



Engagement of Parents with the Aurora Child Health Chatbot: A Conversation Log Analysis Study

Diana Liebetrau^{a,b} , Melissa Densmore^a , and Francisco Nunes^b

^aComputer Science, University of Cape Town, Rondebosch, South Africa; ^bFraunhofer Portugal AICOS, Porto, Portugal

ABSTRACT

Chatbots have the potential to support child health by answering parents doubts and providing tailored information. However, prior work has not studied the deployment of chatbots for this setting. We analysed how parents used the Aurora Facebook Messenger chatbot, designed for Portuguese parents, with an optional subscription to professional support. Our analysis investigated chatbot use, discussed topics, and conversation topics, drawing on user engagement and conversation metrics, text-mining, user satisfaction scores, and conversation content analysis. Results revealed 718 active users (out of 1043), with peak activity during lunchtime and late at night. Most queries pertained to critical situations, including infant sleep (80%), (breast)feeding (13%), or healthcare-related issues (7%). Aurora handled in-domain questions appropriately, but struggled to answer multi-topic queries. Subscription users had 243% more interactions and 162% more extended use of the chatbot. Our research underscores the importance of offering timely and personalised messaging to meet parents' needs.

KEYWORDS

conversation log analysis;
chatbot; parenting chatbot;
data mining; healthcare
chatbot; mHealth

CCS CONCEPTS

Human-centered computing
→ Empirical studies in HCI;
Applied computing →
Health informatics

1. Introduction

A chatbot, often referred to as a conversational agent, is an Artificial Intelligence (AI) program designed to interact with users through human-like conversation (Sanguinetti et al., 2020). The capability of chatbots to provide uninterrupted support and reduce the demand for human resources has generated significant interest in multiple sectors, making them a crucial tool for customer support or user management (Cancel & Gerhardt, 2019; Higashinaka et al., 2014). In healthcare, the potential of chatbots is underscored by recent advances in machine learning (ML), which enhance the ability of conversational agents to deal with complex dialogues and conversational flexibility (Siddique & Chow, 2021). The exponential growth of chatbot solutions spans various applications, including addressing patient health queries, updating health records, and disseminating disease-specific information (Frangoudes et al., 2021; Siddique & Chow, 2021; Xu et al., 2021). Despite the availability of new solutions, there is limited literature on using chatbots in real-world healthcare settings (Laranjo et al., 2018; Vaira et al., 2018). Concerns regarding accuracy, empathy, and the need of human touch remain prominent impediments to broader implementation of chatbots (Go & Sundar, 2019; Rapp et al., 2021), particularly in sensitive areas like child health (Entenberg et al., 2021). Large Language Models (LLMs) like ChatGPT revolutionize the interaction with a chatbot due to their superior ability to manage diverse user queries (Chakraborty et al., 2023), yet, recent studies have

also highlighted ethical concerns associated with the deployment of LLM chatbots because of their unbalanced training data which can lead to misdiagnosis and lower care equity (Chow & Li, 2024).

In child health, chatbots have the potential to deliver timely information that caters to parents' learning needs (Agarwal & Wadhwa, 2020; Coleman et al., 2023). By filtering through information and personalising it to the parents' and children's requirements, chatbots can support parents in learning to care for their children (Agarwal & Wadhwa, 2020). In fact, while medical experts, social media, parenting blogs, and baby care apps like BabyCentre (BabyCenter, 2023) or BabySparks (BabySparks, 2021) provide child health information, they do not enable parents to ask personalised questions about their infants 24/7. Having said this, there are no studies reporting on the use of child health chatbots and only a few (general) design guidelines for creating chatbots for healthcare settings (Entenberg et al., 2023). With the rapid advancement of chatbot technology, particularly hybrid models that combine AI-enabled interactions with human oversight, there remains a significant need for research on ethical implications and efficacy in sensitive healthcare topics (Chow & Li, 2024; Siddique & Chow, 2021).

Hybrid chatbots have demonstrated potential in various health contexts. Recent research on hybrid chatbots, such as Ginger.io (Shih et al., 2022) and Feedpal (Yadav et al., 2019), highlights the enhanced efficacy of combining automated responses with human support. These hybrid models

leverage AI to provide immediate data-driven responses, while allowing for human intervention to address more complex or nuanced queries. Studies have shown that such models can significantly improve user engagement and satisfaction, particularly in sensitive areas like mental health and breastfeeding support (Shih et al., 2022; Yadav et al., 2019). These findings suggest that hybrid models could be particularly beneficial in fields requiring nuanced and empathetic communication, such as child health.

To better understand the potential of chatbots to support child health, we studied the Aurora Facebook Messenger (hybrid) chatbot (referred from now on as Aurora) (Aurora Tech AI, 2023), created in 2018 for Portuguese parents. We analyzed 1043 unique Aurora conversation logs to explore user engagement patterns, identify queried topics, highlight topics outside the chatbot's knowledge domain, and assess the dynamics and quality of the conversations. We analyzed Aurora's chatlogs using a mixed-method approach involving the quantitative analysis of engagement metrics, topic distribution, and content analysis. User conversations mainly revolved around baby sleep, with peak activity during lunchtime and late at night. Despite Aurora's ability to handle in-domain questions, challenges arose with multi-topic inquiries. Paid subscription users demonstrate significantly higher levels of interaction and prolonged usage of the chatbot.

The primary objective of this study was to gauge the overall performance of the Aurora chatbot, particularly in comparison to chatbots that do not use human agents. By integrating human support, Aurora provides an unique hybrid model that offers personalised advice alongside automated responses. This study aims to contribute to the understanding of how such a model can enhance user engagement and satisfaction, and how it compares to purely automated systems. Existing studies predominantly focused on how parents engage with parenting forums, social media platforms, blogs, and applications (Lupton et al., 2016). To our knowledge, this is the first conversation log analysis of a parenting chatbot. Our findings contribute to understanding how parents use parenting chatbots and offer valuable insights to inform the design and improvement of future chatbots for child health, potentially benefiting other applications in social networks and mobile applications for parents.

2. Materials and methods

To understand the engagement of parents with the Aurora child health chatbot, we conducted a mixed-method analysis of Aurora's chatlogs, drawing on user engagement metrics, topic analysis, user satisfaction, and content analysis to make sense of interactions. Qualitative and quantitative data was analyzed in parallel. We could, for example, use message quotes to understand the metrics, as well as investigate how common certain experiences were by analyzing their prevalence in the dataset. The mixed-method approach enabled us to define the nature of the conversations, assess their quality, and identify common and unaddressed topics.

This section documents the dataset used, analysis process, tools, and data privacy and ethics, but before starting we provide additional background on the Aurora chatbot.

2.1. Aurora chatbot

Aurora was a rule-based chatbot that was designed for parents to learn more about child sleep and feeding. The chatbot was created by a Portuguese start-up called AuroraTechAI and was launched in 2018 (Aurora Tech AI, 2023). The company was led by a healthcare professional, who iteratively evolved the content to deal with the doubts of parents that she observed in her counseling practice. Aurora launched with a specialized focus on infant sleep issues and breastfeeding topics. An additional feature was an educational game named "Funzone" (see Funzone: Figure 1), devised as a multiple-choice quiz. The quiz game aimed to debunk common misconceptions about childcare and impart knowledge on a range of topics, including sleep, pregnancy, breastfeeding, and general feeding practices.

Aurora implemented a dual-modal interaction mode, where users could use a menu-based (Figure 1) approach and explore different parts of the content, or explicitly write their question into the chat, which was matched to one of the topics the chatbot could reply about.

The paid subscription model enabled users to access an additional feature – direct communication with its founder, Cláudia. Subscribers could initiate personalised consultations with Cláudia through a typed command ("Cláudia") or by selecting the option from the menu ("Talk with Cláudia": Figure 1). Users would receive daily tips aimed at improving infant sleep patterns (this was only accessible through the paid subscription model). Utilising the chatbot's capability to record the timestamp of the last user interaction, Aurora employed an automated system to send the sleep-improvement tips daily for the subsequent 7 or 30 days (depending on the subscription model).

In technical terms, Aurora operated as a task-oriented, rule-based chatbot that utilised predefined rules and patterns to discern keywords relevant to parenting queries. Aurora did not use a LLM; instead, it relied on a structured rule-based algorithm to provide responses. Aurora's technical architecture comprised Facebook Messenger as the user interface, Chatfuel (Chatfuel, 2023) for predefined responses and menu options using keyword matching, and Dialogflow (Google Cloud, 2023) for handling complex queries through intent matching. Chatfuel (Chatfuel, 2023) managed predefined responses and executed actions, while Dialogflow (Google Cloud, 2023) processed user-typed messages and formulated appropriate responses. Aurora was discontinued in 2020 due to changes in Facebook Messenger's API, which disabled communication between architecture components and hindered its operation.

