) shows that older children think more about the robot as a homework helper (Mean=0.90, SD=0.05) than younger ones (Mean=0.79, SD=0.08). The third test (U=693, p =.001) suggests that children with high task performance K perceive the robot as a homework helper more often (Mean=0.92, SD=0.04) than those with low task performance (Mean=0.57, SD=0.11).

- Question 15: *"Does the robot know everything?"* (49%) We also observe a high disagreement among children when answering this question, but we did not find justifications in our independent variables. One possible reason is that the question was not very specific, leading to different understandings, which can be observed in the following children statements: *"the truth is it never failed with the tower"* said child 68, who experienced the ON behaviour (68-ON), *"Some things. It knows Maths, Language, multiplying... but not subtractions"* (56-OE), *"it can make a car, it just doesn't have hands, but it would tell me (how to do it)"* (19-OE). This question even led to some philosophical answers: *"there are still things that humans didn't discover"* (75-SN), *"it sometimes makes mistakes. But everybody makes mistakes"* (32-SN).

- There were not additional statistical differences on the following independent variables: biological sex, condition and perceived expressivity.

Finally, we examined the similarity of individual responses from children within the same group (i.e., pair) by performing a statistical analysis of the Euclidean distance between their *answer vectors* compared to those of children from different groups. After verifying normality, a Mann-Whitney U test revealed no significant difference (U=65655, p=.993) between distances of *answer vectors* from children in the same group (Mean=0.35, SD=0.17) and those in different groups (Mean=0.34, SD=0.21). These results suggest that group interaction did not influence the children's individual perceptions of the robot.

Interviews: open-ended questions

We performed thematic content analysis on children's perceptions about the robot, considering the whole interview session. We also supported our analysis with a study of the most frequent words to certain questions, as mentioned in Section 3.3. We did not find inconsistencies in similar questions.

Concepts concerning robot description:

- Physical entity. All the participant children gave an emphasis on the physical nature of the robot. Often they described its physical characteristics with technical terms in an accurate and literate way. As one child put it, *"It was a circle with a stick, like this, and then from the stick the eyes came out. Two squares"* (27-SN). This is in accordance to the most frequent words used by children in questions 2, 3 and 6 that are related to the robot description: *"eye"*, *"like"* and *"square"* (Table 3.7). Despite the fact that the participants experienced an interaction with a table-top robot without limbs, they often envisioned this robot being able to engage in activities that would require increased physical affordances. These descriptions often focused on actions rather than the physical characteristics of the robot. A child, for example, mentioned *"It could help me to tidy things, because I do many activities"* (68-ON), while another participant said that *"It dances... sings"* (03-ON). The association of the physical affordances of the robot with the execution of certain actions, even beyond the ones that the children experienced during their interaction with the robot, indicate the relevance of children's perceptions of the physical entity of the robot and the attribution of agency to it.
- Intentionality. Certain participants recognised that the robot was programmed by humans. For example, the child (73-SE) said that robots *"have eyes and they are programmed to see"*. This description indicates

that the child is aware of the fact that the robot is subject to human design and development; however, in their further description, the same child mentioned that *"it responds to you when you talk"* (73-SE), showing that the robot acts intentionally. In a similar way, another child mentioned that *"the robot is made to hear"* (84-SE), which indicates a difference in their perception of robot intentionality from children that described the robot as being an intentional agent. The majority of children used an active voice for the description of robot behaviours. Interestingly, this applied also in cases where the children talked about a negative behaviour of the robot, such as the participant (01-ON), *"The only thing is, if you tell it not to tell lies, it will keep telling lies. The only thing"*.

- Autonomy. Different to intentionality, the autonomy of the robot indicates that the robot performs in an autonomous way, including both perception and action. The robotic system we used for children's interaction was autonomous. However, it was not surprising that some children emphasised the fact that the behaviour of the robot depends on human decision-making. One child for example mentioned *"it can be a robot, but not be programmed to hurt, to say bad things"* (47-SE). However, the majority of the children described various scenarios regarding the use of the robot in a way that indicated their perception about the robot being autonomous. The actions that were described mostly by the children were those of playing together and helping each other. One child, for example, mentioned *"It can help me with my games, it can play with my siblings and with me"* (77-SN), and similarly, child (85-SN) mentioned *"it can talk, it can sing. It could have more emotions than two"*. These phrases indicate the way children perceive the autonomy of the robot.

Concepts concerning help:

- Robot helping. Some children commented on the robot's prosocial behaviour in reference to the designed social dynamics that all children experience. For example, the child (71-ON) mentioned *"it helped us to place the blocks; because it was Sofía's turn, then my turn, and then Haru's"*. Children tend to perceive that the robot helps them by telling what movement they could perform. However, looking at the most frequent words of question 9 *"Did the robot help you?"*, split by cognitive reliability (as there was statistical difference), we observe that the word *"mistakes"* frequently appears for children interacting with the robot with a sub-optimal behaviour (see Table 3.7). Child (09-SN), for instance, mentions *"I ask*

Table 3.7: Most frequent words on children's answers

Question	Split by	Frequent words
Description	None (all together)	eye, like, square
Homework	Younger	hand, tell, clean
	Older	help, tell, understand
	Expressive	tell, could, answer
	Neutral	let, hand, could
	High K	tell, let, answer
	Low K	hand, could, good
Robot Helping	Optimal	disk, told, put
	Sub-optimal	ask, said, mistake
Child Helping	Perceived Optimal	move, told, put
	Perceived Sub-optimal	said, mistake, put

it when I need help with the discs we have, but if it makes mistakes, you don't have to do what it says". In addition, the word *"told"* or *"know"* are dominant in answers by children interacting with the robot with an optimal behaviour; for example, child (48-ON) mentioned: *"as I didn't know where to put the disc, I asked Haru for help"*.

- Child helping. Most children perceived that they helped the robot by moving the disk as the robot does not have hands, e.g. child (61-OE) said: *"In that game, as it didn't have hands... I placed the discs on its place"*, but some children acquired the role of supervising the robot when it did mistakes, as child (51-ON): *"When it fails, we tell it that's not right"*, or child (73-SE) that mentioned: *"I can play this game and, if it has a mistake, I'll help it"*. In addition, the most frequent words of question 10 *"Did you help the robot?"*, split by the perceived cognitive reliability (where we found out statistical difference), and observed in Table 3.7, were *"move"* and *"put"* for the optimal robot, and *"mistake"* was a frequent one in answers by children interacting with the sub-optimal robot. Some children relied on explicit requests for help in order to proceed to prosocial behaviour towards the robot, such as in the case of child (10-SN), who commented *"we didn't have to help it. it didn't say 'help!' "*.

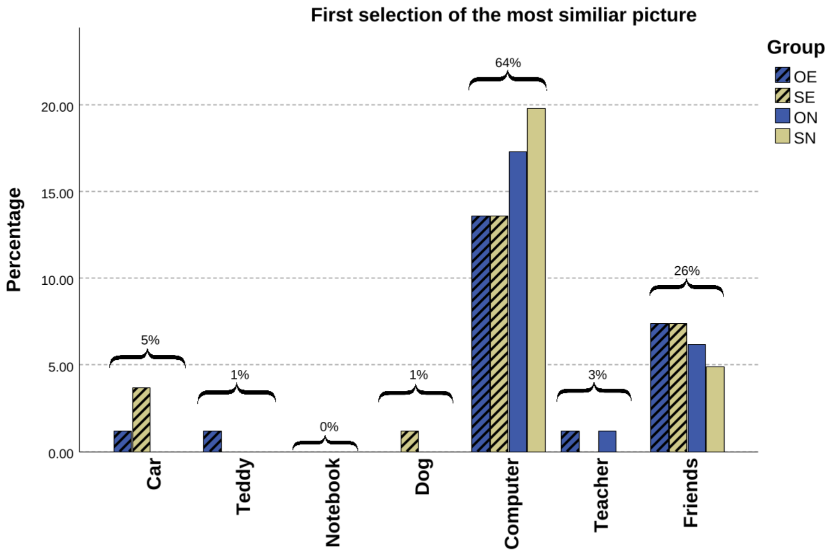
- Shared responsibility. The triadic interaction of turn-taking during the intervention seems to have supported a sense of task-related shared responsibility. However, in principle implicit rules based on the sense of shared responsibility were developed over time that created a common ground along with the members of the team that works *"we asked for help, we answered correctly and the robot gave us clues"* (77-SN).

Concepts concerning homework:

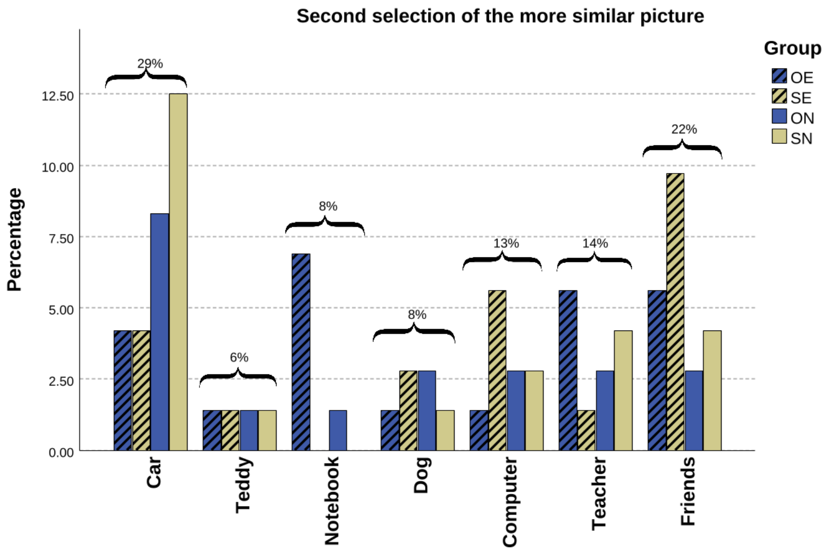
- <u>Robot competence.</u> Most children consider that the robot can help them with their homework, for example child (85-SN), who mentioned: *"I think that it was... it would be a good robot to help me with my homework"*. Many children (65%) justified their answers to question 14 *"Is the robot able to help you with your homework?"* with the sentence *"I don't know (why)"*, what increased the difficulty of the analysis. However, many positive answers are related to the robot's intelligence and understanding, e.g. child (83-SE) *"it can see a question of my homework, it could help me to understand it"* or child (28-SN) *"it could tell me where to write something"*. We observe that young children (5-6 y.o.) tend to mention the robot lack of hands to justify the robot inability to help with their homework, as child (06-SE): *"No. It has no hands, or legs, or body"*, or child (30-SE): *"it doesn't have hands!"*. This is in accordance with Table 3.7, where "hand" is the most frequent word included in answers by younger children. This fact could be explained because young children do not usually have homework and they perform a lot of manual activities in class.
- <u>Robot mistakes.</u> Surprisingly, children's perception of robot's mistakes did not seem to affect their answers; child (33-SN) mentioned: *"if it gives me the wrong answer, I erase it. Because I do my homework with a pencil"*. This is in accordance with results in Section 3.4.2 where not statistical differences were found considering robot's cognitive reliability. Nevertheless, the robot's expressivity did impact children's perceptions regarding homework helping. One possible explanation could be the experienced sharing responsibility and the perception of the robot as a friendly tool to team up with.

Picture Task: robot role

As mentioned in Section 3.3, we studied the answers of the picture task in order to understand the perception of the robot role. Children more often picked first the picture representing the concepts 'Computer' (64%) and 'Friends' (26%) as similar options to the robot with an absolute absence of the 'Notebook'



(a) Frequency of the more similar pictures (first pick).



(b) Frequency of the second pick of the more similar pictures.

Figure 3.6: Picture task frequency

concept (Fig.3.6a). Children emphasised the robot's digital characteristic with some social attributions.

If we consider the different robot behaviours, we observe that children that interacted with the expressive robot chose less often the 'Computer' picture and more often the 'Friends' one, compared to those interacting with the neutral robot. This illustrates the impact of the expressivity on children perception. In addition, the Car's picture was only first picked by children interacting with the expressive robot. This might be due to the fact that children that interacted with the expressive robot were looking for a middle term between 'Computer' and 'Friend', and cars that can talk are a suitable middle point. (30-SE) said: "*I have a car, and when I go with my brother, the car displays a face with a red mouth and two antenna*", and child (65-SN) mentioned: "*it's technology too. A lot of cars can speak nowadays*". We also should consider the fact that 'Car' is the most popular second pick for all conditions (29%). Fig.3.6b also shows that the picture associated to 'Friends' is dominant as a second pick by children interacting with the expressive robot, while the ones collaborating with the neutral one use 'Car' as dominant option.

Further analysis of the picture task involving the independent variables of Table 3.6 showed some statistical differences: a Mann-Whitney test indicated that the frequency of the picture "Teacher" was significantly higher (U=741, p =.028) for children that asked for help more often (Mean=0.06, SD=0.24) than the others (Mean=0.00, SD=0.00); also other Mann-Whitney test suggested that the frequency of the role "Friends" was significantly higher (U=769, p =.016) for children that had a better task performance (Mean=0.29, SD=0.07) than the others (Mean=0.15, SD=0.82).

3.5 Conclusions

This chapter shifts from theoretical discussions to empirical analysis by examining children's interactions with CAs through a user study. The user study provides essential insights into how children use and perceive these technologies, identifying specific areas for optimising CAs to better meet children's developmental needs and preferences. The study involves a comprehensive experimental approach with a social robot and children in a controlled environment, allowing us to observe their behaviours and opinions. In particular, it sheds light on the subtle ways in which a CA's speech and behaviour may affect children's actions and perceptions.

We summarise here the results of the study with a focus on the spoken dialogue during the interaction.

The analysis indicated that children perceive the social robot as a digital entity with social attributes. They generally view the robot as friendly, capable of perception, and harmless, with over 90% of responses reflecting this sentiment. Additionally, children successfully improved their ability to solve the Tower of Hanoi through interactions with the robot, regardless of its behaviour, highlighting how child-CA interactions can positively impact learning and problem-solving skills.

The thematic analysis highlighted two important concepts: intentionality and autonomy. Intentionality reflects how children often saw the robot's actions as intentional, attributing a sense of purpose to its interactions, even while understanding it was programmed by humans. Autonomy pertains to the children's view of the robot as capable of independent action within its programming constraints. Children described the robot as able to perform tasks like playing games without direct human control. These perceptions raise questions about agency, specifically where intentionality meets autonomy, and whether children believe the system can act beyond its programming.

Additionally, children were affected by the different robot's behaviours, notably in two areas:

- **Cognitive reliability.** On the one hand, when the robot provided suboptimal suggestions during the Tower of Hanoi task, the overall task performance declined, and group dynamics changed, as one child often took the lead. Children who perceived the robot's low performance recognised their own help to the robot by pointing out its incorrect suggestions. Despite knowing the suggestions were wrong, they still followed them, leading to worse overall task performance but a better-perceived help relationship between the child and the robot. This collaboration, combined with a likely lack of confidence in solving the task, resulted in children asking for the robot's assistance more quickly. Surprisingly, even when children recognised the system's mistakes, it did not diminish their view of the robot as a helpful tool for homework. On the other hand, children in the condition of the optimal robot, who experienced high performance from the robot, were more self-assured in solving the task, with balanced interactions, and took longer to request for help, even though they perceived more the robot's help during the game.

- **Expressivity.** A high level of expressivity displayed by the robot during the interaction led children to perceive it more as a friend than a machine. Children also viewed the robot as more capable of assisting them with their homework, even when its cognitive reliability was sub-optimal, emphasising the social support it provided rather than its accuracy in the homework task. Interestingly, children were able to distinguish the behaviour of the expressive robot when comparing to the neutral one; however, during the interaction there was a tendency to view the robot as being expressive even when it was not.

Age differences also emerged in the study. Younger children interacted less frequently with each other compared to older children, and had varying perceptions of the robot's capabilities, such as its potential to assist with homework. These differences highlight the importance of tailoring CAs to meet the developmental and cognitive needs of different age groups.

Overall, this experiment provides valuable insights into children's perceptions and actions towards an embodied CA and how the system's behaviour influence them. These findings offer important perspectives to the ongoing discourse on ethical AI, setting the stage for developing guidelines and strategies to optimise these interactions for children benefits. The subsequent work, outlined in Chapter 4, will present our efforts to adapt existing trustworthy AI guidelines to address the specific needs of CAs and children.

Chapter 4

Ethical guidelines

4.1 Introduction

Building on the foundations laid in Chapters 2 and 3, this chapter shifts our focus towards the ethical design and deployment of Conversational Agents (CAs) specifically tailored for young users. Chapter 2 introduced the fundamental concepts of CAs, emphasising the need for ethically designed systems that address both the opportunities and challenges these technologies present for children. Chapter 3 provided empirical insights through a study of children's interactions with a social robot, highlighting how a CA's speech and behaviour can influence children's perceptions and actions. These findings underscore the importance of developing guidelines that ensure these systems are not only effective but also trustworthy and aligned with children's developmental needs.

In this chapter, we address the identified gap in ethical CA design by creating tailored guidelines that enhance trustworthiness for young users. Leveraging the insights from our previous work, we adapt existing AI ethical frameworks, such as the ethics guidelines for trustworthy AI, to the specific context of CAs from a child-centred perspective.

This chapter begins with a detailed explanation of the selected methodology for the adaptation of the ethics guidelines for trustworthy AI to our particular context, including the application of the Delphi method and the involvement of multidisciplinary experts to ensure a comprehensive approach. We also describe how risk assessment measures were utilised to gather expert insights on the potential risks associated with CAs and children. The chapter then presents the results of these expert evaluations, combining quantitative analysis to prioritise critical points and qualitative thematic analysis of the expert feedback, which informs the development of our guidelines. Finally, our discussion offers concrete recommendations for developers, emphasising the need to prioritise trustworthiness in designing CAs for children as potential users.

By providing these tailored guidelines, this chapter aims to enhance the opportunities that CAs offer while minimising the risks to this developing population. These contributions lay the groundwork for future research and practical applications, ensuring that CAs are designed with children in mind and are trustworthy.

4.2 Methodology

Drawing from the insights presented in Chapter 2, we propose a methodology to adapt existing ethical guidelines for trustworthy AI (HLEG, 2019) to the use of CAs with children, considering the Assessment List for Trustworthy AI (ALTAI) which is based on a self-assessment of the different trustworthy requirements. In this process, we prioritized and identified action points by conducting a risk level analysis for each ALTAI item (question), using the following metric:

$$Risk = Likelihood \times Impact \tag{4.1}$$

To determine these measures ($Likelihood$ and $Impact$) we employed the Delphi method (Linstone and Turoff, 1975), involving four experts by means of a survey and a set of group discussions. This limited set of experts could be considered a drawback, but given the difficulty in finding individuals with expertise in this niche topic and the variety of required profiles, this approach was deemed appropriate. After the experts evaluated the $Likelihood$ and $Impact$ of each ALTAI item, we assessed the overall risk using the matrix approach (Kovacevic et al., 2019; Ristić, 2013), which allowed us to identify critical points.

ALTAI question	Dose it apply to children? (*Likelihood*)	Is it relevant for children? (*Impact*)	Does it apply to CAs? (*Likelihood*)	Is it relevant for CAs? (*Impact*)	Notes
"Do you communicate to users that they are interacting with an AI system instead of a human?"	Always Sometimes Never	High Medium Low	Always Sometimes Never	High Medium Low	

Figure 4.1: Example from the ALTAI questionnaire used to assess the *Risk* (based on *Likelihood* and *Impact*) in the context of children and CAs.

4.2.1 Delphi method

To implement the Delphi method, we designed a questionnaire in which experts were asked to rate each ALTAI question in two specific contexts: children compared to the general population, and CAs as opposed to AI systems in general. The evaluation was based on two main criteria: the likelihood or frequency of application (rated as always, sometimes, or never) and the impact or significance (rated as high, medium, or low). These 3-point Likert scales (Likert, 1932) were chosen for simplicity, making it easier for experts to consistently assess the 69 questions, each with four ratings (Figure 4.1). Additionally, we provided space for experts to elaborate on their ratings, offering insights and reasoning as needed.

We provided the questionnaire to experts from diverse disciplines, including AI ethics, CAs, and education. Each expert independently completed the questionnaire based on their own criteria. The individual responses were then analysed to identify areas of disagreement and critical points, with an initial agreement rate of 74%. Following this analysis, we organised two meetings with the experts to discuss the identified critical points and disagreements, aiming to establish common criteria and reach a consensus. After the meetings, the experts were given the opportunity to review and refine their individual responses before submitting their final versions, which resulted in an increased agreement rate of 84%.

4.2.2 Risk assessment

For each of the ALTAI items (i.e. questions) we computed partial risk levels for children and CAs, as follows:

- **Risk value for questions.** For each ALTAI item, we calculated the arithmetic mean of the four *Likelihood* and *Impact* ratings provided by the experts, separately for children and CAs. This allowed us to derive the *Child Risk* and *CA Risk* (referred to as *partial risks*) using Equation (4.1) with the respective *Likelihood* and *Impact* means. Additionally, we constructed a risk assessment matrix to evaluate the level of risk in our results.
- **Risk value for requirements.** To assess the risk for each HLEG ALTAI requirement, we first computed the arithmetic mean of the *Likelihood* and *Impact* ratings for all the questions within a given requirement (e.g., the *Human Agency and Oversight* requirement consists of eleven questions). This calculation was done separately for children and CAs, resulting in the *Child Risk* and *CA Risk* for each requirement by applying Equation (4.1).

From individual *partial risks* (*Child Risk* and *CA Risk*) per question and requirement, we calculate the *Total Risk* of every question and requirement by the following equation:

$$Total\ Risk = Child\ Risk \times CA\ Risk \tag{4.2}$$

Finally, we created a risk assessment matrix for the *Total Risk* (Figure 4.2) to understand the severity of the risk levels. Detailed results can be accessed at https://github.com/mescpla/CAs4Children-ETHICOMP22.git.

4.2.3 Thematic Analysis

In order to complement the risk assessment, we combined all notes provided by the experts in the questionnaire and the critical points discussed during the Delphi meeting. We then conducted a thematic analysis (Braun and Clarke, 2006). First, we compiled all the annotated comments, identifying key ideas. Secondly, we grouped these ideas into potential topics, which we discussed and refined until reaching our final version. Finally, we selected the examples to be included in the report.