2.2. Dataset

We analyzed a conversation log dataset from the second version of Aurora, which was operational from October 2018 to September 2021, and consisted of 1043 pseudoanonymized

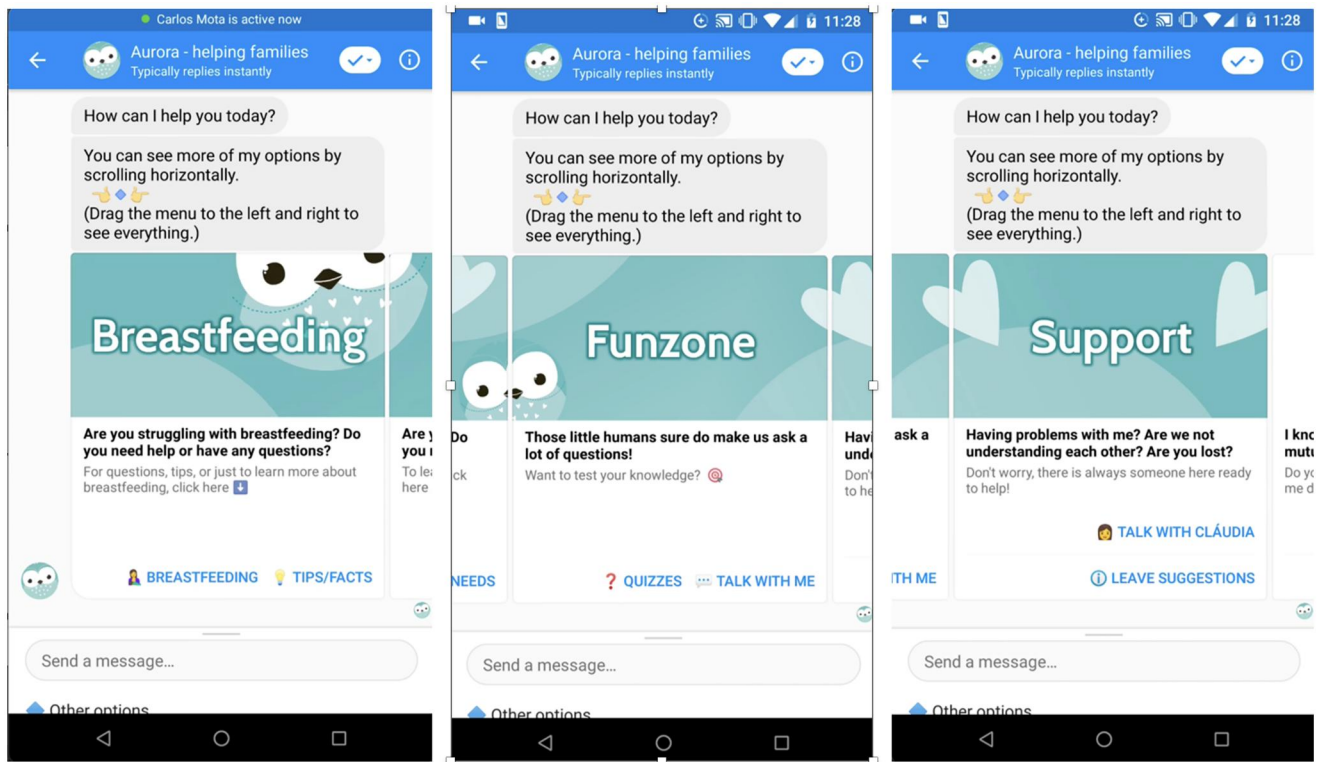


Figure 1. Screenshots of Aurora's features. Showing some of the menu options that users could select: breastfeeding, funzone and support.

unique users. We downloaded user conversations as a Facebook Messenger JSON file and conducted text pre-processing, including lowercasing the text, removing special characters, converting numerical values to textual representations, eliminating stop words, and performing lemmatisation (Hickman et al., 2022). The researcher who read the messages and conducted the analysis is a native speaker in both English and Portuguese language.

2.3. User engagement and conversation metrics

We computed common chatbot log analysis metrics to understand user interactions with the chatbot, including total user count, bounce rate, interaction modality, retention, chat volume, session count, tenure, unique days, dialogue paths, interaction timing, error frequency, and response accuracy (Table 1) (Abd-Alrazaq et al., 2020; Booth et al., 2023; Chen, 2017; Jiang et al., 2024; Lishchynska, 2024; Ma et al., 2021; Maroengsit et al., 2019; Procter et al., 2018). This exploration aimed to uncover trends, challenges, and assess the chatbot's effectiveness in responding to diverse user queries.

While our selection of metrics encompasses a diverse range of dimensions relevant to user engagement and conversation analysis, it is important to acknowledge that alternative evaluation matrices exist (Braun & Clarke, 2006; Casas et al., 2021; El-Ansari & Beni-Hssane, 2023; O'Brien & Toms, 2008). The chosen metrics were prioritised based on their alignment with the research objectives, relevance to the context of the chatbot, and data available (Booth et al., 2023; Chen, 2017; Jiang et al., 2024; Lishchynska, 2024; Ma et al., 2021; Maroengsit et al., 2019; Procter et al., 2018).

2.4. Text-mining

We used text-mining techniques to uncover recurrent topics in user interactions with the chatbot (Jiang et al., 2024). This step involved data exploration, topic identification, word-topic mapping, topic assignment, and topic analysis (Jiang et al., 2024). The first author individually read and classified messages that encompassed multiple topics. This process also helped identify instances where the chatbot failed to respond satisfactorily to users' queries.

2.5. User satisfaction

We used descriptive statistics to analyze star ratings and thus assess user satisfaction with Aurora. Every few messages, the chatbot prompted users to evaluate their interaction using a star rating system that ranged from 1 to 5 stars (highest score). This rating system allowed users to provide feedback and assess their overall satisfaction with the chatbot.

Additionally, we performed a qualitative analysis of users' feedback from star ratings. This allowed us to further understand users' perceptions of the chatbot's performance and information dissemination.

2.6. Content analysis

The qualitative analysis of the content was based on conventional content analysis (Hsieh & Shannon, 2005), whereby the first author read chatbot messages classifying them into categories. This method is commonly used in research to assess various aspects of chatbots, such as their core

Table 1. Computed chatlog analysis metrics used in this study.

Metric	Description
Bounce rate (Lishchynska, 2024)	Number of users who sent very few messages to Aurora (less than 5), divided by the number of total users.
Interaction modality of menu prompts (Ma et al., 2021; Maroengsit et al., 2019)	Number of times the user selected menu prompts to interact with Aurora divided by the total number of messages sent.
Interaction modality of written queries (Ma et al., 2021; Maroengsit et al., 2019)	Number of times the user wrote a query to interact with Aurora divided by the total number of messages sent ^a .
Chat volume (Chen, 2017)	Total number of messages users sent to the chatbot during their tenure.
Session count (Jiang et al., 2024)	Number of periods in which a user exchanged consecutive messages with the chatbot. Messages arriving 30 min later were considered to belong to different sessions.
Tenure (Booth et al., 2023)	Number of days that lapsed from a user's first interaction with the chatbot to their last interaction.
Unique days (Jiang et al., 2024)	Number of days a user has sent at least one message to the chatbot.
User initiated conversation (Jiang et al., 2024)	Conversations initiated with users prompts.
Chatbot initiated conversation (Jiang et al., 2024)	Conversations initiated by the chatbot, for example in the sleep-related program that sent the user a daily message with a tip.
Interaction timing (Jiang et al., 2024)	Specific times and days of the week when users sent messages to the chatbot.
Error frequency (Jiang et al., 2024)	Number of error response messages generated by the chatbot.
Response accuracy (Procter et al., 2018)	Frequency with which the chatbot accurately met user queries and informational requirements of user-typed questions.

^aBased on a manual review of the conversations, it was established that messages with fewer than three words as menu selections and those with more words as typed responses because the JSON files did not explicitly differentiate between the two.

functionalities, the range of responses they provide, and their reliability in disseminating information (Chang et al., 2022; Hsieh & Shannon, 2005; Zhang et al., 2024). Early in the process, we read complete conversations to understand the types of messages and structure of the chatbot. After the familiarization phase, we selected messages to read purposefully, focusing on: (a) first messages sent to the chatbot, (b) last messages sent to the chatbot, (c) messages exchanged before and after error messages, (d) messages relating to common topics identified by text-mining, or (e) messages associated to a specific star rating. We also classified messages as Error, Incomplete, Accurate, and Referreal, depending on based on how well the chatbot responded to an users question.