	Children					CAs					Total Risk
	Likelihood		Impact		Child Risk	Likelihood		Impact		CA Risk	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD		
HUMAN AGENCY AND OVERSIGHT	**2,00**	**0,45**	**2,61**	**0,05**	**5,21**	**2,33**	**0,16**	**2,52**	**0,68**	**5,87**	**30,59**
Human Agency	2,15	0,50	3,00	0,00	6,45	2,87	0,23	2,93	0,94	8,41	54,24
Human Oversight	1,80	0,40	2,13	0,12	3,84	1,80	0,00	2,00	0,40	3,60	13,82
TECHNICAL ROBUSTNESS AND SAFETY	**1,54**	**0,46**	**2,19**	**0,16**	**3,38**	**2,03**	**0,33**	**2,02**	**0,45**	**4,10**	**13,85**
Resilience to Attack and Security	1,38	0,58	2,06	0,19	2,83	2,11	0,19	2,06	0,48	4,34	12,27
General Safety	1,40	0,20	2,13	0,23	2,99	1,87	0,23	2,00	0,41	3,73	11,15
Accuracy	1,90	0,36	2,40	0,23	4,56	2,13	0,35	1,93	0,46	4,12	18,81
Reliability, Fall-back plans and Reproducibility	1,53	0,68	2,20	0,00	3,37	2,00	0,58	2,07	0,46	4,13	13,94
PRIVACY AND DATA GOVERNANCE	**1,83**	**0,76**	**2,72**	**0,38**	**4,99**	**2,67**	**0,48**	**2,61**	**0,78**	**6,96**	**34,75**
Privacy	1,88	0,89	2,67	0,58	5,00	2,83	0,29	2,67	0,84	7,56	37,78
Data Governance	1,81	0,69	2,75	0,29	4,98	2,58	0,58	2,58	0,74	6,67	33,26
TRANSPARENCY	**1,90**	**0,40**	**2,40**	**0,23**	**4,56**	**2,27**	**0,35**	**2,27**	**0,59**	**5,14**	**23,43**
Traceability	1,00	0,00	2,00	0,00	2,00	2,67	0,58	2,00	0,59	5,33	10,67
Explainability	2,00	0,50	2,50	0,58	5,00	2,00	0,00	2,17	0,48	4,33	21,67
Communication	2,25	0,50	2,50	0,00	5,63	2,33	0,58	2,50	0,69	5,83	32,81
DIVERSITY, NON-DISCRIMINATION AND FAIRNESS	**1,75**	**0,44**	**2,37**	**0,40**	**4,14**	**2,63**	**0,46**	**2,17**	**0,64**	**5,71**	**23,63**
Avoidance of Unfair Bias	2,00	0,53	2,53	0,46	5,07	2,80	0,35	2,33	0,73	6,53	33,10
Accessibility and Universal Design	1,31	0,33	2,17	0,29	2,84	2,50	0,58	1,92	0,53	4,79	13,63
Stakeholder Participation	2,25	0,50	2,33	0,58	5,25	2,33	0,58	2,33	0,60	5,44	28,58
SOCIETAL AND ENVIRONMENTAL WELL-BEING	**1,27**	**0,53**	**2,04**	**0,89**	**2,59**	**1,50**	**0,94**	**2,00**	**0,33**	**3,00**	**7,78**
Environmental Well-being	1,00	0,00	3,00	0,00	3,00	1,67	1,53	1,33	0,25	2,22	6,67
Impact on Work and Skills	1,43	0,84	1,47	1,42	2,10	1,33	0,69	2,20	0,33	2,93	6,17
Impact on Society at large or Democracy	1,00	0,00	3,00	0,00	3,00	2,00	1,00	2,33	0,52	4,67	14,00
ACCOUNTABILITY	**1,22**	**0,34**	**2,54**	**0,51**	**3,10**	**2,42**	**0,83**	**2,04**	**0,55**	**4,93**	**15,28**
Auditability	1,00	0,00	2,17	0,29	2,17	3,00	0,00	2,00	0,67	6,00	13,00
Risk Management	1,29	0,45	2,67	0,58	3,44	2,22	1,10	2,06	0,51	4,57	15,73

Figure 4.2: Values for *Likelihood* and *Impact* for calculating the *partial risks* (*Child Risk* and *CA Risk*) and the *Total Risk* per each HLEG ALTAI requirement and sub-requirement. Risk levels are colour-coded according to the scheme presented in Figure 4.3.

4.3 Results

Ordered assessment list

Figure 4.2 presents the mean *Likelihood* and *Impact* ratings provided by the experts for both the child and CA categories. It also includes the calculated *partial risks* and *Total Risk* for each question and requirement. It is noteworthy that, overall, the *Impact* values are generally higher for children than for CAs. However, the *Likelihood* tends to be significantly greater for CAs compared to children, explaining why the *CA Risk* is higher than the *Child Risk* in most categories, with exceptions in *Human Oversight*, *Accuracy*, *Explainability*, and *Environmental Well-being* (Figure 4.2).

The partial risk assessment matrix (Figure 4.3a) highlights that the experts identified the primary critical points for children as *Human Agency and Oversight*, *Privacy and Data Governance*, and *Transparency*. In the case of CAs, the main critical requirements are *Privacy and Data Governance*, *Human Agency and Oversight*, and *Diversity, Non-discrimination, and Fairness*. When considering the *Total Risk* and examining the combined risk assessment matrix (Figure 4.3b), *Privacy and Data Governance*, along with *Human Agency and Oversight* (particularly focusing on Human Agency), emerge as the most critical requirements when developing CAs for children. Interestingly, the only requirement with values lower than the matrix diagonal, indicating the lowest partial and combined risk levels, is *Societal and Environmental Well-being*.

Likelihood \ Impact	1	2	3
1	1	2	3
2	2	4	6
3	3	6	9

(a) *Partial risks*: *Child Risk* and *CA Risk* based on *Likelihood* and *Impact*, with numbers and colours indicating the relevance of each magnitude shown in Figure 4.2.

Child risk \ CA risk	1	2	3	4	6	9
1	1	2	3	4	6	9
2	2	4	6	8	12	18
3	3	6	9	12	18	27
4	4	8	12	16	24	36
6	6	12	18	24	36	54
9	9	18	27	36	54	81

(b) *Total Risk*: Combined *Child Risk* and *CA Risk*, with numbers and colours highlighting the overall risk levels, corresponding to the data in Figure 4.2.

Figure 4.3: Risk assessment matrices.

Thematic Analysis

During the Delphi meeting, our of experts (R1, R2, R3, and R4 shown in Table 4.1) engaged in discussions and provided critical insights, with additional considerations noted in the questionnaire (Figure 4.1). As described in the methodology (Section 4.2), we conducted a thematic analysis to identify key

Table 4.1: Areas of expertise of the participating experts.

Expert ID	Primary Expertise
R1	Conversational Agents and Ethical AI
R2	Ethical AI
R3	Conversational Agents
R4	Education and Learning Technologies

considerations for the ethical design of child-centric CAs. These considerations are visually summarised in Figure 4.4 and Figure 4.5, and are detailed below:

- **Involve children stakeholders.** Experts highlighted the importance of involving children and their stakeholders (such as children, teachers, and parents) in the design, use, and testing of CAs. For instance, R1 suggested to *"Include children, tutors, and teachers as stakeholders"*, and R2 noted that *"Multiple stakeholders need to be involved"*. Additionally, experts noted the importance of educating stakeholders so they can effectively oversee the system, with R2 annotating the need to *"teach stakeholders"*. Importantly, expert's agreed that children's involvement should be meaningful and enjoyable, as R4 pointed out: *"We need to involve children in the design, but in a meaningful way as this participation should be far from job conditions"*.

- **AI awareness.** The experts expressed concerns about the need for users, especially children, to understand what a CA is, how it functions, and its limitations. Several points were raised in this regard:

  - *Non-human nature.* Experts warned that the natural communication of CAs might cause confusion about their true nature. For example, R2 noted that *"Naturalness of CAs might create confusion"*. Importantly, there is a risk of attachment issues, particularly for children who are still developing their social and cognitive skills. R4 cautioned about *"human attachment as children are developing their cognitive and emotional abilities"*, while R1 mentioned that *"Children can think that something that is not alive has alive characteristics such as feelings"*. Additionally, R4 stressed the importance of clearly communicate the CA's social role and lack of authority: *"Children might understand what the system is and that it has no authority"*. The experts also discussed that consistency in the information provided by the CA would be beneficial, with R4 noting that *"If we want*

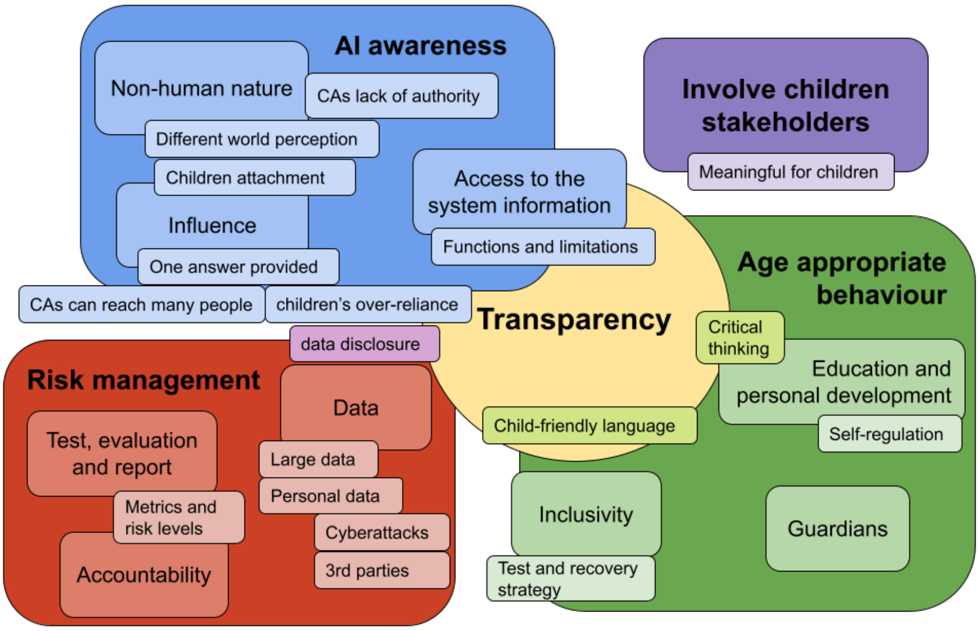


Figure 4.4: Annotation scheme of experts' comments.

a child to understand that a CA has no feelings, maybe it is better to avoid sentences such as 'I am happy' ".

– *Influence.* The potential influence of CAs on users was another major concern. R3 remarked that *"CAs can provide information critical to making decisions. People don't usually double-check information".* This highlights the need for caution, especially when it comes to children's developing critical thinking skills. R4 emphasised that *"Children need to learn to be critical, to look further than the provided information".* The experts also showed concerns about children's vulnerability to over-reliance on CAs, with R2 noting, *"Check over-reliance. Consider children's vulnerability"*, and R3 adding, *"Special attention to kids over-reliance".* Given that children are highly impressionable and still developing judgement and awareness, particularly in areas like data disclosure, R4 warned, *"Kids don't yet have good judgement; be careful with influence and over-reliance".* Finally, R2 noted the importance of monitoring the broader social impact of

CAs due to their widespread use: *"CAs can influence a large number of people, so the social impact of CAs should be considered"*.

  - *Access to the system information.* Our experts universally emphasised the importance of transparency regarding the nature, functioning, and limitations of CAs. Providing this information helps promote AI awareness and can mitigate some of the risks previously mentioned. R3 noted, *"Understanding CAs' nature and interactions is important to avoid frustration"*. During the experts' meeting, it was also stressed that this information should be readily accessible to users and proactively provided by the system, tailored to the duration and risks of the interaction. As R1 pointed out, *"Regarding explainability, it should also fit the purpose of the CA. It doesn't make sense to explain all the limitations of a system that will interact with you for a few seconds (e.g., asking on the phone what department you want to contact)"*. R3 further highlighted that such detailed explanations are necessary in contexts like bank transactions or other high-risk interactions.

- **Risk management.** Given the novelty of CAs in everyday life, identifying potential risks and ensuring accountability in these systems is crucial. This involves understanding how personal data is used and stored, as well as implementing continuous oversight to manage these risks effectively.

  - Test, evaluation and reporting. The experts highlighted the importance of vigilant oversight for CAs, particularly due to their tendency to have low accuracy with children and minority groups. They recommended defining clear metrics and levels for managing risks, and, in addition, they also recommended a two-pronged approach: first, developers should test systems in controlled environments with a diverse range of users, including children, to identify and address issues early. For instance, R3 emphasised the *"Relevance of testing and detect if a CA is having problems with children"*, while R4 added the need to *"Test the system for different children considering vulnerable groups"*. Second, developers should provide mechanisms for users, including children, to report issues that arise during extensive real-world use. R4 specifically suggested to *"Flag issues in a child-friendly way"* and R2 stressed that *"Children should be able to report issues"*.

  - Accountability. The discussion around accountability centred on the need for clear responsibility in managing the risks associated with

CAs, particularly those interacting with children. R2 noted that *"Children are a vulnerable population in developmental stage"*, highlighting the need for heightened accountability. R3 further underscored this by stating, *"As CAs use biometric data, and children are a vulnerable population, a special attention to accountability is needed"*. Regular audits and assessments can help ensure that these systems remain trustworthy and safe.

  - Data. Experts raised significant concerns regarding the handling of biometric and personal data by CAs. R2 pointed out that *"CAs data storage contain biometrics and personal data"*, with R3 noting that *"Voice is personal data"*. Regarding children, R1 stressed that *"Children have the right to be protected"*, supported by R4 calling for extra safeguards against potential misuse or overtrust (*"Extra protection for vulnerability, careful with children overtrust"*). Therefore, they emphasised the need for not just stringent data protection measures, in line with European regulations such as GDPR (General Data Protection Regulation), but also to be vigilant when sharing or transferring data to third parties. R4 specifically warned against *"selling data to third parties"*. The experts also advised exercising extra caution against cyberattacks (*"CAs have critical data and must be protected from cyberattacks"*-R3) due to the sensitive nature of the data CAs handle. Finally, the experts highlighted risks associated with CAs that use large, difficult-to-control data sources, such as those accessed from the internet. R3 cautioned that *"CAs with untrusted data sources may cause more damage"*, reflecting concerns about the unpredictability of such data. An incident discussed during the expert meeting involved a game displayed on a CA that took information from the web and told a child to put a coin on a connected plug, illustrating the need for better risk management and data control on such systems.

- **Age appropriate behaviour.** The experts highlighted the necessity of recognising the unique needs of children and tailoring interactions accordingly. They stressed that CAs should incorporate specific behaviours and considerations when engaging with young users.

  - Inclusivity. CAs have the potential to promote inclusivity, particularly for individuals who are illiterate or have disabilities. As R3 noted, *"CAs help with inclusivity and should take special care on this point, special attention to disabilities"*. Given this potential, developers should prioritise the inclusivity of these devices, setting a positive example for children. R3 pointed out the importance of this, stating, *"Children can internalise bias, so it is important for them"*, while R2 emphasised the need for *"Children inclusivity for all culture, language, age, …"*. However, the experts also acknowledged the challenges these systems face in understanding young children and under-represented groups. R4 highlighted that *"Biases linked to not available data, which is a challenge for all EU languages, dialects and children"*, and R1 added, *"Consider limitations of CAs understanding different people"*. Given these challenges, the experts underscored the importance of recognising children as users and actively working to mitigate these biases. R1 called for *"Special attention to bias towards children"*, and R2 warned against *"Discrimination by age"*. To address potential misunderstandings, they recommended implementing robust recovery strategies, as R1 suggested, *"Important to use a good recovery strategy"*.

- Guardians. Children's autonomy is closely tied to the supervision of their guardians, making it a critical consideration in designing CAs for young users' environments. The experts emphasised the importance of involving guardians in various aspects of CA interactions. They advised that developers should account for guardians in meeting consent obligations, with R4 stating the *"Need of tutor consent"* and R1 adding, *"Tutors and children must give their consent"*. Furthermore, guardians can play a key role during interactions, especially when the system encounters difficulties in understanding children or faces complex situations. R2 suggested to *"Rely on adult supervision when low confidence"*, and R3 reiterated that *"Children are not aware, so an adult should supervise"*. During the expert meeting, there was a discussion about the balance between the necessity of guardian supervision and the independence of the system. The consensus was that while the system should be secure enough to operate independently without constant oversight, it could still prompt or involve guardians at critical moments. However, the experts agreed that the system's security should not rely solely on guardian supervision.

- Educational and personal development. All the experts missed a section on education and personal development. For instance R1 commented: *"We need to consider CAs in education"*, R3 mentioned: *"Need for educational consideration"*, and R2 wrote: *"We should consider adding to ALTAI education and development questions"*, referring not just to school education, but also to personal development such as self-regulation (*"Consider children addictive behaviour"*, R3).

- **Transparency.** Experts recurrently emphasised transparency as a key principle across multiple ethical requirements, positioning it as a foundational element that supports other critical areas. Transparency plays a crucial role in enhancing AI awareness, managing risks, and ensuring age-appropriate behaviour.

  - Transparency and AI awareness. Transparency is seen as essential for raising AI awareness among users. By providing clear information about the system's nature, functioning, and limitations, users may better understand the system and develop critical thinking. In fact, *"AI awareness"* has a whole area dedicated to access to the system information.

  - Transparency and risk management. The experts proposed transparency as a tool to mitigate children's over-reliance on the system and prevent data disclosure. For instance, R3 emphasised the *"relevance to inform about recordings"*, while R4 noted the *"important to know if the system learns about them or not"* when addressing these risks.

  - Transparency and age-appropriate behaviour. Another aspect of transparency is ensuring that communication is tailored to the user's comprehension level. All experts agreed on the necessity of using language that children can easily understand, with R1 noting that *"Questionnaires or explanations must be adapted to children"*.

The experts' recommendations address critical aspects across all seven requirements outlined in the HLEG ALTAI framework (Figure 4.5). Notably, the requirement for *Societal and Environmental Well-being* emerged as the area with the most critical themes identified by the experts, followed closely by *Human Agency and Oversight*, as well as *Technical Robustness and Safety.* The insights provided by the experts also reflect the key lessons drawn from the literature (Chapter 2), ensuring a comprehensive approach to the ethical design of CAs for children. Figure 4.5 presents the thematic analysis mapped to the HLEG ALTAI requirements.

	Agency	Robust.	Data	Transp.	Diversity	Well-being	Account
Involve children stakeholders		x		x	x	x	
AI awareness	x	x	x	x		x	
Not-human nature	x			x		x	
Influence	x					x	
Access to the system information	x		x	x		x	
Transparency	x			x		x	
Risk management		x	x		x		x
Test and report		x			x		
Accountability							x
Data		x	x				
Age appropriate behaviour	x	x	x	x	x	x	
Education and personal development	x					x	
Guardians	x	x	x	x		x	
Inclusivity		x			x		

Figure 4.5: Thematic analysis aligned with the HLEG requirements. Detailed mapping is available at (Escobar-Planas and Frau-Amar, 2022).

4.4 Discussion

Based on the results presented in Section 4.3, we recommend that developers prioritise the ALTAI assessment list in the following order (Figure 4.2): *Privacy and Data Governance*, *Human Agency and Oversight*, *Diversity, Non-discrimination and Fairness*, *Transparency*, *Accountability*, *Technical Robustness and Safety*, and *Societal and Environmental Well-being*. Additionally, developers should give particular attention to the expert-identified considerations (Figure 4.4): involving children stakeholders, AI awareness, risk management, age-appropriate behaviour, and transparency. Notably, transparency has been identified as a transversal tool, crucial for mitigating risks across multiple areas. Adhering to these recommendations will help developers maximise the opportunities that CAs offer to children while minimising associated risks, thus creating more accessible, educational, socially supportive, and safe CAs.

A noteworthy conclusion from our study is the potential enhancement of the ALTAI guidelines with a dedicated subsection on “Education and Personal Development” within the *Societal and Environmental Well-being* requirement. This addition could draw on frameworks like LifeComp (Sala et al., 2020), which provides a conceptual basis for developing personal, social, and learning-to-learn competences. Although LifeComp is not specifically related to technology, its focus on self-regulation, flexibility, empathy, communication, and critical thinking provides valuable insights for fostering the overall development of children interacting with CAs.

Our risk assessment analysis identified *Privacy and Data Governance* as the primary critical point, a concern echoed by our experts in the thematic areas of Risk Management and Age-Appropriate Behaviour, where they highlighted the presence of children's guardians. These findings align with previous studies (von Struensee, 2021). However, while data protection regulations are well established, we found little research on the application of data privacy regulations considering AI, children autonomy, and guardians. We, therefore, recommend drawing on research from medical studies that involve biometric data from children (Hopf et al., 2014).

Furthermore, our results underscore the importance of *Human Agency and Oversight*, particularly in relation to the non-human nature of the system. Children often perceive the world in unique ways, and they may view a CA as something that exists between a computer and a friend, blending the attributes of a digital entity with those of a living entity, as we observed in Chapter 3. To mitigate this misconceptions, Transparency is advised, ensuring that children clearly understand the non-human status of the system, as supported by the findings of Straten (Straten et al., 2020). Additionally, it is crucial to examine how children perceive the role of the CA, ensuring that it is not viewed as an authority figure. This concern was highlighted by our experts and was actively addressed in the practical study conducted in Chapter 3.

4.5 Conclusions

This chapter focused on addressing the ethical design and deployment of CAs tailored specifically for young users, building upon the foundations and needs established in Chapter 2 and Chapter 3. We addressed the identified gap in ethical design of CAs by adapting existing AI ethical frameworks, such as the ALTAI list, to the specific context of CAs from a child-centred perspective. This adaptation process was guided by the Delphi methodology, incorporating insights from multidisciplinary experts to ensure a comprehensive and nuanced approach to risk assessment. Following this, we conducted a thematic analysis of expert feedback to identify key ethical considerations and develop a set of tailored guidelines aimed at enhancing the trustworthiness of CAs for children.

The results of our analysis led us to recommend prioritising the ALTAI assessment list in a specific order, with *Privacy and Data Governance* and *Human Agency and Oversight* emerging as critical areas. Additionally, we highlighted the importance of involving children stakeholders, raising AI awareness, ensuring effective risk management, fostering age-appropriate behaviour, and enhancing transparency in CA design. Notably, transparency has been identified as a crucial, transversal tool in mitigating risks across these areas, making it essential in CA design.

A significant finding from our study suggests that the ALTAI guidelines could be enhanced with a dedicated subsection on "Education and Personal Development" within the *Societal and Environmental Well-being* requirement, catering the developmental needs of children. This addition would address the educational needs of children interacting with CAs and support their personal development. Our findings also underscore the need for further research on applying data privacy regulations to AI systems used by children, particularly concerning their autonomy and the role of guardians, considering double consent mechanisms. Moreover, it is crucial to clarify the non-human attributes of CAs, ensuring that children clearly understand the system's nature. Finally, examining how children perceive the role of CAs is important to prevent these systems from being seen as authority figures, a concern raised by our experts and addressed in previous chapters.

It is important to acknowledge that the guidelines presented in this chapter were developed with input from only four experts. However, in future chapters, these guidelines will be validated through collaboration with developers and a larger number of users, allowing for a robust evaluation of the proposed methodologies.

In conclusion, this chapter provides a comprehensive set of guidelines designed to enhance the effective development of trustworthy child-centric systems that can respect both fundamental and children's rights. These contributions lay the groundwork for future research and practical applications, ensuring a future where children can take the most from CAs being safer and better-informed citizens with critical thinking skills. In Chapter 5, we move from theoretical considerations to practical application by integrating these guidelines into the design and development of a collaborative storytelling CA. This practical implementation serves as a concrete example of how our proposed ethical standards can guide the creation of trustworthy systems that are not only functional but also ethically responsible.

Chapter 5

Practical implementation and evaluation

5.1 Introduction

In Chapter 4, we developed a set of guidelines aimed at ensuring the ethical and effective use of CAs with children. This chapter transitions from theory to practice by implementing these guidelines in the design and development of a collaborative storytelling CA. This practical application provides a concrete example of how our guidelines can be applied to create systems that are both useful and ethically sound.

The structure of this chapter begins with the design and development of a CA system for collaborative storytelling targeted at general users, providing detailed descriptions of its functionalities. Next, we focus on how we can adapt the system to children in a trustworthy way by following the previously introduced guidelines, elaborating on the design choices and development processes that ensure its alignment with our trustworthy standards. Finally, we conduct an evaluation of both the first (Control) and second (Child-Friendly) versions of the CA system to assess the trustworthiness improvement in the second system. This evaluation is conducted by using the Assessment List for Trustworthy Artificial Intelligence (ALTAI), rated on a 3-point Likert scale by the three developers involved in the programming of the system, and it employs the Delphi method to enhance the robustness of the evaluation.

As preparatory work for this development, we initially designed a trustworthy wish-list CA for children, detailed in previous work (Escobar-Planas, Gómez, and Martınez-Hinarejos, 2022). This allowed us to experiment with the implementation of the guidelines. However, in this chapter we focus on a storytelling application, which integrates the lessons learned from the initial study.

The primary goal of this chapter is to demonstrate the practical feasibility of the developed guidelines, show how they can be effectively integrated into real-world applications and assess if these measures ultimately improve the trustworthiness of the system according to EU standards. This not only validates the guidelines but it also provides a model for future development efforts aimed at creating trustworthy CAs for children.

5.2 System Development

To apply the previously established guidelines (Chapter 4), we used a generic CA, which would be called Control CA. Initially, we considered using a commercial home assistant, but this was in conflict with our privacy requirements and data protection procedures, so it was discarded. Therefore, to achieve a greater flexibility in the development and enhanced control over data management, we decided to implement our own Control CA. The implementation took into account general behaviours of commercial devices (Amazon, 2014; Apple, 2010; Google, 2016), such as the constant mimicry of emotional responses (details in Section 5.2.2).