2.7. Analysis tools

The quantitative analysis was supported by Python 3 running on Jupyter Notebooks (Randles et al., 2017). We used Pandas (McKinney, 2010) (version 2.0.2) for data analysis and Matplotlib (Hunter, 2007) (version 3.7.1) to create data visualisations. Facebook conversations were processed from JSON files into CSV (Comma Separated Values) format to enable tabular data analysis methods.

The qualitative analysis was performed on Microsoft Excel, where we categorized messages into categories using columns of the sheet. The selection of messages to read, was purposeful, as mentioned above, and was achieved by entering scripts into Jupyter Notebook.

2.8. Data privacy and ethics

The analyzed dataset was provided by Aurora Tech AI (Aurora Tech AI, 2023), the company behind Aurora. Throughout the data analysis, ethical considerations and best practices were taken into account to ensure that data handling complied with all relevant privacy standards. The

chatlogs were pseudonymized to protect user privacy before beginning the analysis with a script that replaced the username with a participant number. While reading the logs, we manually removed any usernames that the script had not replaced. The study was compatible with the Terms and Conditions of the Aurora chatbot and was approved by the ethics committee of the Human Science Research Council (no. 1/26/04/22) and the University of Cape Town (no. HREC.REF 339/2022).

3. Results

3.1. User engagement

A total of 1043 unique users exchanged messages with the chatbot (Table 2). Active users, those who sent more than five messages, were 718 users (69%). The remaining 31%, referred to as “bounced users,” were excluded from the analysis due to their erratic behaviour, characterised by messages that seemed to be focused on testing the chatbot or playing with it. Active users engaged in 24,350 interactions with the chatbot.

3.1.1. User engagement analysis

The analysis showed that users, on average, interacted with the chatbot 34 times, with a standard deviation of 31. The wide spread of values around the mean suggests that user interactions varied significantly between users. On average, each user interacted with the chatbot in 6 sessions, with moderate variability, as indicated by the standard deviation of 6.

Average user tenure was 82 days, with 127 days as the standard deviation, indicating a considerable variation in user engagement. As per Figure 2, 61.42% of users used the chatbot for 82 days or less. This finding suggests that user engagement was sporadic, with the most active user engaging with the chatbot for 1157 days. The average number of unique days users utilised the chatbot was only 5 days, with

a standard deviation of 4 days. While user tenure showed significant variability, the number of unique days of use was more consistent among users.

3.1.2. Modality used in interacting with the chatbot

User interactions with the chatbot occurred through user-written questions or menu selections, with the latter being the preferred input method, accounting for 87% of interactions (Figure 3). The preference for the menu selection method likely stems from its convenience in accessing or exploring the content of the chatbot, or from the lack of adhesion to the paid subscription service.

3.1.3. Error frequency

When users interacted with the chatbot, the chatbot provided an answer on topic 82% of the time, demonstrating a high response rate. In the remaining instances (18%), the chatbot responded with an error message. Reading what users wrote before the chatbot produced error messages, we understand that there were various factors contributing to errors, including technical issues, backend processing errors, user interruptions, or vaguely worded questions.

3.2. Usage times

The highest user engagement levels were found between 10:00 and 15:00 and from 21:00 to 00:00, with the lowest

number of messages exchanged between 02:00 and 7:00 (Figure 4). Thursdays stood out as the day with the most frequent messages sent to the chatbot, accounting for 18% of all user messages, with 430 (10%) messages sent to the chatbot at 12:00. The overall volume of messages sent to the chatbot declined as the weekend approached, with Sunday registering the lowest number of exchanged messages.

The timing at which parents sent the messages was often aligned with critical parenting moments, such as putting a child to sleep (Figure 4). For example, User P380 wrote to the chatbot: “I cannot get my baby to sleep before 23h30–00h30 ...” at 00:00, which makes us believe they were struggling to put their child to sleep at the moment when they sent the message. Other messages aligned with feeding or napping times as well.

3.3. Topics of questions to the chatbot

Parents queried Aurora about various topics, including baby sleep problems, feeding, and healthcare (Table 3). Sleep problems accounted for 80% of the questions, followed by 11% of questions related to baby feeding. Other healthcare-related conversations, encompassing the utilisation of medical products, constituted the remaining 9% of the conversations. User conversations sometimes included mixed topics, such as questions about sleep and babies’ dependency on falling asleep while feeding at the breast.

3.3.1. Sleep conversations

Most questions from parents focused on baby sleep patterns and parental sleep deprivation. Regarding baby sleep patterns, the chatlogs included messages about baby night-time sleep disruptions (89% of all sleep-related messages), and sleep routine training (7%). The remaining 4% of the messages focused on the emotional impact of sleep deprivation on parents, especially mothers. For example, P75 questioned about her child sleep schedules:

P75: “My little one is almost 7 months old and used to sleep super well ... now she wakes up hourly at night. I don’t know if it is a growth spurt, or do I need to have more patience or change something?” – user message sent at 23:06 on April 2020.

Table 2. Engagement metrics of the users in the analysed dataset.

Users	
Unique user count	1043
Active users	718 (69%)
Bounced users (excluded from analysis)	325 (31%)
Usage interaction	
Average number of interactions	34 (± 31)
Average number of sessions	6 (± 6)
Average tenure	82 days (± 127)
Unique days (user messages only)	5 (± 4)
Total chat volume	24,350
Interaction modality with chatbot	
Interaction modality of menu prompts	87%
Interaction modality of written queries	13%
Error frequency	
Chatbot outputs related response	82%
Chatbot outputs an error response	18%

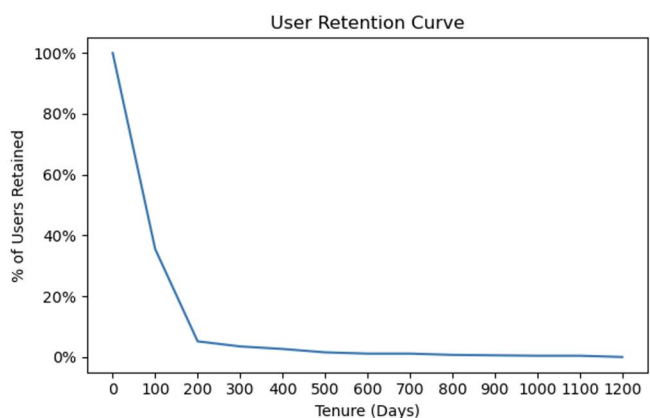
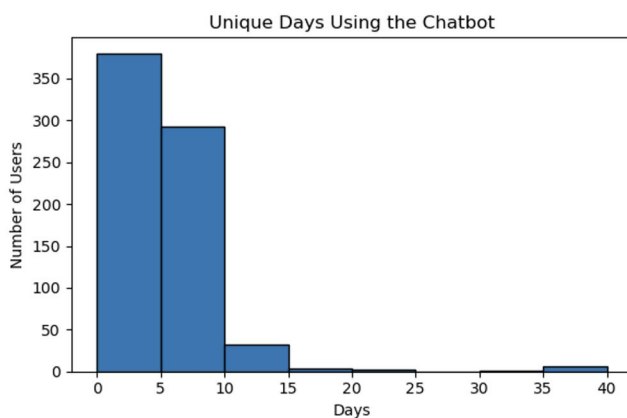


Figure 2. Unique days and user retention of Aurora, consisting of active users only.