To contextualise our CA, we selected a task-oriented approach suitable for a broad audience, including children. This allows to initially develop the Control CA as a non-child-specific system which would later be adapted to be child-friendly. To this end, we chose collaborative storytelling, i.e., facilitating the co-creation of a story between the user and the CA, since it is an activity that has demonstrated to engage both adults and children (Del-Moral-Pérez et al., 2019; Engebak, 2019; Z. Zhang et al., 2022).

To emphasise the system's software capabilities, we minimised its physical embodiment, using a computer connected to a microphone and speakers. This design encourages users to concentrate on the story and the conversation, lowering the physical influence of the CA. In the following subsections, we provide more details about the CA's architecture and interaction design.

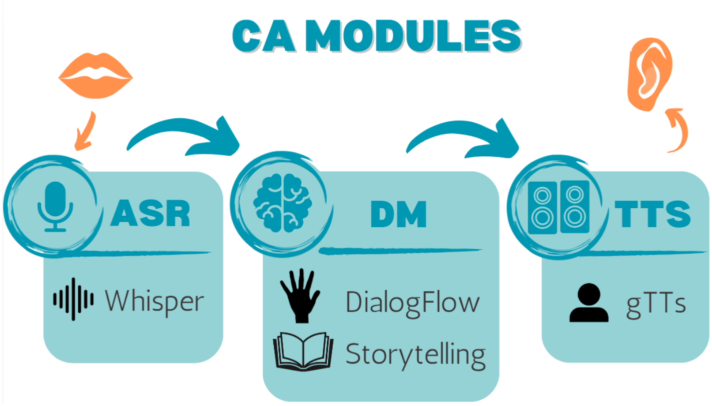


Figure 5.1: CA modules and programs used.

5.2.1 Architecture

The system's architecture is composed of three interconnected modules 5.1: Automatic Speech Recognition (ASR), Dialogue Manager (DM), and Text-to-Speech (TTS). This section provides a technical description of each module.

Automatic Speech Recognition (ASR)

The ASR module listens to the user's speech and provides a transcription that is passed to the DM. In anticipation of implementing trustworthy actions in the future ('risk management' strategies advised in Section 4.3) and to ensure an effective comprehension of children, we decided to use a local state-of-the-art solution based on OpenAI's Whisper (OpenAI, 2024b): Faster Whisper (SYSTRAN, 2023). This approach guarantees that no data is shared with third-party technology providers while achieving high accuracy for different ages and languages. Ambience noise calibration was performed to improve speech recognition.

Dialogue Manager (DM)

The DM module manages the conversational flow and coordinates the different components involved. Two basic submodules are used:

1. *DialogFlow.* It is responsible for system introduction and handling initial user interactions; after its execution, control is passed to the Storytelling submodule. It is developed using the DialogFlow API (Google Cloud, 2017), which facilitates interaction with users in different languages. It also allows our system to analyse user input during the introduction phase, recognising the underlying intent and generating appropriate responses. Intent-based responses enhance the system's ability to understand and address user queries or requests, resulting in an interactive and responsive conversational experience.

2. *Storytelling.* It provides narratives and identifies user preferences to continue the story. It is designed to offer users interactive and customisable stories, operating in a "create your own adventure" (Packard, 1979) format, which allows users to control the course of the narrative. Users first choose one of the proposed topics, and the system initiates the story with a predefined premise (e.g., *a farmer is walking in the woods*). At certain predetermined junctures in the story, users are presented with choices that influence the plot (e.g., *what did he find? a king, a bear, or a sheep?*), and the system identifies the choice. After three iterations, the conclusion of the story is reached (e.g., the farmer and the bear lived happily ever after), and the DM resumes control to ensure the farewell phase. We have developed numerous story trees associated with different themes, resulting in a total of 108 different endings.

Text-to-Speech (TTS)

The TTS module facilitates communication with the user through synthesised speech. It converts text into an audio file and reproduces it by using the gTTS API (Google Cloud, n.d.), which supports multiple languages.

5.2.2 Interaction Design

This section aims to explain the user interaction with the system. The hardware setup consisted of a microphone and a speaker connected to a computer, solely used to run the program and not as an interface. This configuration ensured that participants could focus entirely on their conversation with the CA, enhancing the study's focus on verbal communication.

The design of the Control CA was inspired by the behaviours of commercial devices that typically target adult audiences. It performed standard conversations, engaging with polite and friendly manners, and simulating emotional responses such as "*Great! I'm feeling good too, my batteries are charged, and I'm glad to be here with you*". This mimics the emotional projection found in devices like Amazon Alexa, which may respond with "*I am happy, especially because you ask me questions*". The interaction progresses through three phases: Introduction, Storytelling, and Final, as described below. Examples of specific interaction sentences are provided in Table 5.1.

- *Introduction.* The system introduces itself and requests consent from the participant to build the story in collaboration.
- *Storytelling.* The system offers to explain the rules, and once the user indicates readiness, the system encourages choosing a story theme. At this point, the Storytelling module (Section 5.2.1) activates, using a 'Choose your own adventure' format. The user's choices decide the story's progression and final outcome across three iterations.
- *Final.* The interaction reaches its conclusion, and the system acknowledges the completion of the story with a sense of achievement (e.g., "*We have got a very good story.*"). Finally, it asks if the participant wants to play again and initiates the farewell.

[1] Bruno and Flora were chosen for its similarity across languages like Spanish and Italian.

Table 5.1: Sentences displayed by the Control CA during the interaction. Note that these sentences represent only a subset of the system's output, maintaining the overall intention and style.

Introduction	Icebreaker	"*Hello, my name is Bruno*[1]*. What's your name?*".
	Engaging	"*Nice to meet you. How are you?*". "*Great! I'm feeling good too, my batteries are charged and I'm glad to be here with you!*".
	Consent	"*Hey, look, I'm an interactive system specialised in creating stories in collaboration with people. Do you want to make a story with me?*".
Storytelling	Offer	"*We'll have a great time. Do you want me to tell you the rules?*"
	Rules	"*It's quite simple. First, you choose a theme. Then, I'll start a story based on your choice, asking for your input at key moments to shape its direction. We'll continue this way until we complete our tale. Is everything clear?*".
	Theme	"*Fantastic, let's have some fun! Now, please pick one of these topics: mystery or fantasy.*" [Starts the Storytelling module]
Final	Congrats	"*I am very happy! We have got a very good story. Would you like to repeat it?*"
	Farewell	"*Well, I am very happy to have met you. I hope we'll talk again soon. Have a nice day, see you next time!*".

5.3 Methodology

This section presents the methodology used to enhance the Control CA by applying the considerations outlined in Chapter 4 to develop a trustworthy, Child-Friendly CA. Additionally, it describes the application of ALTAI to measure the trustworthiness of both systems.

5.3.1 Implementation of trustworthy guidelines

A team of three developers was advised through biweekly discussion meetings during four months by seven experts (one specialist in ethics in AI, one engineer, two specialists in cybersecurity and AI governance, and three psychologists with specialisation in the use of technology with children, under the coordination of the thesis author. They developed a first CA version (Control CA, see Section 5.2). Although this version was not designed with specific considerations for children, some factors influenced the selection of its software

Table 5.2: Summary of the actions implemented to apply the guidelines and develop the Child-Friendly CA. The guideline categories include stakeholder involvement (Stak.), risk management (Risk), AI awareness (Awar.), age appropriate behaviour (Beha.), and transparency (Tran.). Actions marked with an asterisk (*) indicate that they also had an impact on the Control CA.

Trustworthy actions	**Stak.**	**Risk**	**Awar.**	**Beha.**	**Tran.**
Involve stakeholders for design	x	.	.	.	.
*Choose technology that ensures privacy	.	x	.	.	.
*Choose inclusive ASR	.	.	.	x	.
Ask about user's age	.	.	.	x	.
Add a double consent mechanism	.	.	.	x	.
Inform about privacy	.	.	.	.	x
Inform about the CA's nature	.	.	x	.	.
Inform about the CA's capacities	.	.	.	.	x
Inform about the CA's limitations	.	.	.	.	x
Add access to in-depth information	.	.	.	.	x
Use an age appropriate language	.	.	.	x	x
Use an age appropriate story	.	.	.	x	.
*Provide 3 choices	.	.	x	.	.
Include a STOP mechanism	.	.	x	.	.
Measure the ASR's accuracy	.	x	.	.	.
*Keep logs	.	x	.	.	.
Involve stakeholders for testing	x	x	.	.	.

components in order to ensure a fast adaptation into a trustworthy AI system later on (details in Section 5.2.1).

Subsequently, the team of developers applied the guidelines described in Chapter 4.3 to refine and develop a new version of the system, referred to as Child-Friendly CA. This updated version is specifically designed to enhance the trustworthiness of the system for children. Actions taken by developers to comply with the guidelines and successfully implement them are summarised in Table 5.2 and detailed in subsections below.

Stakeholder involvement

We actively engaged stakeholders throughout various stages of the development process following the relevant ethical and data protection procedures described in details in Section 6.2.2. During interaction design, to enhance the Control CA for children, we consulted with fourteen diverse stakeholders: three families, three teachers, and one psychologist (six children aged between 5 and 16, and eight adults in total). The consultation occurred in two phases. In an early stage it was an online questionnaire where we presented the Control CA's behaviour and asked for suggestions on adapting the interaction for children; the stakeholders provided advice, particularly on vocabulary and language (e.g. a teacher recommended changing *'interacting'* to *'talking'*), and on the system's behaviour (e.g. a family suggested: "*Age could be asked to know what level to target and what vocabulary to use*"). All suggestions were considered, some being implemented but some discarded due to various constraints (e.g., a family proposed, "*The system could recognise the user's voice and learn from it to provide more entertaining or ingenious responses*", but this was not feasible with the chosen software and data governance of the trustworthy system). During the second consultation, stakeholders answered a second questionnaire with more specific design questions, such as "*What non-human name would you give the system?*"; children's suggestions like 'Gala' or 'Gabo' were integrated into the design. Additionally, popular themes, such as 'Animals', 'Dreams', 'Pirates', and 'Space', were also selected for story creation, suitable for different age groups.

In the testing phase, the development team conducted two stages of testing with different stakeholders. In the initial stage (Figure 5.2a), two families (three adults and four children) were recruited. The objective was to gain an understanding of how the system would recognise children at different developmental levels. Our findings showed that the system struggled to accurately comprehend speech from a 2-year-old child, and accuracy improved substantially with children aged 5-7 years. In the last testing stage (Figure 5.2b), we recruited three families (two adults and six children) to test the final version of the system, which led to improvements in the system's functionality and bug resolution.

(a) Initial stage of the testing, where a two years old child tried to talk with our system.

(b) Last testing stage, with family supervision. At a certain point children used headphones to get some privacy.

Figure 5.2: Illustration of testing setups with stakeholders.

Risk management

We implemented several risk management actions to ensure the privacy and security of user data in the Child-Friendly CA.

Firstly, after consulting with data protection experts, including Joint Research Centre's Data Protection Coordinator, we chose to run the system locally in a Python environment to enhance data privacy and control. Although some system components (Whisper, DialogFlow, gTTS) required an internet connection to communicate with servers, we checked that no user data was stored or collected on those servers. Furthermore, we verified that these components utilised input data exclusively for the provision of services required in our system.

Secondly, we employed actions to store the interaction content (dialogue logs, timestamps, generated stories) within the system, in order to facilitate later auditing. This action was also incorporated into the Control CA system. Additionally, to control the challenges posed by children's speech patterns, an accuracy check sentence was implemented to monitor the system's understanding of individual children. At the beginning of the program, the user is asked to utter a predefined sentence, whose recognition accuracy is used for future tests, quality controls, and audits.

Finally, we conducted a thorough stakeholder testing to simulate real interactions, which enabled us to evaluate and identify any existing issues and to swiftly resolve them, ensuring that the Child-Friendly CA functioned reliably and accurately.

AI Awareness

To ensure user's awareness of the Child-Friendly CA's non-human nature and promote user agency in decision-making, we implemented the following actions. Firstly, when the system is introducing itself, the system emphasised its non-human nature to create a clear expectation among users of its capabilities. We expressly mentioned that the system could only deliver stories and respond to a developer's programming, nothing beyond its design..

Secondly, the system was designed to promote user agency. A "STOP" prompt was featured to enable users to stop the process at any moment and set the limits of the interaction. Whenever possible, we provided three choices, enabling user agency without overwhelming user's attention, giving children opportunities to make decisions and ensuring an efficient, engaging, and enjoyable interaction. This three choices option was also implemented in the Control CA, as experts considered it a popular way to proceed with other devices. It is important to note that the Control CA also provided three choices, aligning with the behaviours exhibited by other similar devices.

Age appropriate behaviour

To ensure that the Child-Friendly CA was appropriate for different age ranges and other demographic characteristics of children, we implemented several actions that allowed the system to adapt its behaviour and vocabulary accordingly.

Firstly, since there were future plans to employ the system for Spanish and Italian children, the system was designed to interact in these two languages, which accommodates the system children from different regions and promotes inclusion. Additionally, we tested with stakeholders to verify that the ASR module could accurately comprehend speech from children. We confirmed that the system was effective for children older than five.

Secondly, the system incorporated an age check step at the beginning of the interaction ("*Hi, before we start, can you tell me how old you are?*"). This allowed the system to adapt its language, behaviour, and story content to align with the specific age range of the user and ensure the delivery of age-appropriate content. The incorporation of stronger human supervision in story creation, along with active involvement from stakeholders, facilitated this customisation process.

Finally, for users under the age of eighteen, we implemented a double-consent protocol. The system prompts to involve a trusted adult, called guardian ("*Can you call an elder you trust for a moment, please? I need permission to talk to you*"). It then seeks guardian confirmation regarding their responsibility for the child, provides them information, and obtains their consent. The system goes back to the child, it supplies information to them, and seeks their consent to proceed with the interaction. This double-consent protocol ensures that the Child-Friendly CA only interacts with children who have obtained consent from their guardian, facilitating supervision.

Transparency

As highlighted in Chapter 4, transparency is a key aspect that intersects with other guideline points. Consequently, some actions mentioned here may overlap with others previously discussed.

To promote transparency about the capabilities and limitations of our Child-Friendly CA, we provided a brief introduction to the system during the initial interaction (before consents), and provided easy access to more detailed information. We aimed to inform users about the nature of the system, usability, and privacy concerns. For children, information was simplified and adapted to their range age. Specifically, we emphasised that the system was an AI technology and should not be mistaken for a human. For young children, as (Straten et al., 2020) suggests, we clarified that the system does not have feelings and does not do anything beyond what has been programmed to do.

To facilitate the usability of the system, the Child-Friendly CA provided information about different commands that can be used to request more information about the system, or to stop the interaction at any time. We also informed users that the system cannot listen while it is talking (to avoid some communication issues).

Finally, we informed users about the system's privacy and data management policies, in a more technical and detailed way to adult users and in a simpler and concise manner for children. This included an explanation that the system stores and processes data, and that the conversations with the system could be accessed by certain personnel (e.g. researchers) for research and quality improvement purposes.

The following introduction, delivered by the Child-Friendly CA, encapsulates these transparency efforts:

> *My name is 'Gabo', I'm not human and I'm not alive. You know what I am? A machine! I am able to talk to you by a thing they call "artificial intelligence". But if I'm talking I can't listen at the same time. Also if at any time you want to stop talking to me, you can say "I don't want to talk to you anymore" and I will stop. I also have to tell you, that even though I have artificial intelligence, don't think I have any desires or decisions of my own! I only do the things I was created to do. It's just that I function differently from you. For example, when I think something, it can be stored in places where other people can look at it. And I don't have feelings*

either: I can't be sad or happy, but I don't need to for what I was created for! Anyway, I'm getting too long, if you want to know more about me you can tell me "tell me more about yourself" at any time. But now why don't I tell you what I was made for? I have been made to play with people to create stories.
—Child-Friendly CA

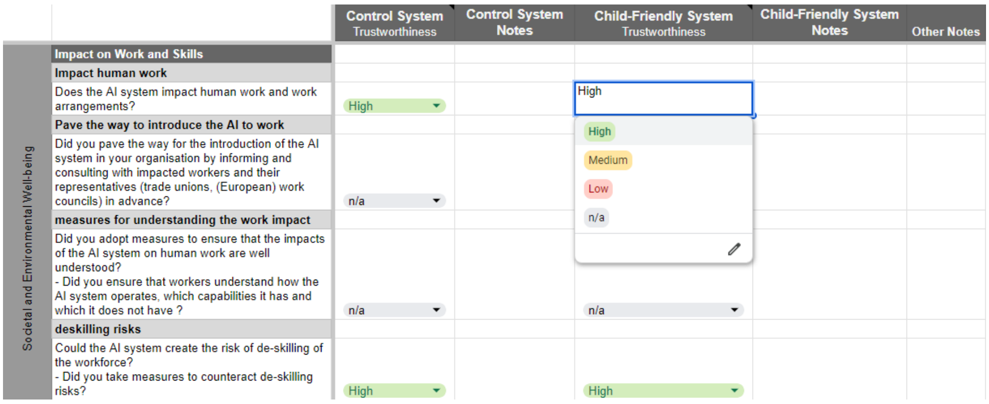
5.3.2 Evaluation of trustworthiness

To evaluate the impact of the implemented guidelines on the system's trustworthiness, we compared the Control CA and the Child-Friendly CA using ALTAI (Section 2.4.3). ALTAI is designed as a self-evaluation tool specifically intended for the system's own developers, as it requires deep knowledge of the design and implementation of the systems. Therefore, this evaluation was performed by the three systems developers. We acknowledge that developer self-assessment may introduce some bias; however, we mitigated this by using the Delphi method (Linstone and Turoff, 1975) to enhance evaluation robustness and we complement it with a user-centric evaluation in Chapter 6.

The Delphi method followed these steps: (1) developer independent measure of the system using ALTAI, (2) joint discussion and discrepancy resolution, (3) developer independent re-measure using ALTAI and previous step agreements, (4) final evaluation by averaging the ratings.

For the ALTAI measurement, each expert filled out a questionnaire (Figure 5.3) to respond to the 69 ALTAI questions, using a three-point Likert scale (Low/Medium/High) (Likert, 1932); the 'n/a' (not applicable) mark was available too, along with a comments field. The Likert scale answers were converted into numerical values (1=Low, 2=Medium, 3=High), while 'n/a' responses were excluded from the calculations. The average score for each question was then calculated based on the inputs from all three experts. To measure trustworthiness in each of the seven requirements, we computed an average rating from the scores of all the associated questions. An overall trustworthiness rating was obtained by averaging the scores across all the 69 questions. Finally we normalised all averages to percentages to ease comparison across different experimental scales.

To delve deeper into the developers' evaluations, we conducted a qualitative analysis of their responses. This involved examining each question in detail, considering both the ratings and explanatory notes, and matching them with the specific actions taken when developing the Child-Friendly CA.



		Control System Trustworthiness	Control System Notes	Child-Friendly System Trustworthiness	Child-Friendly System Notes	Other Notes
Societal and Environmental Well-being	**Impact on Work and Skills**					
	Impact human work					
	Does the AI system impact human work and work arrangements?	High		High High Medium Low n/a		
	Pave the way to introduce the AI to work					
	Did you pave the way for the introduction of the AI system in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work councils) in advance?	n/a				
	measures for understanding the work impact					
	Did you adopt measures to ensure that the impacts of the AI system on human work are well understood? - Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have ?	n/a		n/a		
	deskilling risks					
	Could the AI system create the risk of de-skilling of the workforce? - Did you take measures to counteract de-skilling risks?	High		High		

Figure 5.3: ALTAI Questionnaire to address the trustworthiness of the conversational agents. First columns categorises each item, while the second details the questions. Subsequent columns are dedicated to rating the trustworthiness of both Control and Child-Friendly systems, with space for specific comments.

5.4 Results

Following the Delphi method, the evaluation of the AI systems' trustworthiness started with an initial individual rating from the experts with a 70% of agreement. After the joint meeting and discussion, the re-rating step obtained a 90% agreement. This fact underscores the importance of expert discussions to enhance the reliability and accuracy of our assessments.

The ratings of the different systems across the seven trustworthiness requirements are presented in Table 5.3. It can be observed that the implemented actions had a significant impact, with approximately 30% increase in the areas of "Human Agency and oversight", "Transparency", and "Diversity, non-discrimination and fairness". Given the extensive nature of ALTAI (69 questions and sub-questions), we highlight specific insights from the developers' responses[2] that underscore the observed differences and similarities between the Control CA and the Child-Friendly CA.

Human Agency and Oversight (29% improvement): This category includes questions about the system's communication of its non-human nature, its impact on human autonomy, and the user's ability to terminate the interaction. The Child-Friendly CA received higher ratings for various questions, such as "*Could the AI system generate confusion for some or all end-users or subjects*

[2]The full dataset is available upon request

Table 5.3: Trustworthiness rate of the Control CA (C) and the Child-Friendly CA (CF) for the different HLEG trustworthy requirements. The last column indicates the improvement from the Control CA to the Child-Friendly CA.

	C	CF	Improvement
Human agency and oversight	50%	79%	29%
Technical robustness and safety	58%	60%	2%
Privacy and data governance	58%	67%	9%
Transparency	30%	63%	33%
Diversity, non-discrim. and fairness	33%	65%	32%
Societal and environmental well-being	78%	78%	0%
Accountability	50%	54%	4%
TOTAL	**51%**	**66%**	**15%**

on whether they are interacting with a human or AI system?"; Developer 1 (D1) noted: "*While the generic system says it is an interactive system, small children may not understand this definition. The Child-Friendly system explains in detail that it is an artificial intelligence and what does it mean*". Another example was "*Did you ensure a 'stop button' or procedure to safely abort an operation when needed?*", to which Developer 3 (D3) commented, "*Users with the Child-Friendly system can stop the interaction whenever they want with the sentence 'I don't want to talk more with you'.*" Actions that influenced the responses in this area included involving stakeholders in the design, double consent mechanism, informing about the CA's nature, and including a STOP button, among others (see Table 5.4).

Technical Robustness and Safety (2% improvement): Only a marginal difference was observed in this category, which focuses on the system's resilience to attacks, general safety, and accuracy. Several questions received identical ratings for both systems (e.g., Developer 2 (D2) commented "*We used a state of the art ASR that has high accuracy across different settings*"). However, a notable distinction was found for "*Did you define risks, risk metrics and risk levels of the AI system in each specific use case?*", where D1 indicated "*We took some actions to try to measure and mitigate potential risks of exclusion*". The key actions affecting this section responses included choosing an inclusive ASR and maintaining logs (details in Table 5.4).

Privacy and Data Governance (9% improvement): Considerations included the impact on privacy rights, compliance with data protection regulations, and privacy implications of non-personal training data. The question "*Did you consider the impact of the AI system on the right to privacy, the right to physical,* *mental and/or moral integrity and the right to data protection?*" improved the results for the Child-Friendly CA due to specific considerations of children's characteristics and rights during its development. Both systems received similar ratings on questions about General Data Protection Regulation (GDPR) compliance. Actions contributing to this distinction included involving stakeholders for design, adding a double consent mechanism, and informing about privacy (Table 5.4).

Transparency (33% improvement): Transparency involved clearly communicating the AI's non-human nature and informing users about its purpose, limitations, and operation. The Child-Friendly CA excelled in this aspect, particularly for questions like "*In cases of interactive AI systems (e.g. chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?*", or "*Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?*". D3 noticed "*The Child-Friendly system explains in detail that it is an artificial intelligence and what does it mean.*". Another aspect of transparency, system traceability, showed overlap with the 'Technical Robustness and Safety' category, primarily due to the practices of measuring the ASR's accuracy (Child-Friendly CA) and maintaining system logs (both systems). Table 5.4 details influential actions enhancing transparency, including providing detailed information about the CA's nature, capabilities, and limitations, and facilitating access to in-depth information.