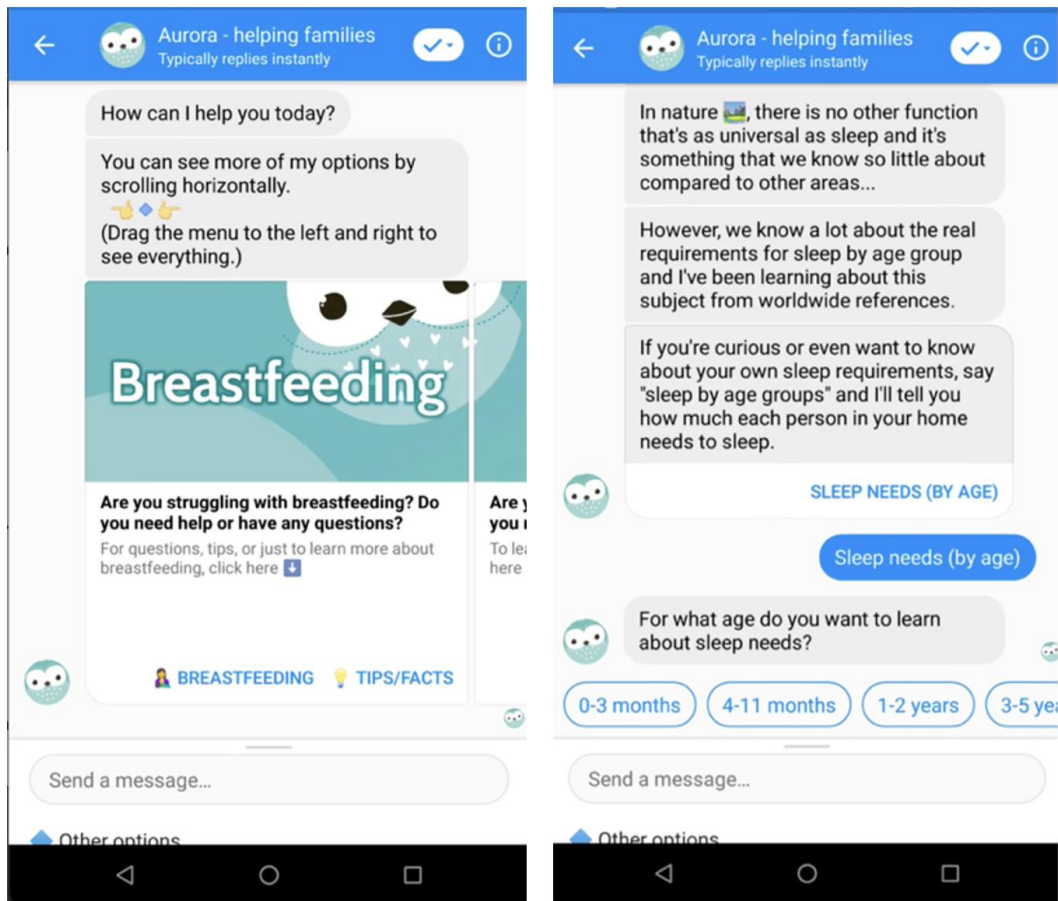


Figure 3. Screenshots of Aurora's user interface, showing content options to explore (left) and menu options for obtaining personalized content (right).

	- 12am	- 1am	- 2am	- 3am	- 4am	- 5am	- 6am	- 7am	- 8am	- 9am	- 10am	- 11am	- 12pm	- 1pm	- 2pm	- 3pm	- 4pm	- 5pm	- 6pm	- 7pm	- 8pm	- 9pm	- 10pm	- 11pm
Monday -	112	114	19	46	16	32	9	57	193	115	279	253	124	221	281	226	129	56	131	99	143	208	241	124
Tuesday -	182	40	35	13	27	36	4	39	82	108	151	168	314	334	255	190	219	151	103	167	135	197	314	205
Wednesday -	108	48	38	49	23	6	51	13	111	103	213	363	261	217	378	258	161	115	203	135	137	310	453	215
Thursday -	154	46	8	28	28	22	51	80	156	144	227	300	430	208	251	327	292	220	189	191	147	387	387	164
Friday -	140	151	12	18	6	22	61	42	118	95	237	188	207	188	233	187	184	133	157	130	97	165	258	136
Saturday -	158	104	26	12	36	28	15	13	71	68	451	171	151	137	250	202	186	198	178	129	191	180	372	320
Sunday -	137	66	60	101	62	16	16	44	153	85	135	178	147	128	232	181	153	110	103	172	102	196	380	178

Figure 4. Time and day that active users sent messages to Aurora. Total count of user messages to the chatbot.

P75 wrote to the chatbot about a recent worsening of her baby's sleep asking if the novel sleep issues could (a) be caused by a growth spur, or (b) be normal for her age, meaning she would have to adapt to these issues. Queries such as this one were common in the dataset. Parents' messages often included very specific details, examples, or stories from their children, to provide context to their questions. The questions were in most cases focused on knowing the reasons behind issues the babies faced, as well as practical strategies they could adopt to address them.

3.3.2. Baby feeding conversations

There were a variety of questions related with feeding and breastfeeding. Commonly raised questions included queries

about breastfeeding tips (45%), latching difficulties or nipple pain (17%), and breastfeeding positions (7%). P245 sent a message about pain while breastfeeding:

P584: "I had cracks in my nipples in the first few days, it's already better, but they're still not good. It continues to hurt me in the first few seconds of breastfeeding. What can I do to make it stop hurting?" – user message sent at 12:25 on December 2018.

P584 struggled with pain during breastfeeding due to nipple cracks. Having pain during the first seconds of breastfeeding is a common experience for some mothers, but continuing to have pain after latching can be a sign of more serious issues. Having this in mind, the healthcare professional, who monitored messages exchanged by the chatbot, decided to write a personalised response to the mother, offering recommendations on how to alleviate breast pain effectively.

Table 3. Examples of messages sent to the chatbot about sleep, breastfeeding, and healthcare-related topics or issues.

Topic	User	Date	Message
Sleep	P412	21:55 November 2018	"I am still struggling she will not go to sleep."
	P244	00:24 November 2018	"I have a routine for her brother. The problem is she constantly wakes up when her brother makes a noise or leaves the room."
	P272	21:47 December 2018	"I have twins, so the waking's are doubled. They are 28 months old."
	P476	21:47 March 2019	"How do I make her sleep the whole night without her asking for milk?"
	P539	08:08 May 2019	"How many hours should a 16-month-old baby sleep? And naps? How to prevent him from waking up crying/moaning?"
	P579	15:26 January 2020	"My son is 5 months old and does not sleep enough during the day nor at night."
Feeding	P253	23:13 June 2019	"The baby wakes up a lot and always cries a lot, he only calms with the breast, my husband is impatient, and so is my other 6-year-old son, and I've been sleep deprived since he was born."
	P375	23:30 January 2019	"How often should I breastfeed?"
	P248	02:07 March 2019	"My baby breastfeeds for a short amount of time."
	P395	13:13 December 2018	"When I lift her up to change, she screams a lot and only breastfeeding calms her down. Now I no longer breastfeed just once a night like I did since she was 2 months old and I'm afraid I'm 'ruining' my night-time routine when it comes to feeding, but the alternative isn't acceptable to me ... In addition to taking a long time to calm down, I don't want to run the risk of her being hungry and me not giving her."
	P245	11:26 November 2018	"Baby is not latching."
	P842	22:13 December 2018	"What can I take to increase my breast milk production?"
Other healthcare topics	P359	08:38 January 2020	"Can I use the breast pump just to stimulate the breast before the baby feeds?"
	P925	15:43 December 2018	"How do I know if he has breastfed enough?"
	P345	11:48 November 2018	"What are reflux symptoms?"
	P793	20:39 March 2019	"I wish the baby slept more hours at a time and didn't have colic."
	P724	14:22 December 2019	"Can I take Alerius (antihistamine) while breastfeeding?"
	P637	22:35 December 2018	"I get tired after waking up so many times to feed my child, and then I find it difficult to fall asleep."
	P152	16:18 December 2018	"I am not sleeping and struggling with post-natal depression."

Other feeding user questions included feeding times and determining whether the baby was full (9%).

P925: "How do I know if he has breastfed enough?" – user message sent at 15:43 on December 2018.

Parents, such as P925, were worried that their babies not receiving sufficient milk or not feeding for adequate durations, so they asked how they could be sure if their babies were eating as well. Many mothers also expressed concerns about having low milk production and sought advice on its causes and ways to increase their breast milk supply (22%).

3.3.3. Other healthcare-related messages

The chatbot also received questions from users regarding illnesses (87%), such as respiratory problems, allergies, and fever, as well as queries about colic and reflux (12%), and medications for addressing these issues (1%).

Another common topic was asking whether specific medications were safe for babies:

P724: "Can I take Alerius [commercial name of desloratadine antihistamine] while breastfeeding?" – user message sent at 14:22 on December 2019.

Multiple users inquired the chatbot about the safety of taking medication while breastfeeding. Allergy-related medication and immune boosters for both the baby and mother being frequent topics of interest. In some instances, such as for P724, the chatbot responded by indicating that it could not understand the user's question.

3.4. Chatbot response

We analyzed Aurora's responses to free-text messages to assess its performance in handling user questions (Table 4).

Notably, we found that 46% of free-text questions elicited an error message in response. Error-prone responses often stemmed from three primary factors: the lengthiness of user messages, the vagueness of questions, and questions out of the chatbot's content corpus.