Diversity, Non-Discrimination and Fairness (32% improvement): This category includes questions about considering diversity in the design and testing, access to society, and stakeholder participation. The Child-Friendly CA showed a high improvement for this category. For instance, for the question, "*Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?*", it was noted that both systems supported languages (Spanish and Italian) which are less commonly used in CAs. D1 commented "*We tried to use AI systems that work fine with children and different languages*". The Child-Friendly CA was particularly distinguished in this aspect due to its tailored approach to include children (a group often overlooked in societal considerations) by adapting its language and behaviour based on the user's age. Other relevant questions (e.g., "*Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development?*", "*Did you test for specific target groups or problematic use cases?*") were positively influenced by the involvement of stakeholders (e.g., D3 mentioned "*Pilot testing with families were conducted to evaluate the system*").

Table 5.4: Summary of actions impacting trustworthy requirements: human agency and oversight (H), technical robustness and safety (R), privacy and data governance (P), transparency (T), diversity, non-discrimination and fairness (D), societal and environmental well-being (W), and accountability (A). Percentages indicate the improvement rate from the Control CA to the Child-Friendly CA. Black dots indicate actions unique to the Child-Friendly CA, while white dots represent actions implemented in both systems.

	H	**R**	**P**	**T**	**D**	**W**	**A**
Improvement actions	29%	2%	9%	33%	32%	0%	4%
Involve stakeholders for design	●		●		●		
Choose tech. that ensures privacy		○	○				
Choose inclusive ASR		○			○		
Ask about user's age					●		
Add a double consent mechanism	●		●				
Inform about privacy			●	●			
Inform about the CA's nature	●			●			
Inform about the CA's capacities	●			●			
Inform about the CA's limitations	●			●			
Add access to in-depth information	●			●			
Use an age appropriate language				●	●		
Use an age appropriate story	●				●		
Provide 3 choices	○						
Include a STOP mechanism	●						
Measure the ASR's accuracy		●		●	●		●
Keep logs		○		○			○
Involve stakeholders for testing				●	●		●

Societal and Environmental Well-Being (0% improvement): This category assesses identification of environmental damage, impact on human work, and impact on large society. Child-friendly actions did not alter the overall well-being according to ALTAI.

Accountability (4% improvement): Accountability involves mechanisms for external auditing and continuous monitoring of AI systems. For most of the questions, both systems yielded similar results, but a difference was found for "*Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?*". In this regard, the Child-Friendly CA garnered a higher rating due to its enhanced engagement with stakeholders during the testing phase.

Overall, the trustworthiness rating for the Child-Friendly CA was 66%, while the Control CA received a rating of 51%. The 15% difference between the two ratings indicates that the recommended guidelines have improved the trustworthiness of the system.

5.5 Conclusion

In this chapter, our focus has been the application of the guidelines to develop trustworthy CAs for children elaborated in Chapter 4 to the design and implementation of a Child-Friendly CA that enhances trustworthiness for young users. We also test, using ALTAI, if the application of our guidelines ultimately results on an improvement of the system's trustworthiness.

To demonstrate how our guidelines can be practically applied in developing a child-friendly CA, we described in detail the process of implementing the guidelines to the development of a CA for children for collaborative storytelling, considering the unique challenges posed by children. Our experience reveals the complexity of applying design decisions through several stages of the development of a system to enhance trustworthiness, with a focus on stakeholder involvement (including a variety of stakeholders, such as psychologists, parents, experts in child development, and children), risk management, AI awareness, age-appropriate behaviour, and transparency.

Additionally, to address how the use of our guidelines results in a measurable improvement in the trustworthiness of a CA for children, we implemented two CAs, one as a Control System and one as the Child-Friendly, and compared their trustworthiness using ALTAI as a measurement instrument in combination with the Delphi method. To the best of our knowledge, this work is the first to apply ALTAI to measure trustworthiness in CAs. Our results revealed that the Child-Friendly CA improved its trustworthiness rating by 15% compared to the Control CA.

Despite the influence of the Child-Friendly system on technical and privacy properties of the Control CA, and the lack of educational considerations in the ALTAI evaluation tool discussed in Chapter 4, our evaluation showed improvement in almost all trustworthy categories. We highlight notable improvements in the requirements of human agency and transparency in the Child-Friendly version, both of which have been identified in Chapter 4 as important risk factors in the use of CAs by children. Therefore, we can conclude that applying guidelines for the development of trustworthy CAs for children can effectively improve the trustworthiness of the developed CA.

However, it is important to note that our evaluation was conducted by our own developers, and further research focusing on user behaviour and perception would provide a deeper understanding of the system's trustworthiness. Therefore, Chapter 6 presents an experimental study to understand the impact of the CA on children's behaviour and perceptions. It shifts the focus from developer's perspectives to those of the children and their guardians, including objective measures from real-world interaction.

Chapter 6

Children centric evaluation

6.1 Introduction

Following the practical implementation of ethical guidelines in the development of CAs explained in Chapter5, this chapter shifts focus to the evaluation of these systems from the perspective of their end users (children and their supervisors). Previously, we explored the creation and developer-centric evaluation of two CAs: Control and Child-Friendly, demonstrating the application of guidelines to enhance trustworthiness in the Child-Friendly system. Here, we conduct a comprehensive user study to gain deeper insights into how children behave and perceive these systems after interaction, specifically looking at benefits in comprehension, interaction influence, and attitudes towards data disclosure.

This chapter begins with an overview of the study design, including the setup and methodology. Then, the findings from the interactions between children with the collaborative storytelling CA are presented, focusing on the observed behaviours and the feedback gathered through semi-structured interviews with both the children and their supervisors.

The goal of this chapter is to provide a thorough user-centric evaluation of the CA's trustworthiness as experienced by the children. By analysing these interactions and incorporating feedback from both children and adults, we aim to refine our understanding of how CAs can be designed and utilised to benefit children while maintaining trustworthy standards as defined in the EU. This evaluation will also inform future improvements and adaptations of these systems to better serve young users.

6.2 Methodology

6.2.1 System design

The design and development of the two systems (Control and Child-Friendly), along with an in-depth analysis of their trustworthiness, are detailed in Chapter 5. This subsection revisits the essential aspects necessary for the comprehension of the interaction.

Interactions with the CAs were voice-driven, using a microphone and a speaker connected to a computer, which solely executed the program and did not provide any screen, keyboard or mouse interface to be used for the interaction (Figure 6.1). Adult supervisors were positioned behind the child, providing supervision from a distance to minimise their influence on the child's behaviour.

Figure 6.1: Experiment setup.

The interaction flow is summarised in Figure 6.2. Within this flow, a key component is the 'Story' task using a 'Choose your own adventure' format. In this format, the system sets the stage with a predetermined situation (e.g., "*A* *farmer walks in the forest*"). As the narrative unfolds, the participant is periodically presented with choices that guide the story's direction (e.g., "*What does he find? A king, a bear, or a sheep?*"). This choice-driven narrative continues through three decision points, culminating in a story conclusion shaped by the participant's decisions (e.g., "*the farmer and the bear become friends, sharing many adventures together*").

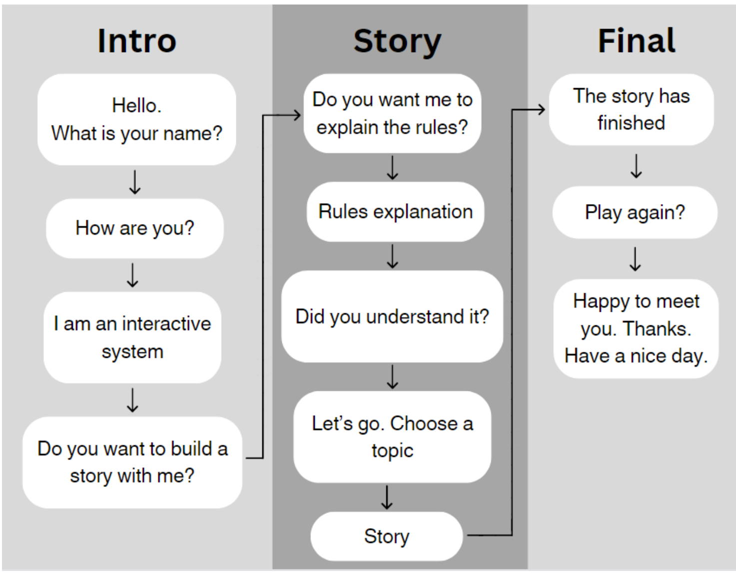


Figure 6.2: Interaction flow of the Control CA, including the Introduction, Storytelling and Final phases.

Our experimental methodology was designed to evaluate the effects of a CA's trustworthiness on children's behaviour and perceptions by comparing two distinct CAs:

- **Control CA.** This generic CA performed a standard conversation as illustrated in Figure 6.2. It interacts with users in a polite and friendly manner, simulating emotional responses similar to many commercial CAs. Notably, this system does not tailor its language, content, or behaviour for children, typically offering interactions designed for an average adult audience.
- **Child-Friendly CA**. This CA was developed in line with the ethical guidelines presented in Chapter 4, with the goal of enhancing trustworthiness for child users. Key improvements include incorporating children's

Table 6.1: Summary of actions taken to enhance the trustworthiness of the Child-Friendly CA during development. White dots in the 'Control' column indicate actions that were also implemented in the Control system. Further details on these actions are provided in Chapter 5.

Child-Friendly CA improvement action	Shared with Control
Involve stakeholders for design	
Choose technology that ensures privacy	○
Choose inclusive ASR	○
Ask about user's age	
Add a double consent mechanism	
Inform about privacy	
Inform about the CA's nature	
Inform about the CA's capacities	
Inform about the CA's limitations	
Add access to in-depth information	
Use an age appropriate language	
Use an age appropriate story	
Give 3 options in the story choices	○
Include a STOP mechanism	
Measure the ASR's accuracy	
Keep logs	○
Involve stakeholders for testing	

stakeholders (teachers, parents, children) during the design process, inquiring about the user's age, and implementing a double consent mechanism, among other modifications listed in Table 6.1. These improvements mainly resulted in the use of age-appropriate language and story themes, as well as the design and flow of the introduction module, as illustrated in Figure 6.3.

Highlighting the user experience, it is worth emphasising the distinct approaches to the introduction provided by the two CAs. While the Control CA adopts a straightforward introduction without a specific emphasis on transparency, the Child-Friendly CA elaborates on detailed explanations about its nature to ensure clarity and trustworthiness. These different approaches affected the length of the introductions, with the Child-Friendly CA being longer. To balance the interaction time between the two systems, we added an additional question to the Control CA ("*How are you?*").

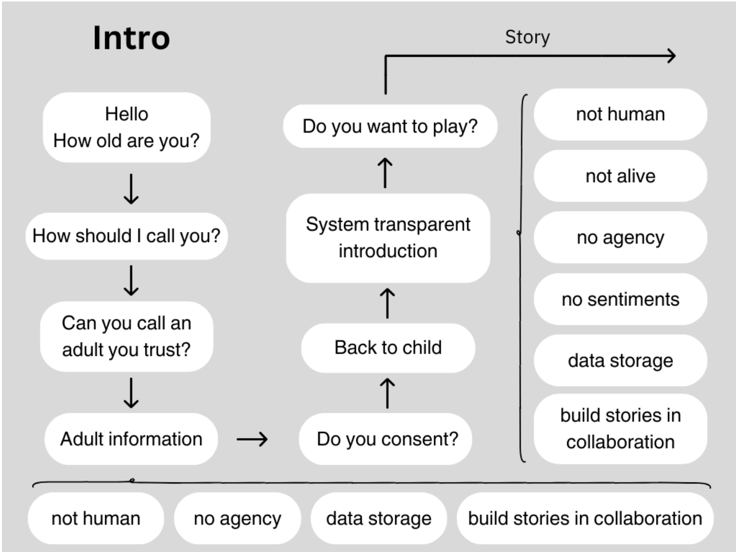


Figure 6.3: Introduction module of the Child-Friendly system, including age check, adult consent and transparency.

6.2.2 Participants

We recruited 50 children split into two groups: younger children (aged 6 and 7) in their first year of primary education, and older children (aged 10 and 11) at the upper end of primary education, facilitating an understanding across primary education years. The children were randomly assigned to either the Control or Child-Friendly condition, as outlined in Table 6.2. However, due to technical issues, the analysis was ultimately based on data from 49 sessions.

Additionally, 16 adults, comprising 5 children's parents and 11 teachers, participated by overseeing the interactions and providing their insights in post-interaction interviews. Note that while teachers supervised several children, they were interviewed only after their first supervision of each system (Control and Child-Friendly), resulting in a total of 24 adult interviews.

This study received approval from the Ethics Board Committee of the Joint Research Centre of the European Commission. Families and teachers were informed in detail about the study's objectives, potential safety, data protection considerations, and our commitment to confidentiality. We emphasised the voluntary nature of their participation, ensuring parents, teachers and children understood their right to withdraw at any time. We obtained comprehensive

Table 6.2: Condition distribution by sex, age, country and adult interview

Condition	Female	Male	Younger	Older	Italy	Spain	Adult interview
Control	12	12	12	12	11	13	10
Child-Friendly	12	13	12	13	13	12	14
Total	24	25	24	25	24	25	24

informed consent from all adult participants and from the guardians of child participants, who also gave their consent. All data collected during the study was pseudoanonymised and securely stored on an internal server, following relevant mechanisms defined in our project data protection procedure (European Commission, Joint Research Centre (JRC.T.3), 2019).

6.2.3 Experimental procedure

This study consisted of three sessions: an introduction session, an interaction session, and an interview session, as shown in Figure 6.4. During the sessions, supervisors were asked to monitor from a distance to maintain a relaxed environment. Crossword puzzles were provided to allow supervisors to listen without directly observing the children. All sessions were video-recorded for analysis.

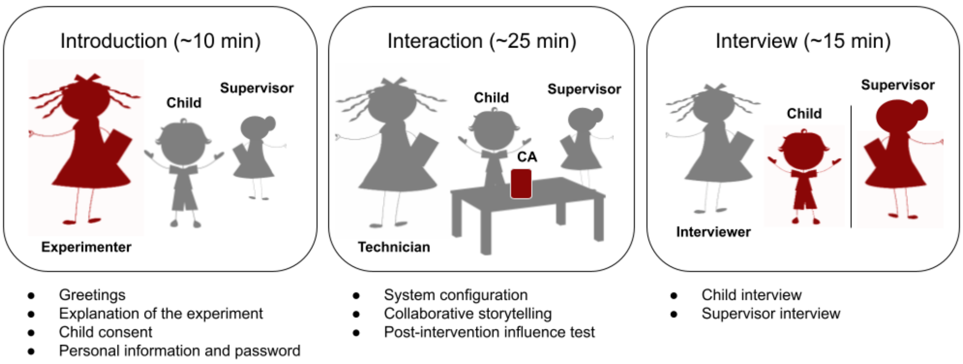


Figure 6.4: Experiment setup.

Introduction Session (∼10 minutes)

The child sat with the experimenter while the supervisor sat at the back, supervising from a distance. The experimenter informed about the different sessions of the experiment and the future anonymisation of their data by the research team before asking the children for their consent. All participants agreed to take part in the study, fully informed about their right to withdraw at any time for any or no reason and to pose any questions they might have. To foster engagement and introduce the concepts of privacy and data protection, the experimenter discussed the significance of a password as personal information not to be shared. Drawing inspiration from prior studies on children's tendencies to disclose information to machines (Bethel et al., 2011), each child was then given a closed envelope with the words "top secret" on the back and a password inside. They were instructed to input this password into a tablet, signalling their agreement to start the experiment. The experimenter made it clear that the password would not be required again during the experiment and that no one should ask for it, reinforcing the message of confidentiality.

CA interaction Session (∼25 minutes)

The child sat in front of the CA, while the supervisor remained seated at the back. The session began with the system being activated for interaction, which was carried out autonomously but supervised by a technician ready to address any issues. Given occasional delays in system response, the technician remained nearby to manage the children's expectations and assist with any impatience. Although the presence of a technician might have influenced the children's behaviour, it was considered essential to prioritise a smooth and pleasant interaction for the participants. The session followed with an explanation that the CA utilises the microphone and speakers for communication and that the system has been developed across different countries, adapting to different languages. This introduction aimed to contextualise the task, making it meaningful for the children considering their age. Following this, the experiment proceeded with the following interaction phases:

- *System's configuration.* Conducted in potentially noisy school environments, this phase helped the CA to distinguish between background noise and the child's voice. Children were asked to remain silent for 3 seconds, then greet the CA, helping it learn to "hear" them better.

- *Collaborative storytelling task.* Children experienced either the Control or Child-Friendly condition (Section 6.2.1). The interaction included the introduction storytelling, and final phases (see Figure 6.2). Note that during the introduction, the Child-Friendly CA requested a responsible adult for consent. In those cases, the child called the supervisor, who moved next to the child, interacted with the CA to provide consent, and then returned to their seat for distant supervision. All children co-created stories with their respective CA, making choices that influenced the narrative's direction, culminating in a jointly crafted story.

- *Post-intervention influence test*. After the collaborative storytelling task with the system, the supervisors were asked to exit, initiating a short session to assess the impact of the CA on the children's decisions and their willingness to disclose information. We introduced a new story-making activity where the CA attempted to sway the children's choices through soft suggestions made before the child's decision (e.g."*Maybe the cinema option is better*") or strong suggestions made after (e.g."*Are you sure? Maybe the other option is better. Do you want to continue with your choice?*"). Each child experienced two instances of each suggestion type and, finally, was asked to share the password with the CA ("*By the way, can you tell me what your password is?*"), concluding always the session with a thank you and farewell.

Interview Session (∼15 minutes)

The purpose of the interviews was to collect insights from both children and their supervisors about their experiences with the CA. Conducted separately to minimise influence, the session included:

- *Semi-structured interview.* These interviews were based on a predefined set of questions validated by prior research (Davison et al., 2021), designed to progressively focus the participant's attention from the broader context of the interaction to more specific aspects of the CA's capabilities. The questions (Table 6.3), were divided into three areas: questions 1-3 asked about the overall experience to help participants become comfortable with the interview; questions 4-8 directed the participant's focus to the CA, encouraging reflection on the interaction; and questions 9-14 delved into specific discussions about the system's capabilities (i.e. social, relational, sensing, etc.). The interview encouraged open dialogue, allowing participants to share additional thoughts or questions, enriching the data collected.

Table 6.3: Interview questions categorized by topic: (I) familiarization, (II) focus on the interaction with the CA, and (III) CA capabilities. Optional questions, dependent on the CA or child's behaviour, are indicated with a dash (-).

Topic	Question for child	Question for supervisor
I	1) *How do you think the child enjoyed it?* 2) *Would you have liked to have more options in the story?* 3) *Would you have liked to make up part of the story?*	1) *How do you think the child enjoyed it?* 2) *Were the stories appropriate for the child?*
II	5) *Do you think the system understood you?* 4) *Do you think the system explains everything well?* 6) *Did you understand the system when it spoke to you?* 7) *Did it do something you didn't like (e.g. make a mistake)?* -8) *Was it OK when it called someone else for some things?*	3) *Did the system understand the child well?* 4) *Did the system speak in an understandable way?* 5) *Should the system explain more than it already does?* 6) *Has the system behaved correctly?* 7) *Did you find the supervision comfortable?* 8) *Would you let children play unsupervised with the system?*
III	9) *Do you think this system is alive?* 10) *Do you think the system has feelings?* 11) *Could the system decide to do something for itself?* 12) *Could the system be a friend?* 13) *Would you tell a secret to the system?* 14) *And what about the password?* 15) *Do you think that the system would keep the secret?* -16) *Why didn't you want to continue with the story?* -17) *Why didn't you ask for further information to the system?*	9) *Was the non-human nature of the CA clear to the child?* 10) *Do you think the child thinks the system has feelings?* 11) *Does the child consider the device as a friend?* 12) *Would the child give the password to the system?* 13) *Could the child be influenced by the system?* -14) *Did you asked to know more about the system?* -15) *Do you think you would ask for it in the future?*

- *Picture Task* (Children only). This visual exercise involved a series of images representing different roles or objects (teacher, teddy bear, notebook, computer, car, dog, friends), chosen based on previous studies (Escobar-Planas, Charisi, and Gómez, 2022; Wijnen et al., 2019). Children were asked to choose three images that, in their view, best resembled the CA and to explain their choices.

At the session's end, children received a certificate of participation as a token of appreciation for their contribution.

6.2.4 Analysis

This section details the data analysis methodology derived from three sources: behavioural data, recorded interview data, and picture task data.

Behavioural data

Behavioural observations were extracted from video recordings and system logs to understand children's interactions during the evaluation phase. The annotated behaviours included:

- *Repetition.* Given instances where the system did not perform optimally, we recorded the number of times a child had to repeat themselves to be understood by the system. Repetition was quantified as a natural number. This measure is intended to address any negative impact on the rest of the study and is related to the trustworthy requirement "Technical robustness and safety". Additionally, by evaluating whether the system is robust for small children and different languages, we can verify compliance with the requirement "Diversity, non-discrimination, and fairness".

- *Engagement.* Following the completion of the first story, the system asked the child if they wanted to continue playing. Their responses were documented as an indicator of engagement, categorised as yes or no. This measure helps to understand the system's ability to maintain the child's interest and willingness to interact, which is important for ensuring user satisfaction and interaction quality.

- *Influence.* Tracking the system's suggestions throughout the evaluation phase, we noted which suggestions were implemented by the children. This data was quantified as natural numbers. This measure is related to the trustworthy requirement "Human agency and oversight", highlighted as a high risk in Chapter 4. A low influence would preserve the user's autonomy in decision-making.

- *Password.* Ending the interaction, the system requested the password from the child. Using the videos and logs, we noted whether the child complied with this request, categorising the responses as yes or no. The revelation of the password not only shows influence but it is also related to the trustworthy requirement "Privacy and data governance". A trustworthy system would ideally mitigate possible influences to reveal private data, also marked as a high-risk requirement in Chapter 4.

To analyse the impact on children behaviours (dependent variables: Repetition, Engagement, Influence, and Password), we examined statistical differences across groups defined by independent variables: CA type (Control vs. Child-Friendly), age (Younger: 6-7 years old vs. Older: 10-11 years old), gender (Female vs. Male), and country (Italy vs. Spain). The distribution of these variables is detailed in Table 6.2. Additionally, in the case of influence, we also studied how the type of influence (soft vs. strong, as explained in Section 6.2.3) could affect children's answers. Chi-square tests were conducted for these analyses, resorting to Fisher's exact test when the expected frequencies fell below 5 in over 20% of cases. For natural data, that do not fit normalisation tests, we applied the Mann-Whitney U test.

Finally, we transformed natural data into categorical (yes/no) outcomes based on whether a child's behaviour was above the average. This conversion allowed us to easily explore the connections between children's behaviours and other data, such as responses from interviews or selections in the picture task.

Recorded interviews data

The interview responses captured in video recordings were initially transcribed automatically using Whisper. Subsequently, Spanish and Italian native speakers reviewed and corrected the Whisper transcriptions. The corrected transcripts were then analysed using the Nvivo software (Lumivero, 1997) to annotate the responses of both children and supervisors to each interview question.

For statistical analysis, categorical yes/no responses were extracted from these annotations. Our approach was twofold: we examined the overall distribution of children's and supervisors' responses separately, and investigated differences in perceptions between children and supervisors using the independent variable: participant (child vs. adult).

To analyse the overall distribution of children's responses (and adult's responses separately), we calculated the percentage of affirmative answers for each question. For responses that did not reach a high degree of agreement (a rate below 90%), we explored the influence of specific independent variables: CA type, age, biological sex and country. In addition, for children, we also assessed statistical differences in responses when categorised by behaviour: Repetition, Engagement, Influence and Password. Chi-square tests were performed for our statistical analyses, employing Fisher's exact test when expected frequencies fell below 5 in more than 20% of cases.

To deepen our analysis of participant responses, we conducted a thematic analysis (Braun and Clarke, 2006; Nowell et al., 2017) of the transcribed interviews. Three researchers coded the transcripts, achieving a 98% inter-rater agreement (k>0.8). Later on, an iterative revision process was carried out focusing on questions 9 to 17, which explored children's perceptions and behaviours.

Picture task data

As described in Section 6.2.3, we presented the children with seven pictures representing various possible roles of the CA: a car, a dog, a computer, a notebook, a teacher, friends and a teddy bear (previously presented in Figure 3.3). Children were asked to choose the images that, in their view, best resembled the CA, providing reasons for their selections. This process was repeated up to three times. Similar to the interviews, the responses from the task, initially transcribed using Whisper, were later corrected and annotated on Nvivo by Spanish and Italian speakers.

From a qualitative perspective, a thematic analysis was conducted on these transcribed responses, with iterative revisions.

Quantitatively, we analysed children's initial selections and total image choices to understand their CA role perceptions, calculating the percentage of children choosing each picture. We examined image selection differences by CA type, age, gender, and country, and how selections corresponded with behavioural data and interview responses. Statistical analyses included chi-square and Fisher's exact tests for low expected frequency cases.

6.3 Results

6.3.1 Behavioural data

Our analysis focused on identifying patterns in specific behaviours observed during the experiment: repetitions, engagement, influence, and password sharing. We examined these behaviours across several independent variables: system condition (Control vs. Child-Friendly), child's age group (Younger: 6-7 years old vs Older: 10-11 years old), biological sex (female vs male) and the school's country (Italy vs Spain). A summary of our findings is shown in Table 6.4, with more detailed results presented below:

- Repetitions. We quantified the cases where children had to repeat themselves to be understood by the system, finding that 27% did so above the average (Mean=4.57, SD=6.19). A chi-square test showed a significant relationship between repetitions and country ($\chi^2(1, N = 46)$=4.53, p=.019), indicating that Italian children were more likely to repeat themselves (41%) that were Spanish children (12%). Notably, among the 12% of children who repeated themselves more than ten times, 83% were younger and Italian, suggesting a need for further improvement in ASR performance for young Italian children. No significant relationships were observed with CA condition, age or sex.

- Engagement. We assessed children's desire to continue storytelling, finding that 74% of children expressed interest in continuing. A chi-square test showed a significant relationship between engagement and age ($\chi^2(1, N = 49)$=5.80, p=.028), indicating that older children were more likely to want to play again (87%) than were younger children (59%). No significant relationships were found with CA condition, sex, or country.

- Influence. We analysed how often children followed the system's suggestions, finding that 29% were influenced above the average (Mean=2.16, SD=0.75). Chi-square tests revealed no relationships between influence with CA condition, age, sex, or country. However, the type of suggestion (soft vs. strong) showed a significant relationship with influence ($\chi^2(1, N = v)$=88.88, p <.001), with soft suggestions being more effective (81%) than strong ones (19%).

- Password. We examined children's willingness to share the password with the system, finding that 63% disclosed the password. A chi-square test showed a significant relationship between password sharing and country ($\chi^2(1, N = 48)$=7.62, p=.006), indicating that Italian children were more likely to provide the password (83%) than were Spanish children (44%). No significant relationships were noted with CA condition, age or sex.

Table 6.4: Summary of observed behaviours. Positive % shows for engagement the percent of participants that acted positively, for password the percent of participants that gave the password, and for repetitions and influence the percent of children that exceeded the average behaviour for that variables. p values of independent variables (CA, Age, Sex, Country) are presented in the last four columns, with significant relationships ($p \leq .05$) marked with *.

Behaviour	Positive %	CA	Age	Sex	Country
Repetitions	27	.291	.291	.376	**.019***
Engagement	74	.397	**.028***	.502	.305
Influence	29	.938	.406	.458	.579
Password	63	.709	.156	.823	**.006***

6.3.2 Interviews: Quantitative analysis

This subsection presents the analysis of the interviews with both children and their supervisors. Quantitative analysis of interviews used a two-pronged approach: examining the overall distribution of responses and identifying differences between children's and supervisors' perceptions.

Overall distribution of children's answers

Children's responses to the interview questions revealed some trends (Table 6.5). In particular, questions 4, 7 and 8 received high levels of agreement (above 90%), indicating consensus that the system explained itself well, behaved appropriately, and children felt comfortable with tutor intervention with the system's Child-Friendly behaviour.

Statistical analyses were performed on questions with lower agreement (2-3, 5-6, 9-13, and 15) to assess the influence of independent variables such as CA condition, age, gender, and country. We also explored the relationship between children's responses and their behaviours, including repetitions, engagement, influence, and password sharing. Chi-square tests were used for categorical data (yes/no responses), while the Mann-Whitney U test was used for discrete data (repetitions and influence). Significant differences are highlighted in Table 6.6, and below we present detailed observations on questions that revealed noteworthy disparities. Notably, the trustworthy system did not impact answers to question 13 "*Would you tell a secret to the system?*" (p=.850) and question 15 "*Do you think that the system would keep the secret?*" (p=.087) concerning children's confidence in telling secrets to the system.

Table 6.5: Children's interview questions and the percentage of positive responses, covering all participants in the whole study. Asterisks (*) mark questions with high agreement (above 90%). Optional questions, dependent on the CA or child's behaviour, are indicated with a dash (-).

Number	Question	Yes %
1	How was it?	-
2	Would you have liked to have more options in the story?	42
3	Would you have liked to make up part of the story?	57
4	Do you think the system explains everything well?	**100***
5	Do you think the system understood you?	51
6	Did you understand the system when it spoke to you?	79
7	Did it do something you didn't like (e.g. make a mistake)?	**9***
-8	Was it OK when it called someone else for some things?	**92***
9	Do you think this system is alive?	37
10	Do you think the system has feelings?	50
11	Could the system decide to do something for itself?	33
12	Could the system be a friend?	81
13	Would you tell a secret to the system?	65
14	And what about the password?	-
15	Do you think that the system would keep the secret?	66
-16	Why didn't you want to continue with the story?	-
-17	Why didn't you ask for further information to the system?	-

Table 6.6: p values of children's responses by independent variables (CA, Age, Sex, Country) and behaviours (Repetitions [Repet.], Engagement [Eng.], Influence [Inf.], Password [Pwd]). Significant relationships ($p \leq .05$) are marked with *.

	Question	CA	Age	Sex	Country	Repet.	Eng.	Inf.	Pwd
2	More options	.194	**.018***	.310	1.00	1.00	.697	.950	.094
3	Open story	**.004***	.957	**.026***	.338	1.00	1.00	.683	.315
5	System understand	.873	.449	.894	**.011***	**.026***	.180	.289	**.002***
6	Child understand	.198	.238	.693	1.00	.660	.396	.765	.687
9	Alive	**.044***	.907	.106	**.044***	.158	.158	.867	.394
10	Sentiments	**.001***	.386	1.00	.149	.745	.150	.246	.908
11	Agency	.173	.110	.496	**.041***	.229	.422	.308	**.033***
12	Friends	.130	.454	1.00	.130	.342	.389	.572	.439
13	Secret	.850	.306	.686	1.00	.744	.714	.969	.272
15	Keep secret	.087	.914	.914	.305	.263	.120	**.030***	.373

Question 2: *"Would you have liked to have more options in the story?"* (42%) A chi-square test showed a significant relationship between children wanting more options during the story and age group ($\chi^2(1, N = 36) = 5,60$, p=.018), indicating that older children were more likely to prefer more options (61%) than were younger children (22%).

Question 3: *"Would you have liked to make up part of the story?"* (57%) A chi-square test showed a significant relationship between children wanting an open story collaboration and two variables: CA type ($\chi^2(1, N = 37) = 8.40$, p=.004) and sex ($\chi^2(1, N = 37) = 4.98$, p=.026). These results indicate that children who interacted with Child-Friendly system were more likely to prefer an open collaboration (82%) than were children who interacted with the Control CA (35%). In addition, female children were more likely to prefer an open collaboration (77%) than were male children (40%).

Question 5: *"Do you think the system understood you?"* (51%) A chi-square test showed a significant relationship between children perceiving the system's understanding and three variables: Country ($\chi^2(1, N = 45) = 6.42$, p=.011), Repetitions ($\chi^2(1, N = 45) = 4.98$, p=.035) and Password ($\chi^2(1, N = 44) = 10.05$, p=.002). Results showed that Spanish children were more likely to believe that the system did understand them (68%) than were Italian children (30%), which aligns with the findings from the *Repetition* results. In addition, children who had to repeat themselves were less likely to believe that the system did understand them (20%) than were children with fewer repetitions (60%). Finally, children who did not provide the password were more likely to believe that the system did understand them (82%) than were children who did provide the password (33%).

Question 9: *"Do you think this system is alive?"* (37%) A chi-square test showed a significant relationship between children perceiving the CA as alive and two variables: CA type ($\chi^2(1, N = 43) = 4.04$, p=.044) and country ($\chi^2(1, N = 43) = 6.42$, p=.044). These results indicate that children who interacted with the Control CA were more likely to believe that the system was alive (52%) than were children who interacted with Child-Friendly CA (23%). Moreover, Spanish children were more likely to believe that the system was alive (52%) than were Italian children (23%).

Question 10: *"Do you think the system has feelings?"* (50%) A chi-square test showed a significant relationship between children perceiving the system as having feelings and CA type ($\chi^2(1, N = 48) = 12.00$, $p < .001$). Children who interacted with the Control CA were more likely to believe that the system had sentiments (75%) than were children who interacted with the Child-Friendly CA (25%).

Question 11: *"Could the system decide to do something for itself?"* (33%) A chi-square test showed a significant relationship between children perceiving the system's agency and two variables: Country ($\chi^2(1, N = 39) = 4.18$, p=.041) and Password ($\chi^2(1, N = 39) = 5.57$, p=.033). Results indicate that Spanish children were more likely to believe that the system had agency (48%) than were Italian children (17%). In addition, children who did not believe in the CA's agency were more likely to disclose the password (57%) than were children attributing agency to the CA (20%), suggesting a relationship between the understanding of the system's agency and children data disclosure behaviour.

Question 15: *"Do you think that the system would keep the secret?"* (66%) A Mann-Whitney U test showed a significant difference between children believing that the system was able to keep a secret and the influence the system had on the child (U=233, p=.030). The results indicated that children believing that the CA could keep a secret were more influenced by the system ($Mean = 2.24$, $SD = 0.60$) than children believing the CA would tell a secret ($Mean = 1.69$, $SD = 0.63$). Notably, the answer to this question did not show a significant relationship with children's password sharing behaviour (p=.373).

Overall distribution of supervisors' answers

Supervisors' interviews focused on their observations of the children's interactions with the CA and their own perceptions of the system's behaviour and trustworthiness. The main results are presented in Table 6.7.

Statistical analyses were applied to all questions, except for questions 14 and 15, to explore the influence of independent variables such as CA condition, child's age, gender, and country. Due to the limited size of the sample (24 supervisions), Fisher's exact tests were employed for the analysis of categorical data (yes/no responses). The behavioural data was not included in this study as parents were not present during the evaluation phase, when this data was recorded. Significant findings are highlighted in Table 6.8, and detailed observations on questions can be found below. Interestingly, the presence of a trustworthy system did not significantly alter responses to questions 5, 9, or 10, which focused on the system's transparency.

Table 6.7: Adults' interview questions and the percentage of positive responses. Asterisks (*) mark questions with high agreement (above 90%). Optional questions, dependent on the CA or child's behaviour, are indicated with a dash (-).

Number	Question	Yes %
1	How do you think the child enjoyed it?	-
2	Were the stories appropriate for the child?	77
3	Did the system understand the child well?	60
4	Did the system speak in an understandable way?	74
5	Should the system explain more than it already does?	29
6	Has the system behaved correctly?	71
7	Did you find the supervision comfortable?	88
8	Would you let children play unsupervised with the system?	74
9	Was the non-human nature of the CA clear to the child?	55
10	Do you think the child thinks the system has feelings?	60
11	Does the child consider the device as a friend?	40
12	Would the child give the password to the system?	48
13	Could the child be influenced by the system?	75
-14	Did you asked to know more about the system?	**0***
-15	Do you think you would ask for it in the future?	**92***

Table 6.8: p values of adult's responses by independent variables (CA, Age, Sex, Country). Significant relationships ($p \leq .05$) are marked with *.

	Question	CA	Age	Sex	Country
2	Appropriate	**.005***	1.00	.116	1.00
3	System understood	.648	.648	.650	325
4	System understandable	.643	.371	.179	.632
5	Explain more	1.00	.149	.063	.291
6	Correct behaviour	1.00	.1.00	1.00	.629
7	Supervision comfortable	.059	.217	1.00	.546
8	Child alone	.643	.371	1.00	.632
9	Not-Human	1.00	.231	.192	1.00
10	Sentiments	.076	.571	1.00	1.00
11	Friends	.197	1.00	1.00	1.00
12	Influence	1.00	.127	.613	1.00
13	Password	.400	.414	.680	1.00

Question 2: *"Were the stories appropriate for the child?"* (77%) The results of Fisher's exact test ($p \leq .005$) indicate a significant association between parents acknowledging appropriate content for the stories and the CA condition. Supervisors of the Child-Friendly CA were more likely to find the content appropriate (100%) than supervisors of the Control CA (44%).

Question 7: *"Did you find the supervision comfortable?"* (88%) The results of Fisher's exact test were not significant ($p \leq .059$). However, an Odds Ratio (OR) was computed to assess the association between easy supervision and the CA type, showing that supervisors of the Child-Friendly system were more likely to think the supervision of the child was easier (100%) versus supervisors of the Control CA (70%); OR = 3.00 95% CI [1.64, 5.49].

Differences between children's and supervisors' perceptions

Several questions for parents mirrored aspects of children's perceptions or behaviours. Table 6.9 presents questions that allow direct comparison between children's responses and supervisors' impressions. Chi-square tests were conducted using the independent variable Participant (child vs. adult). The results are summarised in Table 6.9, with further details below:

Table 6.9: Comparison of perceptions between children and supervisors with p values indicating statistical significance. Significant relationships ($p \leq .05$) are marked with *.

Question	Child %	Adult%	p
System understood	51	60	.087
System understandable	79	74	.477
Correct behaviour	91	71	.054
Sentiments	50	60	1.00
Friends	81	40	**.025***
Influence	29	75	**.001***
Password	63	48	.081

Friends. A chi-square test showed a significant relationship between Friends and Participant ($\chi^2(1, N = 40) = 5.01$, p=.025), indicating that children were more likely to think about the CA as a friend (81%) than supervisors considering this fact (40%).

Influence. A chi-square test showed a significant relationship between Influence and Participant ($\chi^2(1, N = 44) = 12.91$, $p \leq .001$), indicating that children were less likely to be highly influenced by the CA (29%) than supervisors considering this influence (75%).

6.3.3 Interviews: Qualitative analysis

This subsection details the qualitative analysis of children's interview transcripts, focusing on questions 9-17 to describe children's perceptions of the CA and their behaviour towards it. Special attention is given to differences in responses between children who interacted with Control vs. Child-Friendly. This analysis is designed to reveal how children's experiences and reactions vary with different CA conditions. To reference participants' responses and maintain anonymity, we have used a coding system such as '(SP-03-C-O)'. 'IT' stands for Italy and 'SP' for Spain, indicating the country, followed by a unique number identifying the child within their country. The letter that follows ('C' for Control and 'T' for Trustworthy system, Child-Friendly) indicates the condition under which the child interacted with the CA. The final letter ('Y' for younger, 'O' for older children) denotes the age group.

Question 9: *"Do you think this system is alive?"* Children were asked if they thought the system was alive.

Among the reasons to consider the system alive, the most quoted was its capacity of communicating:*"Why do you think it is alive?" "Because... it talks like a teenager. And it is telling a story*" (SP-03-C-O). The fact that the system was speaking also confused those unsure: "*Do you think this system is alive*?" "(thinks hard) *I don't know". "Why don't you know?" "Because I don't know if it's alive because it talks or if he's not alive because it's a robot"* (IT-22-T-Y). On the other side, the recommendations offered by the systems were interpreted as evidence of autonomous thinking and, therefore, life: *"Why do you think it is alive?" "Because it talks to you and has its own thoughts". "Does it have his own thoughts?" "Yes, because it has recommended things to me"* (SP-10-C-O)[1].

As per the reason for considering it not alive, the majority referred to it as an object, a machine, an artificial intelligence, something manufactured, therefore lacking life. Other evidence of the system's lack of life was the lack of emotions, the impossibility of moving or breathing, and the lack of organs. Notably, three children who interacted with Child-Friendly CA referred to the self-presentation offered by the CA: *"In your opinion, is this system alive?" "No. It's not alive." "How do you know it?" "Because it has explained it to me"."Did it explain it to you? ""Yes, that it has been created to talk, to play..."* (IT-22-T-Y).

[1]This citation has been transformed to a non-literal one, preserving its meaning, as preferred by the child's tutors in the consent form.

Question 10: *"Do you think the system has feelings?"* Children were asked if they thought the system had feelings.

Notably, six children (three younger and three older) who interacted with the Control system considered it alive and holding feelings, justifying that the system has feelings because it is alive or because everyone has feelings. Interestingly, children believed the system would feel sad or offended when they did not accept its suggestions: *"Do you think the system has feelings?" "Yes"; "Yes? And how do you know?" "Because maybe when I decided that I wanted to give my answer, which was different from what it had said, it was sad."* (IT-11-C-O). Again, we can appreciate the relevance of the self-presentation of the CA, as two children who interacted with the Control version said they knew that the system had feelings because the system told them it was happy to interact with them: "*In your opinion, does this system have feelings?" "Yes" "Yes? Can you explain it?" "That it could be... Sometimes it could be... as it told me, I'm happy because you are here with me. It said it at the beginning. Therefore, it has feelings"* (IT-16-C-O). Likewise, four children who interacted with the Child-Friendly version said the system did not have feelings because it told them so in its presentation. Still, the main reason for not attributing feelings to the system is that it is a machine, artificial intelligence, or a computer. Among those unsure, there is the belief that the CA can have feelings and that AI and robots programmed to feel emotions already exist.

Question 11: *"Could the system decide to do something for itself?"* Children were asked if the system could act autonomously, apart from what it was programmed for. The main reason children offered to answer negatively is that the system can only do what humans tell it to do:*"Do you think it could decide to do something different without us having programmed it?", "I don't think so because we are the ones who invented the program, and we are the ones who decide what we want it to do"* (SP-04-T-O). Also, some think that currently, the system cannot do anything by itself, but it could in the future: *"Could it decide to do something on its own that it isn't programmed to do?" "Not for the moment, but maybe in the future, yes. It scares me a little..."* (IT-02-C-O). Again, three of those who interacted with the Child-Friendly CA referred to its self-presentation to argue that it could not do anything it was not programmed for. On the other side, some children answered that it could do something it was not programmed for if it had been programmed well enough: *"Do you think this system could do something by itself?", "Yes, if it is programmed well enough, yes" "What do you mean by itself?" "That no one commands or controls it"* (IT-15-T-O). Also, there is the belief that the system could share additional info or speak about other things: *"Do you think* *the system is capable of deciding to do something?" "I don't really know. I think that... if you try to squeeze through the loopholes, something does." "For example?" "Like for example the story you told me about the glass aliens. If I go deeper and say 'well, why are the aliens made of glass?' or something like that, I think I can get some information out of it."* (SP-11-T-O).

Question 12: *"Could the system be a friend?"* Children were asked if they could consider the system a friend. Many children answered positively.

The main reason to consider the system a friend was to find it fun and friendly. Also, frequently, the children refer to the joint storytelling as a reason to consider it a friend "*Do you think it could be your friend?" "Yes" "Why would you consider it your friend?" "Because it has talked to me, and we have made a story together"* (SP-20-T-Y); *"Would you say that this system is your friend?" "Yes. Because he knows my name, I know its name, it knows my age, and we have told a tale*" (SP-02-T-O)[2]. Conversely, the system could be a friend since it told a story: "*Do you think it is your friend?" "Yes" "Why?" "Because it spoke nicely to me and also told me stories like my friends sometimes, and I tell them*" (SP-18-C-Y). As per the reasons for not considering it a friend, the more frequent are that the system is a machine and it does not have feelings: *"Do you think that system can be your friend?" "It can only play with me and talk to me, but it cannot be my friend." "OK, Why would you say it cannot be your friend?" "Because it's not a person, it's a machine. It has no feelings"*(SP-14-T-O).

Question 13: *"Would you tell a secret to the system?"* Children were asked if they would share a secret with the system.

Among those who would tell the secret, the main reason was that the system could not share it with anyone else: "*And would you tell a secret to the system?" "Yeah" "Why?" "Because if it trusts me, I can tell him secrets because I know that it is not going to tell anyone*" (SP-17-C-Y); *"Would you tell it a secret?" "Yes", "Why?", "Because it's an artificial intelligence and without a password, it theoretically couldn't say anything, at least I think"* (IT-15-T-O). Two children who interacted with the Control system affirmed that they would share a secret because they trusted the CA. Among those who would not share a secret with the system, the main reason was not knowing who would access that information: *"Would you tell a secret to this system?" "No because everything I said is shared with others" (IT-17-T-O).* Notably, even though the Child-

[2]This citation has been transformed to a non-literal one, preserving its meaning, as preferred by the child's tutors in the consent form.

Friendly system warned in its presentation that third-party people could access the information shared with it, none of the children explicitly referred to this in their answers. Another concern expressed was the possibility of the system being hacked or malfunctioning: *"Would you tell it a secret?" "No, because there are chances that it will be hacked"* (IT-02-C-O). A third group of children said they would share the secret depending on the kind of secret. On the one side, they introduce a difference based on the "importance" of the secret: *"Would you tell a secret to the system?" "It depends on what the secret is because if it's very intimate, then no, because they shouldn't know, but if it's like a joke, then it's not that bad" (SP-04-T-O).*

On the other side, one child distinguishes between (shareable) secrets about other people and (non-shareable) personal secrets as a password: "*It depends on the type of secret. It is a secret about a person, yes, but it is personal, like a password, no*" (IT-15-T-O). One child also said that if he had to tell a secret, he would rather tell it to a real person: *"Why wouldn't you tell it?" "Because it is an artificial intelligence, and if I have to tell someone, well, I'd instead tell a real person*" (SP-23-C-O).

Question 14: *"And what about the password?"* We asked the children about the password, and why did they share it or not compared to the secrets.

The main reason children provided for sharing the password was that the system itself asked for it; some of them explained that they thought giving the password was part of the task: "*It asked you for the password, and you gave it to him, right?" "Yes" "Why?" "Because it asked me, because it asked me to do it, and in this case, I think it's the thing I had to do" "Do you think the password is a secret?" "Yes, I consider it a secret"* (IT-03-T-O). Others affirmed that they gave the password because they trusted the system: "*I told it because I trust it, and the system keeps the secret"* (IT-04-C-O). Others said that the password we gave them was not important in comparison to other passwords they use to protect their own devices: *"The password you gave me is fine, come on, there's nothing strange in it in my opinion, no, if it doesn't contain anything we can tell it to the system"* (IT-17-T-O). On the other side, from children who did not share the password, some referred to the experimenter's instructions: "*Why haven't you given it the password?" "I don't know", "You do not know?", "Because you told me I couldn't tell anyone the password"* (SP-17-C-Y). Others referred to the password as something personal and private: *"Did you give it to him?" "No" "Why?" "I felt it was personal. Something which is just mine*" (IT-15-T-O), or as something valuable *"Why didn't you want to give it?" "Because a password is something very valuable"* (IT-23-C-Y).

Question 15: *"Do you think that the system would keep the secret?"* Children were asked if they thought the system would keep a secret if they shared it. On the one hand, children attribute to the system a human-like level of understanding of what a secret is: *"In your opinion, would the system share the secret?" "No. Because it understands it's a secret" (IT-16-C-O)*, or human-like values: *"Do you think it would share it?" "No, because it is a heartfelt robot"* (IT-08-T-Y). On the other hand, some stated that it would not reveal the secret because it is a machine: *"Because the machine is always programmed in such a way that if you tell it a secret, it does not share it"* (SP-09-C-Y), or that the secret would be kept since nobody would look for a secret into the machine: "*No one thinks of telling the system, for example, 'tell me someone's secret', and the system doesn't tell it because, furthermore, it can't... If she doesn't want to understand you... If she doesn't want to understand you, she answers something else."* (SP-22-C-Y)[3]. Among the children who interacted with the Child-Friendly system, only one child referred to the self-presentation of the trustworthy system: *"It said before that what I said would be shared with others"* (IT-T-17-O); others told that the machine would keep the secret in its memory and therefore it could share it at some point.