In instances where users submitted excessively long messages, comprising several paragraphs, the chatbot's parsing algorithms struggled to handle the input, resulting in an error response. Similarly, the occurrence of error messages was notable in response to vague questions, exemplified by questions that lacked specificity or clarity and contained ambiguous language. Such questions posed a challenge to the chatbot's natural language understanding capabilities.

P15: "Is it possible to remove the supplement?" – user message sent at 11:56 on May 2020.

Aurora: "You know ... I usually help parents with breastfeeding and sleeping difficulties ... Can I help you with these issues?"

P15 questioned Aurora whether it would be safe to remove "the supplement," an amount of baby formula the parent was probably giving the baby after breastfeeding to compensate for low milk supply. However, without much context it would be hard for a chatbot (or even a person) to answer appropriately to such query. In this particular instance, Aurora did not consider the question to be within its corpus of content and invited the parent to ask about the topics Aurora had most information about.

When subscription users asked questions the chatbot did not understand, it would redirect (*referral responses*) users to engage with the healthcare professional instead of outputting an error message. This referral pathway facilitated direct interaction with the healthcare professional, accounting for 8% of free-text messages responses. These interactions

Table 4. Categorization of chatbot responses.

Response category	Description	Example	% of Total responses
Error	When the chatbot failed to understand the user question, it responded with an error message.	User Question: How to sing to my baby? Chatbot Response: Do you know that every day I'm learning new things? But I don't think I can understand you yet ... :(	46
Incomplete	Responses where the chatbot provided some relevant information but did not fully address the user's question.	User Question: My baby can't sleep for 3 hours and it's most likely because of colic but I'm not sure and I've been to the doctor about it but I haven't had much help and I've been waking up constantly. Chatbot Response: Throughout this program I will explore with you various points related to sleep and the quality of a night's rest. In these first days, I will start with something more related to your main difficulty but this does not replace the entire program. Note: in sleep, sometimes what causes us the most difficulty is just the most visible "symptom." It is always important to understand sleep as a whole. But I'm here to help you with that!	35
Accurate	Responses where the chatbot provided relevant and accurate information that addresses the user's question effectively.	User Question: How do I get my baby to sleep enough during the day? Chatbot Response: (...) Paying attention to the first signs of sleep and helping the baby to fall asleep, for example by moving your finger from the forehead to the nose, thus encouraging the baby to close their eyes, is usually a good tip.	11
Referral	When the chatbot was not able to provide a response it directed users, particularly paid subscribers, to consult healthcare professionals.	User Question: Now she's still sleeping, should I breastfeed her now, in your opinion, should I not wake her up? Healthcare Professional: You never wake up a sleeping baby;)	8

yielded personalised responses tailored to the users' specific healthcare needs.

Approximately 35% of Aurora's responses were categorised as *incomplete*, owing to user questions encompassing multiple topics, which resulted in the chatbot responding to only one of the topics, thus ignoring the nuance or complexity of a multi-topic question. For instance, if an user was to inquire about both sleep routines and feeding schedules in the same message, Aurora defaulted to addressing only the most prominent or easily identifiable topic. This selective response pattern resulted in an incomplete or partial answer that failed to address the entirety of the user's concerns adequately.

User questions solely concerning sleep or breastfeeding, accounting for 11% of (*accurate*) responses, provided what can be considered appropriate or complete responses by the chatbot. These responses typically entailed the dissemination of general information pertaining to the queried topic. For instance, when addressing questions about sleep, the chatbot would offer generic suggestions, derived from its pre-programmed knowledge base, such as establishing a consistent bedtime routine or ensuring a conducive sleep environment. In these instances, the information provided by the chatbot was relevant to the topic at hand, however, it lacked personalisation tailored to the user's specific circumstances or preferences.

A peculiar behaviour became apparent in instances where the chatbot deferred information delivery and prompted a star rating and feedback. An example illustrated in [Figure 5](#), is when user P539 inquired about the ideal duration for their 16-month-old baby's sleep or nap and how to prevent their baby from waking up crying. Aurora provided a generic response and concluded the conversation by requesting an evaluation of the information received. In the case of P539, they rated the interaction with 3 stars, possibly

reflecting dissatisfaction with the chatbot's delayed information delivery.

3.5. Types of users

On analyzing the conversation logs, we categorized four distinct user groups ([Table 5](#)) based on their subscription preferences:

- **Free Content Users:** Users (678 users) who only accessed the free content without opting for any subscription or engaging with the healthcare professional during the 4 day free trail.
- **Free Health Engagement:** Users (27 users) who accessed the free content and chose to engage with the healthcare professional during the 4 day free trial period.
- **28-day Subscription:** Users (9 users) who subscribed for 28 days, benefiting from a comprehensive program offering tips on feeding, sleep, and direct communication with the healthcare professional.
- **7-day Subscription:** Users (4 users) who opted for a 7-day subscription plan, which granted them access to additional sleep information and the ability to interact with the healthcare professional for personalised advice.

Users who engaged with the healthcare professional during the free content period sent an average of 53 messages per user, which was higher than the rest of the free content users who sent an average of 30 messages per user. The free content users interacted approximately 50% less and engaged approximately 86% less compared to the 7-day plan users, and approximately 77% less and 88% less compared to the 28-day plan users.

The 28-day plan users had the highest average session count with 62 sessions per user, followed by the 7-days plan

users with 27 sessions per user. Both subscription groups had higher session counts (approximately over 244% more) compared to users who engaged with the healthcare professional (15 sessions per user) during the free trail and the rest of the free content users (18 sessions per user). However, users who engaged with the healthcare professional for free and the rest of the free content users exhibited relatively consistent session frequencies, with lower standard deviations compared to the subscription groups.

Users who subscribed to the 28-day plan had the longest average tenure, with an average of 300 days, followed by users subscribed to the 7-day plan, with an average of 204 days. In contrast, users who engaged with the healthcare

professional during the free trial period and the rest of the free content users had shorter average tenures of 71 days and 78 days, respectively. Interestingly, users who engaged with the healthcare professional for free exhibited a shorter range of tenure, as indicated by the lower standard deviation, suggesting variability in the duration of engagement among this group.

The 7-days plan users sent an average of 103 messages per user, while 28-days plan users sent an average of 207 messages per user, reflecting a substantial increase of approximately 100% in message volume between the two subscription groups. This discrepancy in message volume suggests that the duration of the subscription plan may influence the intensity of user interaction, with longer subscription periods allowing for more prolonged engagement and interaction with the chatbot.

Additionally, an examination of the conversations among users who engaged with the healthcare professional revealed that their main concerns revolved around specific baby sleep problems, with fewer questions about breastfeeding, and general baby health. Users were able to explain their concerns in greater detail, enabling the healthcare professional to gather additional information about the user's problem. Unlike the chatbot's automated responses, which were constrained by pre-programmed responses, the healthcare professional possessed the flexibility to probe further and ask clarifying questions to better grasp the nuances of the user's problem, and in turn provide a personalised response. Some users also expressed curiosity about the identity of the healthcare professional, indicating a desire for reassurance and credibility in the advice provided. The message exchange between users and the healthcare professional often assumed a conversational tone, allowing for a more fluid and interactive dialogue.

However, it is important to acknowledge that users sometimes encountered delays in receiving responses, as the healthcare professional was not always immediately available to attend to queries. Despite this limitation, the hybrid-interaction enabled the healthcare professional the opportunity to follow up with users to assess the effectiveness of recommendations provided and provide further assistance to the users.

3.6. Star rating analysis

As mentioned before, Aurora prompted users to rate their interactions with the chatbot every few sessions. The star ratings of Aurora were positive. Most ratings (52%) assigned Aurora a 5-star rating, and 34% gave 4 stars (Figure 6). Fewer ratings were assigned 1, 2, or 3 stars (1, 2, and 11%, respectively).

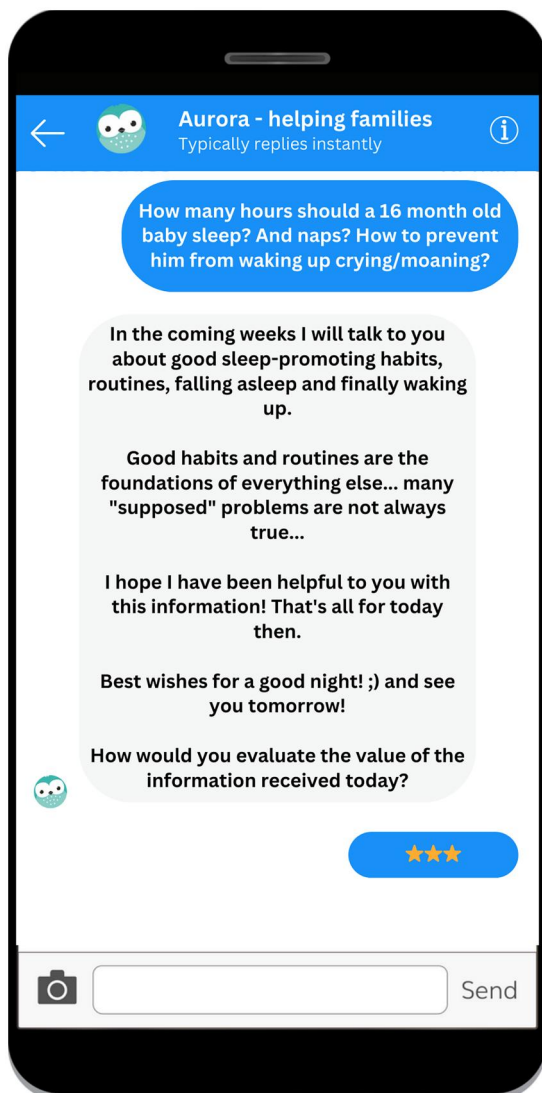


Figure 5. Conversation between Aurora and P539 at 08:08 on April 2019. Demonstrating how Aurora responded to a user query about how often a 16-month-old baby should sleep and nap.

Table 5. User groups that engaged with the Aurora chatbot and the healthcare professional.

	7-Day subscription	28-Day subscription	Free health engagement	Free content users
Users (%)	4 (0.6%)	9 (1.25%)	27 (3.8%)	678 (94.35%)
Tenure	204 (± 81)	301 (± 349)	71 (± 91)	78 (± 120)
Interactions	103 (± 20)	207 (± 75)	53 (± 35)	30 (± 20)
Sessions	27 (± 6)	62 (± 14)	15 (± 3)	18 (± 2)

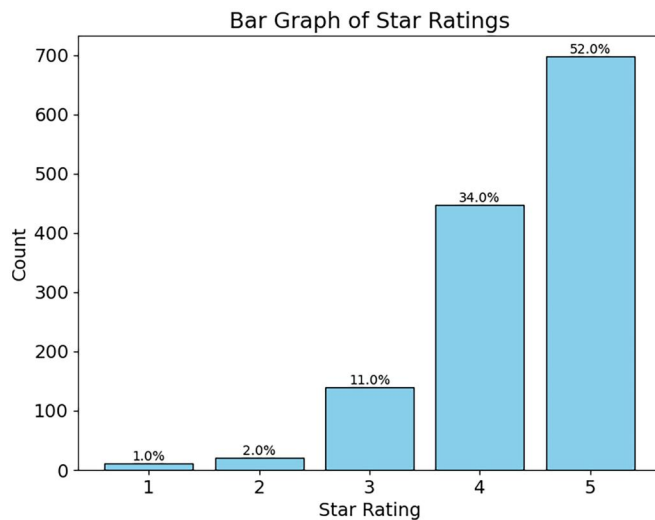


Figure 6. Bar chart illustrating the distribution of all star ratings provided by users to Aurora.

The analysis indicated a variance in the frequency of user ratings. 20% of users rated Aurora only once. 12% of users issued two ratings, suggesting more interaction with the chatbot. A further 25% of users contributed three ratings, while 33% gave four-star ratings. Remarkably, 10% of users demonstrated a strong interest in the chatbot assessment by rating their experience five times or more.

We found that as users engaged with the chatbot more frequently, their average ratings increased, indicating higher satisfaction. However, these observed variations among other user groups are relatively minor and are closely related to overall star rating. The cumulative average star rating of 4.37 stars signifies a generally favourable perception of the chatbot among this user base.

3.7. User feedback analysis

Upon examining the user-written feedback accompanying star ratings, we observed that only 7% of users (74 entries) provided comments. These comments were grouped and categorised based on the nature of the feedback. The qualitative analysis shed light on the types of feedback users provided regarding their interaction with the chatbot and the type of information provided.

There was several positive feedback (21 instances), where users expressed appreciation for the chatbot and stated that the program was helpful and offered valuable tips. Secondly, there were expressions of dissatisfaction (24 instances) with specific aspects, such as paying to use it, finding the information to be common sense and delayed information dissemination. The dissatisfaction regarding delayed information dissemination was due to Aurora's practice of only sending short tips or tips on limited topics. When users sought further clarification on how to implement the tips, the chatbot responded by indicating that additional information would be provided in the following days. This chatbot response left users feeling lost or unable to implement the suggestions provided. Lastly, some users expressed confusion (29 instances), indicating that they did not fully understand

the program's purpose or felt that they needed additional information about the credibility of the chatbot responses.

4. Discussion

4.1. Main findings

Users engaged most actively with Aurora between 10:00–15:00 and 21:00–00:00, with a decline in messages between 02:00–07:00. On average, users interacted 34 times with the chatbot over 6 sessions with a tenure of 82 days. Parents predominantly sought guidance on sleep issues (80%), followed by breastfeeding and baby feeding (11%) and health-care topics (9%). Feedback on the chatbot was mostly positive, with 52% of users giving a 5-star rating. It was evident that users who subscribed to the paid service had more interactions and used the chatbot more extensively.

This study revealed a consistent emphasis on sleep-related issues among parents seeking support, a trend found in parenting chatbots such as Mamabot (Vaira et al., 2018) and Feedpal (Yadav et al., 2019). The analysis revealed that parental concerns frequently coincided with other issues, such as sleep disturbances caused by a health problem or breastfeeding. This finding highlighted the limitations of a rule-based chatbot design and underscored the challenge of addressing multifaceted parenting questions, as corroborated by Yadav et al. (2019). Aurora's design revealed that a rule-based and domain-specific architecture might restrict its ability to engage in fluid dialogues, aligning with broader research indicating that such constraints often lead to user dissatisfaction (Brandtzaeg & Følstad, 2017; Janssen et al., 2021; Luger & Sellen, 2016). Integrating human support with automated responses was shown to enhance user experience and trust, suggesting that future chatbot designs should consider hybrid interaction models to address diverse user needs comprehensively.

4.2. Engagement patterns with the Aurora chatbot

While social media, parenting blogs, and baby care applications provide valuable childcare information, they may lack the real-time interactivity and personalised responses characteristic of chatbots (BabyCenter, 2023; BabySparks, 2021). This limitation highlights the unique advantage of chatbots in offering parents personalised information tailored to their specific needs, potentially making a significant positive social impact (Yadav et al., 2019).

Comparing our findings with existing literature revealed important insights into parenting preferences. While previous studies have noted the evolving needs of parents as their children develop (Vaira et al., 2018; Yadav et al., 2019), our analysis indicates a clear preference among users for personalised responses tailored to their specific circumstances. This aligns with research by Coleman et al. (2023), which emphasized the importance of contextually relevant guidance in addressing the unique challenges faced by parents.

In contrast to chatbots who are solely reliant on pre-existing content (Booth et al., 2023; Vaira et al., 2018; Yadav

et al., 2019), Aurora's hybrid model, which integrated automated responses with human intervention, appeared to be effective in enhancing user engagement. This finding is consistent with the observation that users often disengage from chatbots due to lack of human contact (Potts et al., 2021).

Moreover, our study highlighted that parents tend to utilise the chatbot during critical parenting moments, such as when they are facing challenges with their child's sleep or feeding, emphasizing the importance of delivering timely advice when it is most needed (Brandtzæg et al., 2021). This aligns with existing research suggesting that parents often seek support and information during times of need (Brandtzæg et al., 2021; Cancel & Gerhardt, 2019; Higashinaka et al., 2014). The low usage periods during early morning hours could be attributed to different factors. One possibility is that parents prioritised sleep during these hours, recognizing its importance in managing the demands of parenting. Another explanation could be that parents encounter fewer parenting challenges during this time, leading to reduced engagement with the chatbot. Additionally, parents may have preferred to address parenting issues during waking hours when they had more energy and focus, contributing to the lower usage observed during early morning hours.

4.3. Limitations of the chatbot technology stack

Despite the availability of human intervention, a significant portion of user free-text interactions with the chatbot resulted in error or incomplete responses. This highlighted significant challenges faced by Aurora, particularly in the areas of content expansion and intent recognition. The organic expansion of Aurora's content through collaboration with the chatbot developer raised two significant concerns: dependence on a single source for content growth and limited data availability for Chatfuel (Chatfuel, 2023) and Dialogflow (Google Cloud, 2023), resulting in challenges for keyword matching and intent recognition. This dependency raised concerns about the sustainability and resilience of the chatbot's knowledge base.