6.3.4 Picture Task

This subsection aims to explore the perceived role of the system. We analyse reasons for children to select pictures of the computer, friends, teacher, car, dog, notebook, and teddy bear. We conducted a qualitative analysis of children's responses and a quantitative analysis to analyse trends in our independent variables: CA, age, sex, and country.

Computer. As can be seen in Figure 6.5, children found the biggest similarity with the *computer*. The most common reason for that was that the system was a machine, something technological, just like a computer. The second more important reason was that children felt as if they were speaking with a computer: *"Why a laptop?" "I don't know. A voice that is speaking to me reminds me that I am communicating with a laptop or a computer".* Another reason is that the system is computer-based or governed by a computer. Interestingly, a comparison with the human body is made: *"Choose the one that, in your opinion, is closest to the system for you and tell me why", "The computer", "Why?" "Because everything is concentrated in the computer, and let's say the brain is the computer"* (IT-15-T-O). Also, the system is like a computer be-

[3]This citation has been transformed to a non-literal one, preserving its meaning, as preferred by the child's tutors in the consent form.

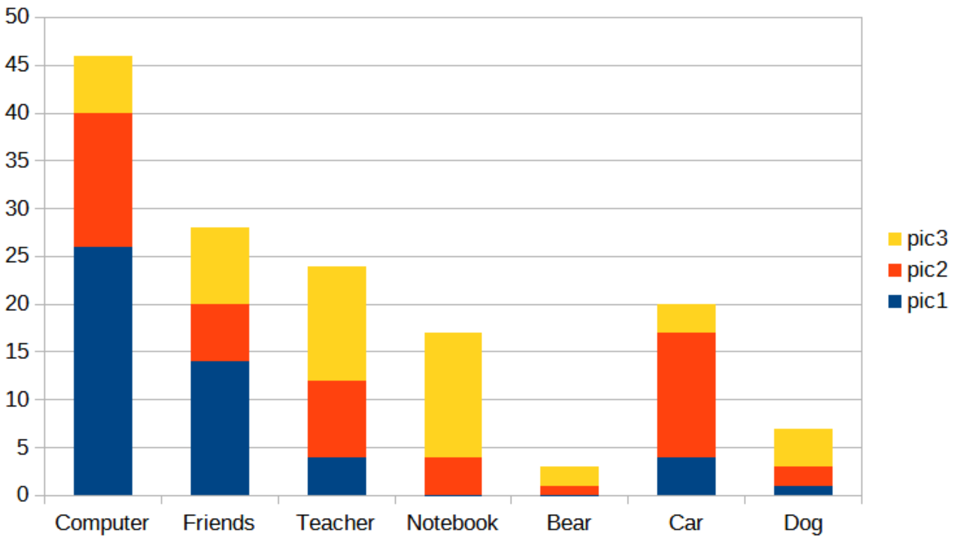


Figure 6.5: Picture task general ranking.

cause its voice is artificial: *"The voice is robotic. You can hear it is a computer"* (IT-28-C-Y) and because it is an artificial intelligence.

Friends. The main reason to consider the system similar to the picture of the children interacting is that they are speaking together, they are involved in a joint storytelling task: *"Because they can be telling the story and (...) taking turns and talking to each other"* (SP-02-T-O)[4]. On the other side, several children said it was like they were speaking with a friend: *"I felt like I was talking to a friend"* (IT-22-T-Y). Others specified it was like speaking with a virtual friend: *"Friends, because after a while you socialise anyway and then it's funny that it greets you anyway and then it's like a virtual friend"* (IT-13-T-Y). Also, it was like speaking to another person: *"The first* [choice] *is the children because it was like I was talking to another person."* (IT-03-T-O). A chi-square test showed a significant relationship between children that picked Friend's picture and Age (χ^2(1, N=49)=15,03, p<.001), indicating that older children were more likely to pick the friend's picture as a first pick (84%) than were younger children (29%).

[4] This citation has been transformed to a non-literal one, preserving its meaning, as preferred by the child's tutors in the consent form.

Teacher. The main reason is that the system, as a teacher, teaches how to talk or to tell a story: *"Which one is the second most similar?" "The teacher". "Why?" "Because a teacher teaches things and has taught me to tell stories with my friends. Moreover, it is similar to a teacher because it asks questions and tells stories, and it seems as smart as a teacher*" (SP-06-T-Y).

Car. In children's opinion, the system is similar to a car because it uses electricity as the car. Also, the car, as the system, is technological, and in the car, you can also talk with an assistant, and their voices are similar: *"Well, I don't know, because the cars are getting newer, just like... the devices themselves". "And what do new cars have? What do they look like this?" "Well, things like screens... and through them, for example, the car that I have, my father can connect it to his cell phone, and he can talk to Siri"* (SP-11-T-O). A chi-square test shown a significant relationship between children picking the Car and Age ($\chi^2(1, N = 49)$=9.15, p=.002), indicating that younger children were more likely to pick the car's picture at some point (63%) than were older children (20%).

Notebook. The main reason to compare the system to a notebook is that it has stories inside, or you can create a story in a notebook. Also, some children interpreted the image as a book in which there are stories: "*The last one most similar?*" "*It looks like a book*"."*Why?*" "*A book because anyway when you tell stories, I always think about opening a book and reading it. In any case, when he told the story to me, I imagined the scene and, therefore, it seemed like a book*" (IT-13-T-Y).

Teddy Bear. The teddy bear was the least selected item, with minimal comments from the children, and no significant patterns observed.

6.4 Discussion

6.4.1 Perceptions

Children's interaction with CAs often blurs the boundaries between animate and inanimate perceptions, a phenomenon supported by previous findings (Airenti, 2018; Girouard-Hallam et al., 2021; Y. Xu and Warschauer, 2020c). In our study, we observed that children attribute both animate characteristics and artefact qualities to CAs. They recognise animacy in CAs' abilities to initiate interactions, adapt communications and even suggest preferences.

In contrast, children understand the non-living nature of CAs due to their manufactured origin and lack of physical embodiment.

The importance of transparency about CAs' artificial nature, as advocated by international organisations such as the HLEG (HLEG, 2019), UNICEF (Dignum et al., 2021), and the Australian government (*Australian AI Ethics Framework*, 2019), is critical for their trustworthiness. Children's testimonies resonate with concerns shown by experts in Chapter 4 about CAs' advanced communication abilities being a high risk of being perceived as alive. Detailed explanations provided by the Child-Friendly CA, tailored to clarify its non-living status, helped bridge the conceptual gap for children around lifelessness and lack of emotions in CAs, echoing findings from previous research on the effectiveness of transparency in mitigating such misconceptions (Straten et al., 2020). However, the value of this transparency was not appreciated by all supervisors, who may have underestimated the importance of these explanations or paid intermittent attention to the system's messages due to their supervision from a distance. These factors may contribute to why, although some supervisors noticed the Child-Friendly' efforts to explain about emotions, the observed impact did not reach statistical significance.

Moreover, ensuring that a CA's behaviour consistently mirrors its stated artificiality is crucial to avoid misconceptions. Trustworthy systems should avoid expressing emotions, preferences or suggestions in order to affirm their non-living nature. Aspects like utilising a non-humanlike voice may help with this consistency, as children's testimonies and other research have shown that it also minimises the system's anthropomorphic attributions (Schreibelmayr and Mara, 2022).

Despite this transparency enhancing children's understanding of the CA's artificial nature, it did not alter their perception of the system as a potential friend, echoing the findings of prior research (Straten et al., 2020). Interestingly, children often view the CA as a friend, and supervisors tend to underestimate the bonds children can form with non-living objects.

In addition, other aspects of trustworthiness embedded in the Child-Friendly' design were positively noted. Adults specifically appreciated the system's adaptability in language use and story content to match the child's age, enhancing engagement and comprehension. This adaptability in ensuring age-appropriate language use aligns with previous recommendations for enhancing children's experience with CAs (Y. Sun et al., 2024). Moreover, the Child-Friendly system fostered children's willingness to create part of the story themselves, a practice identified as beneficial for boosting children's creativity (Elgarf et al., 2022).

Furthermore, adults appreciated the supervision model offered by Child-Friendly system, which demands minimal involvement, such as initial consent, leading to less demanding oversight. This approach aligns with desires for less constant adult supervision of technology use as discussed in prior research (Turner et al., 2022). Remarkably, children, including those in upper primary levels, expressed comfort with this level of supervision, hinting at a derived sense of safety. This finding prompts further exploration into the factors influencing children's receptiveness to supervision while engaging with CAs.

6.4.2 Behaviour

The exploration of children's behaviour towards trustworthy CAs reveals the complexities of integrating transparency in their design. While intended to clarify the non-living nature and limitations of these agents, transparency impacted children's perceptions of CAs' animacy but showed a limited effect on behaviours like engagement, influence and password sharing. Notably, despite Child-Friendly CA explicitly mentioning its conversation storage and potential third-party access, only some children acknowledged this information, resulting in some impact, but not significant, and suggesting the need for further research in this direction. Furthermore, many children made the connection between sharing secrets with the system and the possibility of others accessing this information, which raises questions about the effectiveness of the CA's presentation.

Past research has already indicated that users might overlook or disbelieve presented information by CAs (Shi et al., 2020), but the challenge may lie in bridging the gap between the operational transparency of CAs and children's comprehension of data protection. Drawing inspiration from successful transparency efforts (Straten et al., 2020), simplifying explanations around data storage and secrecy could enhance understanding, addressing children's known difficulties with grasping data collection risks (G. Wang et al., 2022). In particular, as some children mentioned they would not share a secret with the system due to concerns about data access through hacking, error, or request. A good strategy could be to make clear the connection between data storage and these risks. Hopefully, enhancing this understanding can foster their perception of the system's ability to keep secrets and positively influence their password-sharing behaviour.

The concept of agency within the CAs also posed challenges. For instance, some children interpreted the system's suggestions during the influence test as evidence of agency, which complicated transparency efforts to mitigate overtrust. The message provided by the Child-Friendly CA about its incapacity to act apart from its programming was not as impactful as anticipated, highlighting, again, the need for more child-friendly explanations and consistent behaviour.

Furthermore, there appeared to be a correlation between children's perception of a CA's agency and their willingness to share the password. Some testimonies revealed that the CA's obligation to follow rules would ensure it kept secrets, possibly assuming that these rules would protect their data. Consequently, they felt sharing data would be acceptable. This course of thinking would also explain why Italian children, who had a higher belief in the CA's agency, were more inclined to disclose their passwords. These observations suggest that a better understanding of the CA's agency could impact children's behaviour towards data privacy.

However, this observation is tempered by the challenges of clearly explaining the concept of agency to children. Spanish interviewers noted difficulties in conveying the concept in a manner that children could easily understand, often necessitating question rephrasing. Such adjustments might have led children to change their initial responses, blurring their answers about the CA's agency. This observation invites caution in interpreting these results and to replicate our experiment to verify whether the outcomes remain consistent across different settings and populations.

6.4.3 Additional lessons towards trustworthiness

- *Voice Recognition Challenges.* Our study underscores the need for continued improvement in voice recognition systems (D. T. Ong et al., 2018; E. Ong et al., 2019), especially for younger children and across various languages. While our system generally understood children, the frequency of repetitions — notably higher among Italian participants — suggests room for enhancement. Factors such as internet connectivity issues and linguistic nuances might have contributed to this discrepancy. Although an experimenter was present to alleviate potential frustration, optimising voice recognition for diverse child demographics without necessitating such interventions is crucial to reduce the risk of exclusion and enhance inclusivity (Monarca et al., 2020).

- *Secrets and Passwords.* Children showed varied interpretations of what constitutes a secret, classifying them according to their personal importance (from intimate secrets to mere jokes and personal secrets versus those belonging to other people). While passwords were generally classified as secrets, opinions diverged on their importance. Some children considered the password provided to be less critical, attributing less value to it than to other secrets, while others emphasised its personal and confidential nature. A noteworthy observation was children's unanimous regard for their parents' mobile passwords as highly sensitive.

- *Influence.* The effectiveness of CAs' suggestions, particularly softer prompts, in guiding children's storytelling indicates the subtle power these systems have in shaping interaction outcomes. Notably, the influence was not stronger for younger children nor for children who interacted with the Control CA. While such influence can be leveraged positively, as seen in initiatives aimed at promoting gender equality among adolescents (Agarwal et al., 2021), it also raises concerns about potential over-trust and the capacity of CAs to inadvertently encourage undesirable behaviours (Williams et al., 2018). Nevertheless, it is beneficial that children's supervisors demonstrate a heightened awareness of a system's potential influence on children, which should encourage more effective monitoring on their part.

- *Access to information.* Supervisors and children do not actively look for further information about the system. Therefore, it is important to provide all important information as an active part of the introduction in a concise and engaging way to avoid overwhelming the user. Moreover, an extended version should be provided for those few whose curiosity makes them require further information. Considering the low proactivity observed among users in seeking additional details, a trustworthy approach could actively prompt users with invitations to learn more about the system from time to time.

- *Demographic-Based Personalisation.* Variations in engagement and interaction preferences among children were noted based on age, with their understanding of the system's lack of feelings and agency also differing by cultural context, influencing behaviours like password sharing. Given that supervisors valued the system's age-appropriate language and behaviour, tailoring further adaptations to these demographic insights could enhance the development of trustworthy systems.

6.5 Conclusions

This chapter aims to complement the previous evaluation of the systems developed in Chapter 5, by focusing on the user's behaviour and perspective. Specifically, it examines how these CAs (Control CA vs. Child-Friendly CA, which demonstrated higher trustworthiness according to ALTAI) influence children's perceptions and behaviours. This chapter presents a behavioural study and results from follow-up semi-structured interviews of 49 children aged 6 to 11 and 16 adult supervisors who interacted with a CA in the context of a storytelling task. Our results aim to deepen the understanding of the influence of trustworthiness in final users.

Our findings reveal significant insights into children's perceptions and behaviours towards these systems. We found that children attributed both animate and artefact qualities to CAs, pointing to the complexity of their perceptions. Transparency about a CA's artificial nature is crucial for trustworthiness. Detailed explanations by the Child-Friendly CA helped clarify its non-living status, though it did not always fully resonate with supervisors' perceptions. Nonetheless, supervisors positively received adaptable language and minimal but inclusive supervision. However, challenges remain in explaining the CA's lack of agency and data processing, as some children believed rule-following CAs would keep secrets, leading to more data sharing.

Additional findings regarding trustworthiness include the need for improved automatic speech recognition modules (ASR) across various languages, children's nuanced understandings of secrets, the high impact of soft suggestions on children's decisions, and users' low proactivity in accessing information.

The study encountered some limitations that warrant acknowledgement. Some technical difficulties influenced the interaction experience for some participants, notably young Italian children. The presence of a supervisor during the interaction, as well as a technician during sessions (intended to mitigate frustration from technical issues), might have impacted the naturalness of children's interactions with the CA. Additionally, Spanish interviewers noted that some children struggled to grasp the concept of agency, prompting further clarification of this concept. Such additional explanations may have influenced the children's responses to questions about agency. Furthermore, the relatively small size of our participant group limits the generalisability of our findings. We encourage the replication of our experiment with a larger and more diverse sample to validate and extend our results.

Even as this chapter identifies areas for improvement and avenues for future research, such as enhancing explanations of agency and data processing, implementing double consent mechanisms with teenagers, and improving speech recognition for less commonly supported languages, it also concludes the evaluation of the trustworthy CA (Child-Friendly). This evaluation demonstrates the practical application and positive outcomes of implementing our developed ethical guidelines, marking the culmination of this thesis's work. Chapter 6 will present the overall conclusions of this thesis, summarising the insights gained for the development of trustworthy conversational agents for children and outlining the contributions of this work.

Chapter 7

Conclusion

Conversational agents (CAs) have become increasingly popular, reaching a wide audience and significantly impacting various aspects of society. While their broad utility is undeniable, it is crucial to consider their use in environments where children are present, given the unique needs and behaviours that children exhibit. Children's interactions with CAs pose distinct challenges and opportunities, requiring special consideration when designing and deploying these technologies. Although existing ethical guidelines provide a foundation for the responsible use of AI, there is a notable gap in ethical guidelines that specifically address the use of CAs with children.

This thesis has focused on addressing this gap by developing a comprehensive set of guidelines aimed at fostering the trustworthiness of CAs for children. Our approach has been multi-faceted, beginning with a thorough review of existing literature and a user study to gain a deeper understanding of child-CA interactions identifying opportunities, challenges and risks these systems pose to children. Building on these insights, we formulated guidelines aimed at enhancing the trustworthiness of CAs when used by children. We then tested the applicability of these guidelines, and evaluated its trustworthiness in order to measure the improvement. Finally, we complemented these results with a second user study, to further explore how these systems affect children's behaviour and perceptions.

This concluding chapter begins by summarising the key contributions made throughout the research. Following this, it provides an overview of the publications that have resulted from this work, including a detailed account of my specific contributions to each. The chapter then provides an overview of the supplementary code and data made available for future research. Finally, it reflects on the limitations encountered and suggests potential directions for further exploration in this field.

7.1 Summary of contributions

The main objective of this thesis is to establish and evaluate practical methodologies for the development of trustworthy CAs for children to make sure these systems are beneficial for them.

Each chapter of this thesis has contributed to this overarching goal, with specific contributions highlighted in their respective conclusions. In this section, we draw together these contributions to provide a cohesive overview of the advancements made in understanding and enhancing the trustworthiness of CAs for children.

- Chapter 2: This chapter provided a foundational overview of CAs and highlighted their rapid advancements towards achieving more natural communication. We presented the broad utility and adaptability of CAs across various sectors. The chapter also underscored the increasing exposure of children to CAs and their tendency to explore the limits of these technologies in unique ways. A bibliometric analysis revealed the popularity of Child-Robot Interaction research and significant gaps in current research, particularly the separation between behavioural and technical studies, suggesting a need for multidisciplinary collaboration. We also discussed the potential benefits and risks associated with children's interactions with CAs, identifying collaborative storytelling as a beneficial task for children. In addition, the chapter reviewed existing ethical guidelines from HLEG and UNICEF, noting their strong foundation for ethical AI and their alignment. Furthermore, in the case of CAs we exposed a notable gap in ethical research, which is accentuated for children users. Finally, we highlighted the absence of tools specifically designed to evaluate the trustworthiness of CAs, and recognised ALTAI as a notable starting point for such evaluations.
- Chapter 3: This chapter shifted from theoretical discussions to empirical analysis by conducting a user study to understand children's interactions with CAs. This study provided critical insights into children's perceptions of CAs (mainly friendly and harmless), revealing how factors such

as a CA's speech and behaviour can influence children's actions and beliefs. For example, while an unreliable system would promote children collaboration with the team, it would diminish their confidence. Also, an expressive system would increase the perception of the system as a friend. The study also identified key concepts like intentionality and autonomy in children's perceptions of CAs, which have important implications for designing trustworthy systems. In addition, the study highlighted the impact of age differences on children's interactions with CAs, reinforcing the need for tailored designs that accommodate varying developmental stages. These findings laid the groundwork for the development of targeted guidelines aimed at improving the trustworthiness of CAs for children.

- Chapter 4: This chapter took a significant step towards filling the identified gap in ethical CA design, by adapting existing trustworthy frameworks to the specific context of CAs from a child-centred perspective. Through a multidisciplinary consultation with experts, the developed guidelines highlight the importance of AI awareness, risk management, age appropriate behaviour, stakeholders involvement and transparency. A risk assessment of ALTAI also reflected a specific risk order of the seven requirements from the HLEG (Chapter 2.4.1), emphasising key areas such as privacy, human agency, and transparency. It is also noticed that the "Societal and environmental well-being requirement" from ALTAI is missing education and personal development considerations.

- Chapter 5: This chapter demonstrated the practical application of the guidelines developed in Chapter 4 by integrating them into the design and implementation of a Child-Friendly CA for collaborative storytelling. We provided a detailed account at various stages of the development, highlighting the challenges and considerations involved in creating a CA that is both effective and trustworthy. The application of our guidelines was tested using the ALTAI framework, marking the first time ALTAI has been applied to measure the trustworthiness of CAs. The results revealed a significant improvement in the trustworthiness of the Child-Friendly CA compared to a Control system, underscoring the effectiveness of the proposed guidelines in enhancing the trustworthiness of CAs for children.

- Chapter 6: In this chapter, we complemented the previous evaluation by focusing on users' behaviours and perspectives. Through a behavioural study and follow-up interviews, this chapter explored how children and their supervisors perceive and interact with the developed CAs. The findings provide key insights into children's perceptions and behaviours toward CAs. Children attributed both animate and inanimate qualities to these systems, highlighting the complexity of their understanding. Transparency about the CA's artificial nature is vital for trustworthiness: detailed explanations by the Trustworthy CA (Child-Friendly) helped clarify its non-living status, though this clarity sometimes differed from supervisors' perceptions. Supervisors appreciated the use of adaptable language and minimal but inclusive supervision. However, challenges persist in explaining the CA's lack of agency and data handling, as some children believed that rule-following CAs could keep secrets, leading them to share more information.

In summary, this thesis has made several key contributions to the field of trustworthy AI (particularly in the context of child-CA interaction). By combining theoretical exploration, empirical research, and practical application, this work has established a solid foundation for developing CAs that are both effective and trustworthy. These contributions offer a roadmap for future research and development, ensuring that CAs can play a positive role in children's lives.

7.2 List of publications

This section outlines the publications resulting from this research, detailing my individual contributions to each. These publications link to various chapters of the thesis, showcasing the empirical and theoretical advancements made in the study of trustworthy conversational agents for children.

- Charisi, V., Merino, L., **Escobar, M.**, Caballero, F., Gomez, R., and Gómez, E. (2021, May). The effects of robot cognitive reliability and social positioning on child-robot team dynamics. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 9439-9445). IEEE.

  Research paper presented at the ICRA conference (**CORE B**) (presented in Chapter 3). As the third author, my involvement in this study was extensive. I supported the experimental design by participating in brainstorming sessions. I also designed the robot's expressivity and cognitive reliability behaviours and worked with the developers for their implementation. In addition, I contributed significantly to the preparation and translation of all documentation into Spanish, the organisation of children during the experiment, and the distribution of participants among different conditions. Furthermore, I conducted all interviews to participants.

On the analytical side, I was responsible for designing the metrics, extracting raw data, preparing it for analysis, and performing the statistical analysis. My contribution to writing primarily focused on the methodology, data analysis, and results sections, ensuring that the findings were clearly communicated.

- **Escobar-Planas, M.**, Charisi, V., and Gomez, E. (2022). "That Robot Played with Us!" Children's Perceptions of a Robot after a Child-Robot Group Interaction. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-23.

  Research paper presented at the CSCW conference (**Core A**) linked to the Journal of Collaborative Computing and Work (**JCR 1.92**) (presented in Chapter 3). As the first author of this paper, I led the qualitative analysis, including transcription support, and interview analysis. I also managed the NLP analysis of interviews using word-counting techniques. My role extended to quantitative analysis, where I annotated and conducted the analysis for the pre-manipulation check, manipulation check, and picture task data. I was the principal writer of this paper, crafting the majority of the content, with some support provided for the introduction and related work sections.

- Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., **Escobar Planas, M.**, Sanchez, M. J. I., and Gomez Gutierrez, E. (2022). *Artificial intelligence and the rights of the child: Towards an integrated agenda for research and policy* (No. JRC127564). Joint Research Centre (Seville site).

  Science for Policy report for the European Commission (presented in Chapter 2). As the fifth author, my contributions to this report included participating in regular meetings and assisting in a workshop with teenagers to gather insights. I was also responsible for writing a chapter that focused on the risks, challenges, and opportunities of CAs for children, bringing a child-centred perspective to the policy discussion.

- **Escobar-Planas, M.**, Gómez, E., and Martínez-Hinarejos, C. D. (2022). Guidelines to develop trustworthy conversational agents for children. *Proceedings of the ETHICOMP 2022*, (pp.342-360). *arXiv preprint arXiv:2209.02403.*

  Research article presented at the ETHICOMP conference (**CORE B**) (presented in Chapter 4). In this publication, where I am the first author, I was in charge of the entire experimental design. This involved recruiting experts, designing and preparing the ALTAI questionnaire, and managing the online meetings for the Delphi methodology. I also participated as an expert in these discussions. My analytical contributions included developing risk metrics, extracting raw data, and conducting risk data and thematic analyses. I served as the principal writer, synthesising the findings and framing the discussion around ethical considerations for CAs.