Furthermore, ambiguity and uncertainty in intent recognition led to instances of misinterpretation of user queries and suboptimal responses from Aurora. Given that Dialogflow depends on pattern-matching (Zhan et al., 2023), the insufficient data to cover a wide range of possible user intents hindered Dialogflow's generalisation capabilities, limiting its adaptability to new or unseen queries (Keszocze & Harris, 2019).

4.4. Reflection on the ethical implications of analyzing pseudoanonymized chatlogs

The analysis of personal chatlogs, even with a chatbot, can raise issues of privacy and consent. We reflected in-depth if we should analyze the dataset and consulted with two ethics committees on this issue. We considered asking for consent from users, however, it was not really possible. Changes to Facebook's API disabled Aurora, and, consequently,

inhibited the chatbot from messaging users to acquire their explicit consent. Asking users by manual messaging would be possible but would highly restrict the amount of analysis possible (due to non-responses from past users).

After in-depth discussion we considered written informed consent was not required for four reasons. First, conversations with the chatbot agent are structured in specific workflows, encouraging users to click on certain menu options to access certain content. Very quickly users understand that the chatbot has limited skills in understanding messages, and thus it is unlikely that highly sensitive information is shared in conversations. Second, users were probably expecting their conversations to be monitored/analyzed. All conversations with Aurora occurred on Facebook Messenger, a chat window that appears on Facebook website or that is installed as an app together with Facebook. Assuming users attribute the same level of privacy to Facebook messenger as they attribute to Facebook itself, they would not expect their data to escape research analysis. Third, the automatic anonymization protects the privacy of participants, and we collected no additional data from users whose conversations were analyzed. Finally, Aurora users agreed to terms and services sheet that mentions the possibility of their conversations being studied to study and improve future versions of the chatbot. These arguments were welcomed by the ethics committee.

4.5. Limitations and future directions

Considering the limitations in query responses during free-text instances, it can be argued that Aurora's conversations in Portuguese required robust language support to understand and respond to user queries effectively. This finding underscored an ethical imperative that chatbots like Aurora should not enable conversational capabilities without the supporting architecture of robust language models, such as ChatGPT-4 (OpenAI, 2022), especially in a domain as sensitive as child healthcare (Floridi & Cowls, 2019; Sezgin, 2024). This aligns with Sezgin (2024) study of the transformative impact of large language models on healthcare virtual assistants, which emphasises the benefits of incorporating such models in sensitive domains like child healthcare. The same author highlighted GPT-4's heightened accuracy and clinical relevance in addressing complex healthcare inquiries like postpartum depression. However, the trade-offs associated with larger language models like ChatGPT-4 extend beyond language-specific contexts.

Having said this, while LLMs can enhance the chatbot's ability to handle diverse queries (Mannekote et al., 2023), the potential dangers of providing inaccurate or unsafe healthcare information must be carefully considered and mitigated (Palanica et al., 2019; Sezgin, 2024; Zhan et al., 2023). The expensive maintenance requirements, resource-intensive data needs (Vanian, 2023), and potential biases inherent in training datasets (Zhan et al., 2023) are also issues to consider as one considers moving to an LLM chatbot technology stack. The integration of a hybrid model, such as the one used in Aurora, allowed users to communicate with a healthcare professional, represents a strategic approach to addressing these

limitations. By offering personalised human intervention in response to inadequacies in automated chatbot responses, the hybrid model bridges the gap between technology and human expertise, offering few risks.

While technology, including sophisticated language models, is not a panacea and has inherent limitations, the true challenge lies in striking a balance between technological capabilities and ethical considerations to establish an accurate and appropriate support system. Sezgin (2024) highlights the need for a cautious approach, advocating for collaborative platforms to ensure the robust addressing of ethical guidelines, regulatory oversight, governance principles, and privacy measures.

The findings in this study, although focused on the Aurora chatbot, have broader implications that can be translated to other settings (Porcheron et al., 2020; Wang et al., 2021). The user engagement patterns, predominant topics of inquiry, and interaction modalities observed in this study provide valuable insights that can inform the design and implementation of similar systems in different contexts. For instance, the high engagement during specific times of the day suggests that timing and accessibility are crucial factors for parental support tools, whether they are chatbots, social network platforms, or mobile applications for parents.

One important limitation of this study is the lack of demographic or socioeconomic data about the users. This data was not collected as part of the chatlogs, making it difficult to analyze how factors such as income or education level might influence the use of the Aurora chatbot. While we know that some companies offered subscriptions to their employees, suggesting that the service was not exclusive to high-income individuals, the dataset does not include sufficient information to thoroughly assess this. Additionally, the number of paid users in this dataset is minimal, which should mitigate any significant distortion in our analysis. Future studies should aim to collect and consider such demographic data to better understand the broader applicability and impact of chatbot-based healthcare support.

Future research could explore the application of these insights in the development of social networks and apps dedicated to parental support. By adapting the engagement strategies and content personalisation techniques, designers can create more effective tools for various demographics and cultural contexts. Additionally, the hybrid model of integrating automated responses with human support could be explored further in different settings to enhance user satisfaction and trust.

While this study provided valuable insights into the short-term use of the Aurora chatbot, it is essential to consider the potential long-term impacts of using chatbots in child health. This is not possible with the Aurora system, which is no longer operational. Future research on other parenting chatbots could focus on conducting longitudinal studies to understand how users interact with child health chatbots over extended periods. Such studies could examine changes in user behavior, the evolution of questions asked, and any learning or adaptation that occurs over time. Additionally, researchers could investigate the lasting impact

on child health outcomes and parental satisfaction with healthcare support.

5. Conclusion

This study analyzed user conversations of Aurora, a child healthcare chatbot. The study unveiled that the majority of parenting queries occurred during critical moments, notably revolving around baby sleep. The observed usage patterns underscore the need for tailored support and personalised advice, as evidenced by the increased engagement with the hybrid model. While multi-topic user questions posed challenges, the prevailing positive user ratings suggest an overall satisfactory experience with the chatbot. Additionally, the study highlighted the ethical imperative for chatbots in sensitive domains, stressing the importance of robust language models and the balance between technological capabilities and ethical considerations. It is crucial that these technologies ensure reliability and safety in their responses, as misinformation can adversely affect user trust and well-being.

Furthermore, the findings underscore the limitations of the rule-based chatbot architecture, as Aurora often struggled with multi-topic questions and lacked the contextual understanding necessary for nuanced interactions. Such limitations highlight the need for future chatbot designs to adopt a more holistic, context-aware interaction model. Incorporating human support alongside chatbots could provide personalised assistance and address complex questions more effectively, thus improving user engagement and satisfaction. This approach could offer an alternative method for continuous improvement and refinement of chatbot functionalities.

A novel aspect of this study is the integration of human support with automated chatbot responses, creating a hybrid model that bridges the gap between automated assistance and professional healthcare advice. Additionally, the methodology of combining quantitative engagement metrics with qualitative content analysis offers a comprehensive understanding of user interactions and the chatbot's effectiveness.

This research contributes to the field of chatbots for parenting by analyzing previously unassessed chatlogs of Aurora, providing insights into user challenges, engagement patterns, and performance. Additionally, this research contributes to expanding the scholarly corpus on how parents interact with chatbots. Future studies should aim to investigate the long-term effects of chatbot usage on parental decision-making and child health outcomes, thereby furthering our understanding of technology's role in modern parenting.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by project ParentCoach (FCT AGA-KHAN/541742216/2019) funded by Fundação para a Ciência e Tecnologia and Aga Khan Development Network.