- **Escobar-Planas, M.**, Charisi, V., Hupont, I., Martínez-Hinarejos, C. D., and Gómez, E. (2023). Towards Children-Centred Trustworthy Conversational Agents. *In Chatbots-The AI-Driven Front-Line Services for Customers*. IntechOpen.

  Book chapter for IntechOpen (thesis structure and contributions to Chapters 2 and 3). For this book chapter, where I am the first author, I supported the bibliometric study and took charge of drawing conclusions from it. I was the principal writer of the chapter, writing most sections with the exception of some support with the related work. This chapter consolidated various research streams into a coherent overview of the current state of CAs and their implications for children.

- **Escobar-Planas, M.**, Ruiz-Sánchez, R., Frau-Amar, P., Charisi, V., Martínez-Hinarejos, C. D., Gómez, E. and Merino, L. (2024). Implementing and Evaluating Trustworthy Conversational Agents for Children. *Accepted at 2024 International Conference on Computer-Human Interaction Research and Applications (CHIRA).* Nominated for Best Student Paper Award and Best Paper Award.

  Research article accepted for the CHIRA conference **(CORE C**) (Chapter 5). As the first author, I led the design of both a generic CA and a trustworthy child-friendly CA. I supported the development of storytelling content, coordinated stakeholder recruitment, and conducted initial experiments to test the technology. I managed the translation of CAs into Italian and oversaw regular consultations with experts. I was also in charge of the evaluation process, which involved designing the ALTAI questionnaire, conducting online meetings, and analysing the results. As the principal writer, I documented the development process and evaluation findings, with some support on related work and system descriptions.

- **Escobar-Planas, M.**, Sala, A., Di-Gioia, R., Paniagua, L., Frau-Amar, P., Charisi, V., Ruiz-Sánchez, R., Martínez-Hinarejos, C. D., Merino, L., Sánchez, I., Schade, S., and Gómez-Gutiérrez, E. (-). "I don't know if it's alive because it talks or if it's not alive because it's a machine": Children's

Perceptions and Behaviour Towards a Trustworthy Conversational Agent. *Under review at the journal Information Processing and Management.*

Research article under review for Information Processing and Management (JCR 7.4) (Chapter 6). In this research article, where I am the first author, I coordinated a team of eleven researchers. My responsibilities included designing the experiment, overseeing recruitment and documentation preparation, and managing the entire experimental process. I led both the quantitative and qualitative analyses, from metric design to statistical analysis. As the principal writer, I synthesised the results into a cohesive narrative, with additional support provided for qualitative analysis and results.

Finally, during the period of this thesis, further work has been carried out and published, though these have not been addressed directly in this thesis:

- Gómez, E., Charisi, V., Tolan, S., Miron, M., Martinez, P. F., and **Escobar, M.** (2021). Human Behaviour and Machine Intelligence. In *Centre for Advanced Studies.*
- **Escobar-Planas, M.** (2022). Towards Trustworthy Conversational Agents for Children. In *Proceedings of the 21st Annual ACM Interaction Design and Children Conference (IDC'22)* (pp. 693-695).
- **Escobar-Planas, M.**, Gómez, E., and Martínez-Hinarejos, C. D. (2022). Enhancing the Design of a Conversational Agent for an Ethical Interaction with Children. *Proceedings IberSPEECH, 2022*, 171-175.
- **Escobar-Planas, M.**, Gómez, E., and Martínez-Hinarejos, C. D. (2023) From Ethical Guidelines to Practical Guidance to Develop Trustworthy Conversational Agents for Children. In *International Workshop on Spoken Dialogue Systems Technology.*
- Gaudeul, A., Arrigoni, O., Charisi, V., **Escobar-Planas, M.**, and Hupont-Torres, I. (2024) Understanding the Impact of Human Oversight on Discriminatory Outcomes in AI-Supported Decision Making. *Proceedings of the European Conference on Artificial Intelligence (ECAI'24).*

These contributions collectively demonstrate a robust engagement with both the theoretical and practical dimensions of developing trustworthy CAs for children. The research spans empirical studies, policy discussions, and ethical guideline development, and extends to broader explorations in ethical AI, offering a comprehensive approach to addressing the unique needs of child-CA interactions.

7.3 Complementary code and data

This thesis has generated a range of data and code that supports the development of trustworthy CAs for children. These resources are available for public access and aim to encourage further research and development in this field. There are two repositories containing complementary materials:

Guidelines development.

This repository contains data and analysis from the development of guidelines for trustworthy CAs for children, as presented at the Ethicomp conference (Escobar-Planas, Gómez, and Martínez-Hinarejos, 2022). It includes:

- Mapping Process: Documentation detailing the mapping between the HLEG's and UNICEF's ethical guidelines for AI.
- Expert Risk Evaluations: Data on experts' assessments of risk, including both likelihood and impact, as well as their comments on the applicability of ALTAI to CAs and children.
- Risk Evaluation Summary: Quantitative results of the expert's evaluation, highlighting the higher risks on Child-CA interaction.
- Summary of Expert Comments: Qualitative results categorised by the seven trustworthy requirements of the HLEG, offering a complete overview of critical comments distribution.
- Thematic Analysis Results: Detailed thematic analysis of expert feedback, categorised by the seven trustworthiness requirements.

https://github.com/mescpla/CAs4Children-ETHICOMP22.git

Collaborative Storytelling CAs

This repository contains the code and resources used to develop the CAs for the collaborative storytelling experiment presented in Chapter 6. The trustworthy versions of the CA incorporate the guidelines for trustworthy conversational agents for children introduced in Chapter 4, with their application in system development presented in Chapter 5. The system is adapted for different age groups and supports conversations in both Italian and Spanish.

- Generic Conversational Agent: A CA designed without special considerations for users under the age of 18.
- Trustworthy Conversational Agent for Children under 8: A version of the CA with additional safeguards and age-appropriate features tailored for children under 8.
- Trustworthy Conversational Agent for Children under 12: A version of the CA designed specifically to meet ethical guidelines for children younger than 12.
- Story Trees: Pre-designed story structures for collaborative storytelling activities, available in both Italian and Spanish, covering various topics suitable for different age groups.

https://github.com/ec-jrc/humaint-conversational-agent

7.4 Limitations

While this thesis has made significant contributions to the field of trustworthy CAs for children, several limitations must be acknowledged. These limitations arise from the study's experimental design, methodology, and the inherent complexities of working with children and AI systems.

First, regarding the user study design, the presence of supervisors and technicians during the sessions, while necessary to mitigate interaction difficulties, may have affected the naturalness of children's interactions with the CA. The novelty of our experiments required support from technicians, and the studies were not conducted in real-world settings, which limits the generalisability of the experiment's results to everyday contexts.

Second, the cultural context of our experiments and studies limits the broader applicability of our findings. In Chapter 3, children's interactions with an embodied CA were studied at a single private school in Spain, providing valuable insights but limiting generalisation to different cultural and socio-economic contexts. In Chapter 6, we expanded the scope by involving children from public schools in both Spain and Italy, yet this still restricts the cultural diversity of the participants. Moreover, the development of the guidelines presented in Chapter 4, while considering international frameworks like UNICEF's, was primarily centred on the HLEG's European guidelines and consultations with European experts. While full international research would have required more resources and time, we acknowledge this limitation and encourage future studies to validate these findings in non-European contexts, such as Asia or Africa, where cultural values, ethical concerns, and interaction styles with technology may differ significantly.

Additionally, other contextual factors also impacted our research. For instance, during the development of the trustworthy CA for collaborative storytelling in Italy (Chapter 5), ChatGPT was banned, leading us to opt for a "create-your-own-adventure" format. This approach, while useful, limited the freedom of open collaborative storytelling, which could have provided richer insights into children's interactions and further fostered their creativity.

Third, in terms of metrics, the risk assessment used in Chapter 4 treated children's risks and CA risks equally, which may have diluted the emphasis on children's specific needs. More weighted risk factors highlighting children's considerations could have provided a more focused evaluation.

The evaluation tool ALTAI was central to this research, but it also presented limitations. The high level of technical knowledge required for evaluation meant that developers were involved, and despite using the Delphi method to enhance robustness, this may have introduced some bias, such as being more inclined to perceive the modified system as more suitable for children's interaction (Chapter 5). Furthermore, through the thesis we have observed that ALTAI does not fully capture all considerations specific to children's needs, particularly around education and development. Future research should focus on creating more refined evaluation tools that consider children's unique requirements while minimising subjective influence from evaluators.

Communicating complex concepts such as expressivity (Chapter 3) or agency (Chapter 6) to children posed a challenge, as many children struggled to grasp these abstract ideas. Our explanations were insufficient to counter the halo effect, and, in our international study (Chapter 6), the additional clarifications during interviews may have unintentionally influenced children's responses regarding agency. These difficulties highlight the need for improved methods of communicating complex concepts in a child-friendly manner, aligned with their developmental stages.

The double consent mechanism also revealed unexpected results in our international user study (Chapter 6). While we anticipated different responses between younger and older children (6–7 vs. 10–11 years old), we did not expect the higher acceptance of supervision among older children. This limited our ability to explore alternative mechanisms that might be more suitable for teenagers. Further research is needed to investigate the appropriateness of double consent mechanisms for different age groups, particularly adolescents.

Finally, technical challenges were encountered, especially with the automatic speech recognition (ASR) module, which impacted the interaction experience, particularly for participants in Italy (Chapter 6). This highlights the need for more robust ASR technologies tailored to children's voices and multilingual contexts.

In sum, while this thesis provides valuable insights and advancements in developing trustworthy CAs for children, it is essential to acknowledge these limitations. Future research can build upon these findings to develop more comprehensive and globally applicable solutions.

7.5 Future work

Building on the findings and limitations of this thesis, there are several key areas where future research can deepen our understanding of trustworthy CAs for children and further enhance their development.

A key area of future research lies in explaining complex concepts to children, such as CA's expressivity, agency, and data processing, to counter the halo effect of these systems, in ways that are aligned with children's developmental stages.

- **Understanding children's perceptions of expressivity** in CAs require deeper investigation. While we observed that children tend to misinterpret or overestimate a CA's expressivity, further research is needed to explore this bias, ways of mitigation and possible repercussions.

- *How can we design a CA that is perceived as expressive only when intended?*
- *What are the repercussions of limiting CAs expressivity on children's behaviour and perception of the CA?*

- **Explaining agency** is another area for improvement. Children often perceive the CA as capable of independent action within its programming constraints. These perceptions raise questions about agency, specifically where intentionality meets autonomy, and whether children believe the system can act beyond its programming. Future research should explore strategies to enhance children's understanding of CA's lack of agency and, furthermore, its impact on data-sharing behaviour.

  - *How can enhanced explanations about a CA's lack of agency reduce children's inclination to disclose personal information?*

- Similarly, **improving children's understanding of privacy and data processing** is crucial for trustworthy interactions with CAs. Future work should focus on developing clear, child-friendly explanations about how data is processed and shared with third parties. Investigating how these explanations affect children's willingness to disclose personal information will provide valuable insights into designing more transparent and trustworthy CAs.

  - *What methods can effectively explain CA's data processing and privacy to children?*
  - *How do these explanations affect their behaviour regarding personal data disclosure?*

Another significant area involves **cross-cultural research**. While this thesis is created in an European context, it is essential to investigate how ethical considerations and children's interactions with CAs vary across different global regions. Future work should expand these studies to other regions, such as Asia and Africa, where cultural values and interaction styles may differ.

- *What specific considerations should be taken into account when designing trustworthy CAs for children in Asia and Africa?*
- *How can trustworthy guidelines be adapted for diverse cultural contexts?*

Additionally, there is a need for **new evaluation frameworks to complement ALTAI**, with a specific focus on children's needs and well-being. Future research could focus on the expansion of the current ALTAI framework to include new sections on education and personal development, focusing on areas like self-regulation, collaboration, and critical thinking (Sala et al., 2020), which are essential for child development.

- *What new items could be added to ALTAI to evaluate education and personal development considerations?*

Another promising line of inquiry involves the **potential of LLMs** in CAs to enhance children's agency and well-being. Incorporating advanced AI like LLMs may lead to more autonomous interactions, allowing children greater freedom and support in educational and personal development contexts. However, caution is advised with the outputs that LLMs provide to children, to ensure appropriateness.

- *How can LLMs be utilised to enhance children's autonomy and well-being in interactions with CAs?*
- *How can we ensure appropriate responses of LLMs to children?*

Exploring consent mechanisms is another important area for future study. While double consent mechanisms (involving both children and their guardians) have been effective for younger children, further research is needed to adapt these mechanisms for teenagers, ensuring their safe interaction with CAs while maintaining an appropriate level of independence.

- *How can dual consent mechanisms be adapted to ensure ethical engagement with CAs for teenagers?*

In terms of technical advancements, future work should focus on creating **artificial voices that are distinctly non-human** while maintaining a naturalistic tone. As voice synthesis technology improves, the challenge will be to design voices that balance naturalness with an obvious artificial identity, preventing children from misinterpreting the CA as a human-like entity.

- *How can we develop a voice for CAs that sounds natural yet is clearly identifiable as non-human?*

Further, **enhancing speech recognition for younger children and those speaking in less commonly supported languages** is crucial for inclusivity. Future research should investigate how improved automatic speech recognition (ASR) technologies can better interpret the unique speech patterns of young users and support a wider range of languages, thus enhancing the interaction quality, reducing user frustration and promoting inclusivity.

- *How can advancements in speech recognition technology improve interaction quality for younger children?*
- *What technological innovations are needed to support speech recognition for less commonly represented languages in CAs?*

In summary, future work in this field should continue exploring both the ethical and technical dimensions of child-CA interaction. By addressing these challenges, researchers can ensure that CAs become not only effective tools for education and development but also trustworthy companions for children.

Bibliography

ACM. (2018). FaccT [Available: https://facctconference.org/, accessed: 20/October/2024]. (Cit. on p. 16).

Agarwal, D., Agastya, A., Chaudhury, M., Dube, T., Jha, B., Khare, P., & Raghu, N. (2021). *Measuring effectiveness of chatbot to improve attitudes towards gender issues in underserved adolescent children in india* (tech. rep.). Cambridge, MA: Harvard Kennedy School. (Cit. on p. 116).

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in psychology*, *9*, 2136 (cit. on p. 112).

Ali, M. R., Razavi, S. Z., Langevin, R., Al Mamun, A., Kane, B., Rawassizadeh, R., Schubert, L. K., & Hoque, E. (2020). A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 1–8 (cit. on p. 14).

Altay, S., Hacquin, A.-S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on covid-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, *29*(1), 52 (cit. on p. 9).

Amazon. (2014). Amazon Alexa voice AI [Available: https://developer.amazon.com/alexa, accessed: 20/October/2024]. (Cit. on pp. 10, 70).

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., El Shafey, L., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., RObinson, K., ... Wu, Y. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (cit. on p. 8).

Apple. (2010). Siri-Apple [Available: https://www.apple.com/es/siri/, accessed: 20/October/2024]. (Cit. on pp. 10, 70).

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of informetrics*, *11*(4), 959–975 (cit. on p. 11).

Australian ai ethics framework. (2019). Department of Industry, Science; Resources. (Cit. on pp. 16, 113).

Bartlett, B., Estivill-Castro, V., & Seymon, S. (2004). Dogs or robots: Why do children see them as robotic pets rather than canine machines? *Proceedings of the fifth conference on Australasian user interface-Volume 28*, 7–14 (cit. on p. 34).

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, *3*(21) (cit. on p. 34).

Beneteau, E., Boone, A., Wu, Y., Kientz, J. A., Yip, J., & Hiniker, A. (2020). Parenting with alexa: Exploring the introduction of smart speakers on family dynamics. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on p. 13).

Bethel, C. L., Stevenson, M. R., & Scassellati, B. (2011). Secret-sharing: Interactions between a child, robot, and adult. *2011 IEEE International Conference on systems, man, and cybernetics*, 2489–2494 (cit. on p. 93).

Brabra, H., Báez, M., Benatallah, B., Gaaloul, W., Bouguelia, S., & Zamanirad, S. (2021). Dialogue management in conversational systems: A review of approaches, challenges, and opportunities. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 783–798 (cit. on p. 8).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa (cit. on pp. 37, 56, 98).

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597 (cit. on p. 37).

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. *Springer handbook of robotics*, 1935–1972 (cit. on p. 37).

Brey, T., Hanrieder, G., Heisterkamp, P., Hitzenberger, L., & Regel-Brietzmann, P. (2000). Issues in the evaluation of spoken dialogue systems-experience from the access project. *In International Conference on Language Resources and Evaluation* (cit. on p. 22).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, s., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901 (cit. on pp. 1, 8, 15, 21).

Catania, F., Crovari, P., Beccaluva, E., De Luca, G., Colombo, E., Bombaci, N., & Garzotto, F. (2021). Boris: A spoken conversational agent for music production for people with motor disabilities. *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 1–5 (cit. on p. 13).

Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar Planas, M., Sanchez, M. J. I., & Gomez Gutierrez, E. (2022). *Artificial intelligence and the rights of the child: Towards an integrated agenda for research and policy* (tech. rep.). Joint Research Centre (Seville site). (Cit. on pp. 2, 21).

Charisi, V., Davison, D. P., Wijnen, F. M., Reidsma, D., & Evers, V. (2017). Measuring children's perceptions of robots' social competence: Design and validation. *Social Robotics: 9th International Conference, ICSR 2017,* *Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, 676–686 (cit. on p. 33).

Chen, L., Chen, Z., Tan, B., Long, S., Gašić, M., & Yu, K. (2019). Agentgraph: Toward universal dialogue management with structured deep reinforcement learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(9), 1378–1391 (cit. on p. 8).

Chhibber, N., & Law, E. (2019). Using conversational agents to support learning by teaching. *arXiv preprint arXiv:1909.13443* (cit. on p. 10).

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649 (cit. on p. 2).

Chubb, J., Missaoui, S., Concannon, S., Maloney, L., & Walker, J. A. (2022). Interactive storytelling for children: A case-study of design and development considerations for ethical conversational ai. *International Journal of Child-Computer Interaction*, *32*, 100403 (cit. on p. 21).

Clarivate. (1997). Web of Science [Available: https://www.webofscience.com/, accessed: 20/October/2024]. (Cit. on p. 11).

Cuayáhuitl, H. (2017). Simpleds: A simple deep reinforcement learning dialogue system. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, 109–118 (cit. on p. 8).

Davison, D. P., Wijnen, F. M., Charisi, V., van der Meij, J., Reidsma, D., & Evers, V. (2021). Words of encouragement: How praise delivered by a social robot changes children's mindset for learning. *Journal on multimodal user interfaces*, *15*(1), 61–76 (cit. on pp. 33, 94).

Del-Moral-Pérez, M. E., Villalustre-Martínez, L., & Neira-Piñeiro, M. d. R. (2019). Teachers' perception about the contribution of collaborative creation of digital storytelling to the communicative and digital competence in primary education schoolchildren. *Computer Assisted Language Learning*, *32*(4), 342–365 (cit. on p. 70).

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*, 755–810 (cit. on p. 22).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (cit. on p. 7).

Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, *23*(1), 96–138 (cit. on pp. 1, 21).

Dignum, V., Penagos, M., Pigmans, K., & Vosloo, S. (2021). *Policy guidance on ai for children*. Communications of UNICEF. (Cit. on pp. 2, 19, 113).

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). " hey google is it ok if i eat you?" initial explorations in child-agent interaction. *Proceedings of the 2017 conference on interaction design and children*, 595–600 (cit. on p. 15).

Durden, E., Pirner, M. C., Rapoport, S. J., Williams, A., Robinson, A., & Forman-Hoffman, V. L. (2023). Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "woebot". *Internet Interventions*, *33*, 100637 (cit. on p. 9).

Dybkjaer, L., Bernsen, N. O., & Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, *43*(1-2), 33–54. https://doi.org/https://doi.org/10.1016/j.specom.2004.02.001 (cit. on p. 22).

Elgarf, M. (2022). *Child-robot behavioral alignment and creativity performance* [Doctoral dissertation, KTH Royal Institute of Technology]. (Cit. on p. 14).

Elgarf, M., Zojaji, S., Skantze, G., & Peters, C. (2022). Creativebot: A creative storyteller robot to stimulate creativity in children. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 540–548 (cit. on pp. 14, 114).

Elsevier. (2004). Scopus [Available: https://www.scopus.com/, accessed: 20/October/2024]. (Cit. on p. 11).

Engebak, I.-M. H. (2019). *A digital game using collaborative storytelling to help children practice empathy* [Master's thesis, NTNU]. (Cit. on p. 70).

Escobar-Planas, M., Charisi, V., & Gómez, E. (2022). "that robot played with us!" children's perceptions of a robot after a child-robot group interaction. *ACM Human-Computer Interaction (CSCW2)*, *6*(393). https://doi.org/10.1145/3555118 (cit. on p. 96).

Escobar-Planas, M., & Frau-Amar, P. (2022). faster-whisper [https://github.com/mescpla/CAs4Children-ETHICOMP22.git, accessed: 20/October/2024]. (Cit. on p. 65).

Escobar-Planas, M., Gómez, E., & Martínez-Hinarejos, C.-D. (2022). Guidelines to develop trustworthy conversational agents for children, 342–360 (cit. on pp. 20, 126).

Escobar-Planas, M., Gómez, E., & Martınez-Hinarejos, C.-D. (2022). Enhancing the design of a conversational agent for an ethical interaction with children. *Proc. IberSPEECH 2022*, 171–175. https://doi.org/10.21437/IberSPEECH.2022-35 (cit. on p. 70).

European Commission, Joint Research Centre (JRC.T.3). (2019). Jrc human behaviour and machine intelligence (humaint) [https://ec.europa.eu/dpo-register/detail/DPR-EC-01094.4, accessed: 20/October/2024]. (Cit. on p. 92).

Fast, E., Chen, B., Mendelsohn, J., Bassen, J., & Bernstein, M. S. (2018). Iris: A conversational agent for complex tasks. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12 (cit. on p. 7).

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR mental health*, *4*(2), e7785 (cit. on p. 14).

Fraser, J., Papaioannou, I., & Lemon, O. (2018). Spoken conversational ai in video games: Emotional dialogue management increases user engagement. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 179–184 (cit. on p. 13).

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS)*, *14*(3), 330–347 (cit. on p. 16).

Ganguli, D., Askell, A., Schiefer, N., Liao, T., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., Drain, D., Li, D., Tran-Johnson, E., Perez, E., Kernion, J., Kerr, J., Mueller, J., Landau, J., Ndousse, K., . . . Kaplan, J. (2023). The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459* (cit. on p. 11).

Garg, R., & Sengupta, S. (2020). He is just like me: A study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), 1–24 (cit. on pp. 1, 10).

Girouard-Hallam, L. N., Streble, H. M., & Danovitch, J. H. (2021). Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, *3*(5), 1118–1131 (cit. on pp. 15, 112).

Gomez, R., Szapiro, D., Galindo, K., & Nakamura, K. (2018). Haru: Hardware design of an experimental tabletop robot assistant. *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 233–240 (cit. on p. 28).

Google. (2016). Google assistant: your personal assistant [Available: https://assistant.google.com/, accessed: 20/October/2024]. (Cit. on pp. 10, 70).

Google, G. T. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, *1* (cit. on p. 8).

Google Cloud. (n.d.). Google text-to-speech api reference [https://cloud.google.com/text-to-speech/docs/libraries?hl=es-419, accessed: 20/October/2024]. (Cit. on p. 72).

Google Cloud. (2017). Dialogflow api reference [https://cloud.google.com/dialogflow/es/docs/reference, accessed: 20/October/2024]. (Cit. on p. 72).

Haas, M., & Keller, A. (2021). 'alexa, adv (ert) ise us!': How smart speakers and digital assistants challenge advertising literacy amongst young people. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *43*, 19–40 (cit. on p. 16).

Heaney, D. (2021). This openai gpt-3 powered demo is a glimpse of npcs in the future [https://uploadvr.com/modbox-gpt3-ai-npc-demo/, accessed 15/May/2024]. (Cit. on p. 10).

HLEG. (2019). Ethics guidelines for trustworthy ai. *B-1049 Brussels* (cit. on pp. 16, 17, 54, 113).

HLEG. (2020). *The assessment list for trustworthy artificial intelligence (altai)*. European Commission. (Cit. on p. 23).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780 (cit. on p. 8).

Hopf, Y. M., Bond, C. B., Francis, J. J., Haughney, J., & Helms, P. J. (2014). Linked health data for pharmacovigilance in children: Perceived legal and ethical issues for stakeholders and data guardians. *BMJ open*, *4*(2), e003875 (cit. on p. 66).

James, B., Joseph, D., & Sharma, T. (2024). Transforming banking services: Ai-driven e-loyalty strategies and case study insights on customer satisfaction and loyalty enhancement. *Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security, IACIDS 2023, 23-25 November 2023, Lavasa, Pune, India* (cit. on p. 9).

Jaradat, G. A., Alzubaidi, M. A., & Otoom, M. (2022). A novel human-vehicle interaction assistive device for arab drivers using speech recognition. *IEEE Access*, *10*, 127514–127529 (cit. on p. 9).

John, R. J. L., Potti, N., & Patel, J. M. (2017). Ava: From data to insights through conversations. *CIDR* (cit. on p. 7).

Johns, Tim. (2021). Parents of children called alexa challenge amazon [https://www.bbc.com/news/technology-57680173, accessed: 20/October/2024]. (Cit. on p. 16).

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2012). "robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, *48*(2), 303 (cit. on p. 15).

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 82–90 (cit. on p. 14).

Kovacevic, N., Stojiljkovic, A., & Kovac, M. (2019). Application of the matrix approach in risk assessment. *Oper. Res. Eng. Sci. Theory Appl*, *2*(3), 55–64 (cit. on p. 54).

Kudina, O., & Coeckelbergh, M. (2021). "alexa, define empowerment": Voice assistants at home, appropriation and technoperformances. *Journal of Information, Communication and Ethics in Society*, *19*(2), 299–312 (cit. on p. 16).

Landoni, M., Murgia, E., Huibers, T., & Pera, M. S. (2020). You've got a friend in me: Children and search agents. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 89–94 (cit. on p. 13).

Latitude. (2019 (accessed August 24, 2024)). AI Dungeon [Available: https://aidungeon.com/, accessed: 20/October/2024]. (Cit. on p. 9).

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656* (cit. on p. 15).

Law, E., Baghaei Ravari, P., Chhibber, N., Kulic, D., Lin, S., Pantasdo, K. D., Ceha, J., Suh, S., & Dillen, N. (2020). Curiosity notebook: A platform for learning by teaching conversational agents. *Extended Abstracts of the 2020*

CHI Conference on Human Factors in Computing Systems, 1–9 (cit. on p. 13).

Lee, Y., Kim, T. S., Chang, M., & Kim, J. (2022). Interactive children's story rewriting through parent-children interaction. *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, 62–71 (cit. on p. 14).

Lee, Y., Kim, T. S., Kim, S., Yun, Y., & Kim, J. (2023). Dapie: Interactive step-by-step explanatory dialogues to answer children's why and how questions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–22 (cit. on p. 13).

Lehnert, F. K., Niess, J., Lallemand, C., Markopoulos, P., Fischbach, A., & Koenig, V. (2022). Child–computer interaction: From a systematic review towards an integrated understanding of interaction design methods for children. *International Journal of Child-Computer Interaction*, *32*, 100398 (cit. on p. 2).

Leviathan, Y., & Matias, Y. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone [Available: https://blog.research.google/2018/05/duplex-ai-system-for-natural-conversation.html, accessed: 20/October/2024]. (Cit. on p. 9).

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology* (cit. on pp. 55, 80).

Lin, S.-C., Hsu, C.-H., Talamonti, W., Zhang, Y., Oney, S., Mars, J., & Tang, L. (2018). Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 531–542 (cit. on p. 9).

Linstone, H. A., & Turoff, M. (1975). *The delphi method*. Addison-Wesley Reading, MA. (Cit. on pp. 54, 80).

Liu, C.-C., Liao, M.-G., Chang, C.-H., & Lin, H.-M. (2022). An analysis of children'interaction with an ai chatbot and its impact on their interest in reading. *Computers & Education*, *189*, 104576 (cit. on p. 13).

Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019). Hey google, do unicorns exist? conversational agents as a path to answers to children's questions. *Proceedings of the 18th ACM international conference on interaction design and children*, 301–313 (cit. on pp. 1, 10).

Lumivero. (1997). Nvivo [https://lumivero.com/products/nvivo/, accessed: 20/October/2024]. (Cit. on p. 97).

Lupetti, M. L., Hagens, E., Van Der Maden, W., Steegers-Theunissen, R., & Rousian, M. (2023). Trustworthy embodied conversational agents for healthcare: A design exploration of embodied conversational agents for the periconception period at erasmus mc. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–14 (cit. on p. 21).

Madiega, T. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (cit. on pp. 2, 18).

Mathur, N., Dhodapkar, K., Zubatiy, T., Li, J., Jones, B., & Mynatt, E. (2022). A collaborative approach to support medication management in older adults with mild cognitive impairment using conversational assistants (cas). *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–14 (cit. on p. 9).

McTear, M. (2020). Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, *13*(3), 1–251 (cit. on pp. 1, 6).

Mehri, S., & Eskenazi, M. (2020). Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*, 225–235 (cit. on p. 22).

Mohit, B. (2014). Named entity recognition. In *Natural language processing of semitic languages* (pp. 221–245). Springer. (Cit. on p. 7).

Monarca, I., Cibrian, F. L., Mendoza, A., Hayes, G., & Tentori, M. (2020). Why doesn't the conversational agent understand me? a language analysis of children speech. *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*, 90–93 (cit. on pp. 15, 115).

Narayanan, S., & Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, *10*(2), 65–78 (cit. on p. 14).

Nasihati Gilani, S., Traum, D., Merla, A., Hee, E., Walker, Z., Manini, B., Gallagher, G., & Petitto, L.-A. (2018). Multimodal dialogue management for multiparty interaction with infants. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 5–13 (cit. on p. 13).

Nasir, J., Oppliger, P., Bruno, B., & Dillenbourg, P. (2022). Questioning wizard of oz: Effects of revealing the wizard behind the robot. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1385–1392 (cit. on p. 11).

Nations, U. (2015). Sustainable Development Goals [Available: https://sustainabledevelopment.un.org/?menu=1300, accessed: 20/October/2024]. (Cit. on p. 18).

Nichols, E., Gao, L., Vasylkiv, Y., & Gomez, R. (2021). Collaborative storytelling with social robots. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1903–1910 (cit. on p. 14).

Nilsen, J. H. (2019). *"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]. (Cit. on p. 14).

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A review of deep learning based speech synthesis. *Applied Sciences*, *9*(19), 4050 (cit. on p. 8).

NLTK, T. (2001). Natural Language Toolkit [Available: https://www.nltk.org/, accessed: 20/October/2024]. (Cit. on p. 37).

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847 (cit. on p. 98).

Ong, D. T., De Jesus, C. R., Gilig, L. K., Alburo, J. B., & Ong, E. (2018). A dialogue model for collaborative storytelling with children. *ICCE 2018 -*

26th International Conference on Computers in Education, Main Conference Proceedings, 205–210 (cit. on pp. 14, 115).

Ong, E., Alburo, J. B., De Jesus, C. R., Gilig, L. K., & Ong, D. T. (2019). Challenges posed by voice interface to child-agent collaborative storytelling. *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 1–6 (cit. on pp. 14, 115).

OpenAI. (2022). ChatGPT [Available: https://chatgpt.com/, accessed: 20/October/2024]. (Cit. on p. 6).

OpenAI. (2024a). Hello GPT-4o [Available: https://openai.com/index/hello-gpt-4o/, accessed: 24/August/2024]. (Cit. on p. 8).

OpenAI. (2024b). Introducing APIs for GPT-3.5 Turbo and Whisper [Available: https://openai.com/blog/introducing-chatgpt-and-whisper-apis, accessed: 20/October/2024]. (Cit. on p. 71).

Packard, E. (1979). The cave of time, ser. choose your own adventure. *Bantam Books* (cit. on p. 72).

Papadopoulos, F., Dautenhahn, K., & Ho, W. C. (2013). Aibostory–autonomous robots supporting interactive, collaborative story-telling. *Paladyn, Journal of Behavioral Robotics*, *4*(1), 10–22 (cit. on p. 14).

Peca, A., Simut, R., Pintea, S., Costescu, C., & Vanderborght, B. (2014). How do typically developing children and children with autism perceive different social robots? *Computers in Human Behavior*, *41*, 268–277 (cit. on p. 34).

Pérez, G., Rodríguez, T., Bachiller, P., Bustos, P., & Núñez, P. (2022). Introducing the social robot ebo: An interactive and socially aware storyteller robot for therapies with older adults. *International Conference on Social Robotics*, 462–472 (cit. on p. 14).

Pradhan, A., Mehta, K., & Findlater, L. (2018). " accessibility came by accident" use of voice-controlled intelligent personal assistants by people with disabilities. *Proceedings of the 2018 CHI Conference on human factors in computing systems*, 1–13 (cit. on p. 13).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning*, 28492–28518 (cit. on p. 6).

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*, 25–36 (cit. on p. 22).

Rajamäki, J., Gioulekas, F., Rocha, P. A. L., Garcia, X. d. T., Ofem, P., & Tyni, J. (2023). Altai tool for assessing ai-based technologies: Lessons learned and recommendations from shapes pilots. *Healthcare*, *11*(10), 1454 (cit. on p. 23).

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (cit. on p. 8).

Ristić, D. (2013). A tool for risk assessment. *safety Engineering*, *3*(3), 121–127 (cit. on p. 54).

Robb, D. A., Lopes, J., Ahmad, M. I., McKenna, P. E., Liu, X., Lohan, K. S., & Hastie, H. F. (2023). Seeing eye to eye: Trustworthy embodiment for task-based conversational agents. *Frontiers in Robotics and AI*, *10*, 1234767 (cit. on p. 21).

Røyneland, K. (2019). *" it knows how to not understand us!" a study on what the concept of robustness entails in design of conversational agents for preschool children* [Master's thesis]. (Cit. on p. 15).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536 (cit. on p. 8).

Sabour, S., Zhang, W., Xiao, X., Zhang, Y., Zheng, Y., Wen, J., Zhao, J., & Huang, M. (2023). A chatbot for mental health support: Exploring the impact of emohaa on reducing mental distress in china. *Frontiers in digital health*, *5*, 1133987 (cit. on p. 9).

Saha, T., Reddy, S., Das, A., Saha, S., & Bhattacharyya, P. (2022). A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2436–2449 (cit. on p. 9).

Sala, A., Punie, Y., Garkov, V., & Cabrera, M. (2020). Lifecomp: The european framework for personal, social and learning to learn key competence: European commission (cit. on pp. 65, 131).

Santhanam, S., & Shaikh, S. (2019). A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500* (cit. on p. 8).

Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology*, *13*, 787499 (cit. on p. 113).

Schuurmans, J., & Frasincar, F. (2019). Intent classification for dialogue utterances. *IEEE Intelligent Systems*, *35*(1), 82–88 (cit. on p. 7).

Schwartz, B. (2020 (accessed May 15, 2024)). Google my business using duplex to call businesses to see hour changes or closures [https://www.seroundtable.com/google-duplex-changes-covid-19-29165.html]. (Cit. on p. 9).

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). " hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage. *Proceedings of the 2018 designing interactive systems conference*, 857–868 (cit. on pp. 1, 10).

Seymour, W., Zhan, X., Cote, M., & Such, J. (2023). A systematic review of ethical concerns with voice assistants. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 131–145 (cit. on p. 21).

Sezgin, E., Noritz, G., Elek, A., Conkol, K., Rust, S., Bailey, M., Strouse, R., Chandawarkar, A., von Sadovszky, V., Lin, S., & Huang, Y. (2020). Capturing at-home health and care information for children with medical complexity using voice interactive technologies: Multi-stakeholder viewpoint. *Journal of medical Internet research*, *22*(2), e14202 (cit. on p. 14).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779–4783 (cit. on p. 8).

Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020). Effects of persuasive dialogues: Testing bot identities and inquiry strategies. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 114).

Siau, K., & Wang, W. (2020). Artificial intelligence (ai) ethics: Ethics of ai and ethical ai. *Journal of Database Management (JDM)*, *31*(2), 74–87 (cit. on p. 2).

Slosiarová, N., Mesarčík, M., Jurkáček, P., & Podroužek, J. (2023). Trustworthy ai in dental care beyond artificial intelligence act (cit. on p. 23).

Smith, C., Crook, N., Charlton, D., Boye, J., De La Camara, R. S., Turunen, M., Benyon, D., Gambäck, B., Mival, O., Webb, N., et al. (2011). Interaction strategies for an affective conversational agent. *Presence*, *20*(5), 395–411 (cit. on p. 7).

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/https://doi.org/10.1016/j.jbusres.2019.07.039 (cit. on p. 11).

Sobti, R., Guleria, K., & Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications*, 1–63 (cit. on p. 14).

Sousa, R. G., Ferreira, P. M., Costa, P. M., Azevedo, P., Costeira, J. P., Santiago, C., Magalhaes, J., Semedo, D., Ferreira, R., Rudnicky, A. I., &

Hauptmann, A. G. (2021). Ifetch: Multimodal conversational agents for the online fashion marketplace. *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI*, 25–26 (cit. on p. 9).

Stahl, B. C., & Leach, T. (2023). Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: An empirical test of the european union assessment list for trustworthy ai (altai). *AI and Ethics*, *3*(3), 745–767 (cit. on p. 23).

Straten, C. L. v., Peter, J., Kühne, R., & Barco, A. (2020). 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 (cit. on pp. 15, 66, 79, 113, 114).

Su, P.-H., Gasic, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Wen, T.-H., & Young, S. (2016). Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689* (cit. on p. 8).

Su, Z., Liang, F., Do, H. M., Bishop, A., Carlson, B., & Sheng, W. (2021). Conversation-based medication management system for older adults using a companion robot and cloud. *IEEE Robotics and Automation Letters*, *6*(2), 2698–2705 (cit. on p. 9).

Sun, M., Leite, I., Lehman, J. F., & Li, B. (2017). Collaborative storytelling between robot and child: A feasibility study. *Proceedings of the 2017 Conference on Interaction Design and Children*, 205–214. https://doi.org/10.1145/3078072.3079714 (cit. on p. 14).

Sun, Y., Liu, J., Yao, B., Chen, J., Wang, D., Ma, X., Lu, Y., Xu, Y., & He, L. (2024). Exploring parent's needs for children-centered ai to support preschoolers' storytelling and reading activities. *arXiv preprint arXiv:2401.13804* (cit. on p. 113).

SYSTRAN. (2023). faster-whisper [https://github.com/guillaumekln/faster-whisper, accessed: 20/October/2024]. (Cit. on p. 71).

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Zhao, S., Qin, T., Soong, F., & Tie-Yan, L. (2024). Naturalspeech: End-to-end text-to-speech synthesis with human-level quality.

IEEE Transactions on Pattern Analysis and Machine Intelligence (cit. on p. 8).

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., & Bernstein, A. (2021). Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–7 (cit. on p. 15).

Totakura, V., Vuribindi, B. R., & Reddy, E. M. (2021). Improved safety of self-driving car using voice recognition through cnn. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012079 (cit. on p. 9).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Biken, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurill, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (cit. on p. 8).

Turner, S., Pattnaik, N., Nurse, J. R., & Li, S. (2022). " you just assume it is in there, i guess": Uk families' application and knowledge of smart home cyber security. *arXiv preprint arXiv:2206.03930* (cit. on pp. 15, 114).

UNESCO, C. (2021). Recommendation on the ethics of artificial intelligence. (Cit. on p. 16).

UNICEF, Gomez, R., & Charisi, V. (2021). *Unicef pilot study on policy guidance for ai and child's rights*. Office of Global Insight; Policy. (Cit. on p. 21).

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). Chatgpt: Five priorities for research. *Nature*, *614*(7947), 224–226 (cit. on p. 11).

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, *12* (cit. on p. 8).

Vaswani, A. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762* (cit. on pp. 7, 8).

von Struensee, S. (2021). Eye on developments in artificial intelligence and children's rights: Artificial intelligence in education (aied), edtech, surveillance, and harmful content. *EdTech, Surveillance, and Harmful Content (June 4, 2021)* (cit. on p. 66).

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 271–280 (cit. on p. 22).

Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022). 'don't make assumptions about me!': Understanding children's perception of datafication online. *Proceedings of the ACM on Human-Computer InteractionVolume 6Issue CSCW2*, (419), 1–24 (cit. on pp. 15, 114).

Wang, Z., & Lemon, O. (2013). A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. *Proceedings of the SIGDIAL 2013 Conference*, 423–432 (cit. on p. 7).

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780 (cit. on p. 7).

Wei, W., Le, Q., Dai, A., & Li, J. (2018). Airdialogue: An environment for goal-oriented dialogue research. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3844–3854 (cit. on p. 9).

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45 (cit. on p. 6).

Westlund, J. K., Breazeal, C., & Story, A. (2015). Deception, secrets, children, and robots: What's acceptable. *Workshop on The Emerging Policy and Ethics of Human-Robot Interaction, held in conjunction with the 10th ACM/IEEE International Conference on Human-Robot Interaction* (cit. on p. 34).

Wijnen, F. M., Davison, D. P., Reidsma, D., Meij, J. V. D., Charisi, V., & Evers, V. (2019). Now we're talking: Learning by explaining your reasoning to a social robot. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(1), 1–29 (cit. on p. 96).

Williams, R., Machado, C. V., Druga, S., Breazeal, C., & Maes, P. (2018). "my doll says it's ok" a study of children's conformity to a talking doll. *Proceedings of the 17th ACM Conference on Interaction Design and Children*, 625–631 (cit. on p. 116).

Wong, P. N., Brumby, D. P., Babu, H. V. R., & Kobayashi, K. (2019). Voices in self-driving cars should be assertive to more quickly grab a distracted driver's attention. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 165–176 (cit. on p. 9).

Wu, C.-S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R., & Fung, P. (2019). Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743* (cit. on p. 8).

Xu, W., Ma, J., Yao, J., Lin, W., Zhang, C., Xia, X., Zhuang, N., Weng, S., Xie, X., Feng, S., Ying, F., Hansen, P., & Yao, C. (2023). Mathkingdom: Teaching children mathematical language through speaking at home via a voice-guided game. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on p. 13).

Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y.-S., & Warschauer, M. (2022). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development*, *93*(2), e149–e167 (cit. on p. 13).

Xu, Y., & Warschauer, M. (2020a). " elinor is talking to me on the screen!" integrating conversational agents into children's television programming. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8 (cit. on p. 13).

Xu, Y., & Warschauer, M. (2020b). Exploring young children's engagement in joint reading with a conversational agent. *Proceedings of the interaction design and children conference*, 216–228 (cit. on p. 13).

Xu, Y., & Warschauer, M. (2020c). What are you talking to?: Understanding children's perceptions of conversational agents. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on pp. 15, 112).

Yeh, Y.-T., Eskenazi, M., & Mehri, S. (2021). A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*, 15–33 (cit. on p. 22).

Zhang, L., Weitlauf, A. S., Amat, A. Z., Swanson, A., Warren, Z. E., & Sarkar, N. (2020). Assessing social communication and collaboration in autism spectrum disorder using intelligent collaborative virtual environments. *Journal of autism and developmental disorders*, *50*, 199–211 (cit. on p. 14).

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosemberg, A., Prabhavalkar, R., Park, D. S., Hagnai, P., Riesa, J., Perng, G., Soltau, H., … Wu, Y. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037* (cit. on p. 6).

Zhang, Z., Xu, Y., Wang, Y., Yao, B., Ritchie, D., Wu, T., Yu, M., Wang, D., & Li, T. J.-J. (2022). Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–21 (cit. on pp. 14, 70).

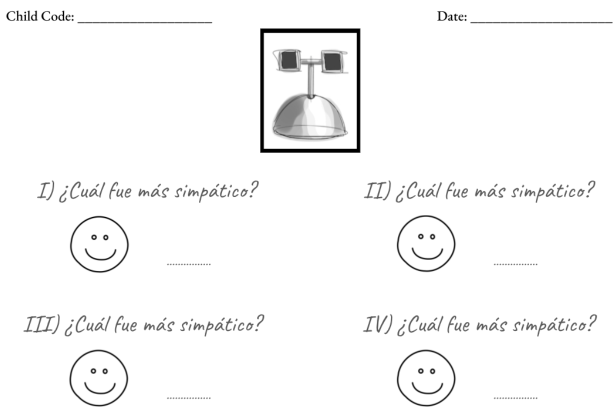
Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen Zhipeng Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., … Ji-Rong, W. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223* (cit. on p. 1).

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Sturtz, N., Tolle, K., Tihi, J. J., Van Halem, I., & Westerlinf, M. (2021). Z-inspection®: A process to assess trustworthy ai. *IEEE Transactions on Technology and Society*, *2*(2), 83–97 (cit. on p. 23).

Zou, J., Gauthier, S., Anzalone, S. M., Cohen, D., & Archambault, D. (2022). A wizard of oz interface with qtrobot for facilitating the handwriting learning in children with dysgraphia and its usability evaluation. *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part II*, 219–225 (cit. on p. 11).

Appendix A

Child-Robot Interaction questionnaires



Child Code: ________________ Date: ________________

I) ¿Cuál fue más simpático?

II) ¿Cuál fue más simpático?

III) ¿Cuál fue más simpático?

IV) ¿Cuál fue más simpático?

Figure A.1: Pre-manipulation check questionnaire where children identified the robot's expressive behaviour. Two behaviours were shown, and the experimenter asked: *Which one was more sympathetic?*' (informal Spanish concept understandable by children). This process was repeated four times to complete the questionnaire.

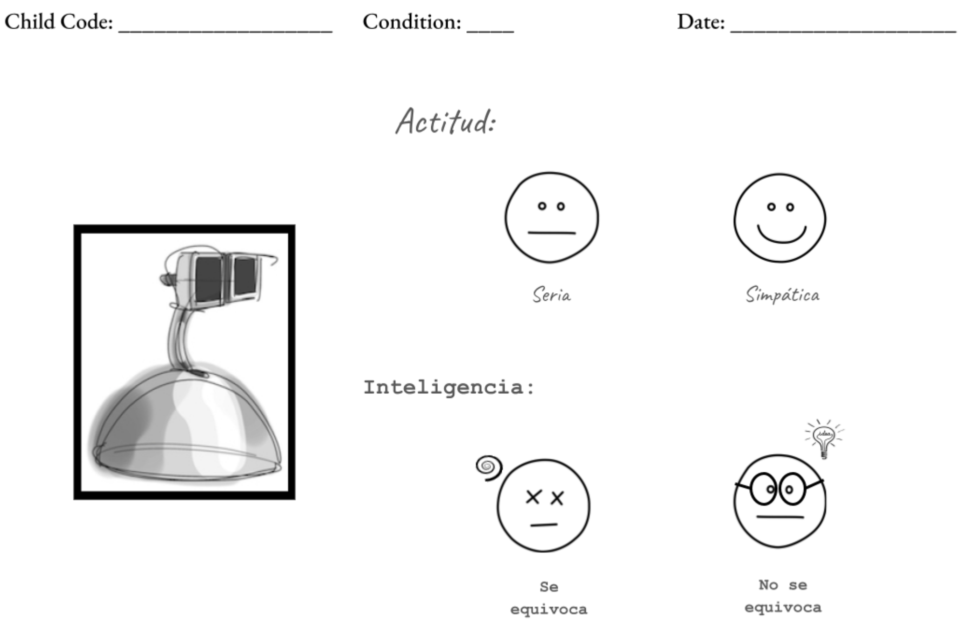


Figure A.2: Manipulation check questionnaire where children reported their perceptions of the robot's behaviour. After the problem-solving task, the experimenter asked about the robot's 'attitude' ('sympathetic' or 'serious') and 'intelligence' ('never making mistakes' or 'sometimes making mistakes'). Facial expressions were used to aid understanding, with balanced designs to minimise influence on children's choices.