ORCID

Diana Liebetrau  <http://orcid.org/0009-0003-6596-2905>
 Melissa Densmore  <http://orcid.org/0000-0003-1733-2653>

References

- Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical metrics used to evaluate health care chatbots: Scoping review. *Journal of Medical Internet Research*, 22(6), e18301. <https://doi.org/10.2196/18301>
- Agarwal, R., & Wadhwa, M. (2020). Review of state-of-the-art design techniques for chatbots. *SN Computer Science*, 1(5), 1–12. <https://doi.org/10.1007/s42979-020-00255-3>
- Aurora Tech AI. (2023). *Aurora for families*. <http://auroratechai.com/pt/familia/> (accessed May 15, 2023).
- BabyCenter. (2023). *BabyCentre free app: My pregnancy & baby today*. <https://www.babycentre.co.uk/mobile-apps> (accessed May 15, 2023).
- BabySparks. (2021). *What is BabySparks*. <https://babysparks.com/about-us/> (accessed May 15, 2023).
- Booth, F., Potts, C., Bond, R., Mulvenna, M., Kostenius, C., Dhanapala, I., Vakaloudis, A., Cahill, B., Kuosmanen, L., & Ennis, E. (2023). A mental health and well-being chatbot: User event log analysis. *JMIR mHealth and uHealth*, 11, e43052. <https://doi.org/10.2196/43052>
- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In: I. Kompatsiaris, et al. (Ed.), *Internet Science. INSCI 2017. Lecture Notes in Computer Science* (pp. 377–392).
- Brandtzaeg, P. B. B., Skjuve, M., Kristoffer Kristoffer Dysthe, K., & Følstad, A. (2021). *When the social becomes non-human: Young people's perception of social support in chatbots* [Paper presentation]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (pp. 1–13), New York, NY, USA. <https://doi.org/10.1145/3411764.3445318>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Cancel, D., & Gerhardt, D. (2019). *Conversational marketing: How the world's fastest growing companies use chatbots to generate leads 24/7/365 (and how you can too)*. John Wiley & Sons, Inc.
- Casas, J., Tricot, M.-O., Khaled, O. A., Mugellini, E., & Cudré-Mauroux, P. (2021). *Trends & methods in chatbot evaluation* [Paper presentation]. Companion Publication of the 2020 International Conference on Multimodal Interaction, New York, NY, USA (Virtual Event, Netherlands) (ICMI '20 Companion) (pp. 280–286). Association for Computing Machinery.
- Chakraborty, C., Pal, S., Bhattacharya, M., Dash, S., & Lee, S. S. (2023, October). Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Frontiers in Artificial Intelligence*, 6, 1237704. <https://doi.org/10.3389/frai.2023.1237704>
- Chang, I.-C., Shih, Y.-S., & Kuo, K.-M. (2022, September). Why would you use medical chatbots? interview and survey. *International Journal of Medical Informatics*, 165, 104827–104811. <https://doi.org/10.1016/j.ijmedinf.2022.104827>
- Chatfuel. (2023). *Facebook Messenger automation | Chatfuel*. <https://chatfuel.com/facebook> (accessed September 11, 2023).
- Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*. Addison-Wesley Professional.
- Chow, J. C. L., & Li, K. (2024, November). Ethical considerations in human-centered AI: Advancing oncology chatbots through large language models. *JMIR Bioinformatics and Biotechnology*, 5(6), e64406. <https://doi.org/10.2196/64406>
- Coleman, T., Till, S., Farao, J., Shandu, L., Khuzwayo, N., Mbombi, M., Muthelo, L., Bopape, M., Van Heerden, A., & Mothiba, T. (2023, October). *Reconsidering priorities for digital maternal and child health: Community-centered perspectives from South Africa* [Paper presentation]. Proceedings of the ACM on Human-Computer Interaction (vol. 7, p. 31), Minneapolis, MN USA. <https://doi.org/10.1145/36100>
- El-Ansari, A., & Beni-Hssane, A. (2023, February). Sentiment analysis for personalized chatbots in e-commerce applications. *Wireless Personal Communications*, 129(3), 1623–1644. <https://doi.org/10.1007/s11277-023-10199-5>
- Entenberg, G. A., Areas, M., Roussos, A. J., Maglio, A. L., Thrall, J., Escoredo, M., & Bunge, E. L. (2021). Using an artificial intelligence based chatbot to provide parent training: Results from a feasibility study. *Social Sciences*, 10(11), 426. <https://doi.org/10.3390/socsci10110426>
- Entenberg, G. A., Mizrahi, S., Walker, H., Aghakhani, S., Mostovoy, K., Carre, N., Marshall, Z., Dosovitsky, G., Benfica, D., Rousseau, A., Lin, G., & Bunge, E. L. (2023, January). AI-based chatbot micro-intervention for parents: Meaningful engagement, learning, and efficacy. *Frontiers in Psychiatry*, 14, 1080770. <https://doi.org/10.3389/fpsyt.2023.1080770>
- Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 14. <https://doi.org/10.1162/99608f92.8cd550d1>
- Frangoudes, F., Hadjaros, M., Schiza, E. C., Matsangidou, M., Tsivitanidou, O., & Neokleous, K. (2021). An overview of the use of chatbots in medical and healthcare education. In Panayiotis Zaphiris & Andri Ioannou (Eds.), *Learning and collaboration technologies: Games and virtual environments for learning* (pp. 170–184). Springer International Publishing.
- Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97(2019), 304–316. <https://doi.org/10.1016/j.chb.2019.01.020>
- Google Cloud. (2023). *Dialogflow Messenger*. <https://cloud.google.com/dialogflow/es/docs/integrations/dialogflow-messenger> (Accessed July 28, 2023).
- Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1), 114–146. <https://doi.org/10.1177/1094428120971683>
- Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., & Matsuo, Y. (2014). *Towards an open-domain conversational system fully based on natural language processing* (pp. 928–939). <https://aclanthology.org/C14-1088>
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Janssen, A., Grütznert, L., & Breitner, M. (2021). Why do Chatbots fail? A critical success factors analysis. In Joe S. Valacich, Anitesh Barua, Ryan T. Wright, Atreyi Kankanhalli, Xitong Li, & Shaila Miranda (Eds.), *Proceedings of the 42nd International Conference on Information Systems, ICIS 2021, Building Sustainability and Resilience with IS: A Call for Action*, Austin, TX, USA, December 12-15 2021 (vol. 6, p. 17). Association for Information Systems, Austin, Texas, USA
- Jiang, T., Guo, Q., Wei, Y., Cheng, Q., & Lu, W. (2024, September). Investigating the relationships between dialog patterns and user satisfaction in customer service chat systems based on chat log analysis. *Journal of Information Science*, 50(6), 1541–1556. <https://doi.org/10.1177/01655515221124066>
- Kesozcze, O., & Harris, I. G. (2019). Chatbot-based assertion generation from natural language specifications. In *2019 Forum for specification and design languages (FDL)* (pp. 1–6). IEEE. <https://doi.org/10.1109/FDL.2019.8876925>
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018, September). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248–1258. <https://doi.org/10.1093/jamia/ocy072>

- Lishchynska, D. (2024). *Chatbot analytics: 14 Chatbot metrics to track in 2024*. <https://botscrew.com/blog/chatbot-metrics/> (Accessed February 27, 2024).
- Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents [Paper presentation]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16) (pp. 5286–5297), San Jose, CA, USA. <https://doi.org/10.1145/2858036.2858288>
- Lupton, D., Pedersen, S., & Thomas, G. M. (2016). Parenting and digital media: From the early web to contemporary digital society. *Sociology Compass*, 10(8), 730–743. <https://doi.org/10.1111/soc4.12398>
- Ma, Y., Kleemann, T., & Ziegler, J. (2021). Mixed-modality interaction in conversational recommender systems. Interfaces and Human Decision Making for Recommender Systems 2021: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (CEUR Workshop Proceedings (vol. 2948, pp. 21–37). RWTH Aachen, Aachen.
- Mannekote, A., Celepkolu, M., Wiggins, J. B., & Boyer, K. E. (2023). Exploring usability issues in instruction-based and schema-based authoring of task-oriented dialogue agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)* (pp. 1–6). Association for Computing Machinery. <https://doi.org/10.1145/3571884.3604310>
- Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A survey on evaluation methods for chatbots [Paper presentation]. Proceedings of the 2019 7th International Conference on Information and Education Technology, New York, NY, USA (ICIET 2019) (pp. 111–119). Association for Computing Machinery, Aizu-Wakamatsu, Japan.
- McKinney, W. (2010). *Data structures for statistical computing in python* [Paper presentation]. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference, SciPy Conferences* (pp. 56–61). Austin, Texas, USA.
- O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6), 938–955. <https://doi.org/10.1002/asi.20801>
- OpenAI. (2022). *Introducing ChatGPT*. <https://openai.com/blog/chatgpt> (accessed August 13, 2023).
- Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of Medical Internet Research*, 21(4), e12887. <https://doi.org/10.2196/12887>
- Porcheron, M., Clark, L., Jones, M., Candello, H., Cowan, B. R., Murad, C., Sin, J., Aylett, M. P., Lee, M., Munteanu, C., Fischer, J. E., Doyle, P. R., & Kaye, J. (2020). *CUI@CSCW: Collaborating through conversational user interfaces* [Paper presentation]. Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing, New York, NY, USA. (Virtual Event, USA) (CSCW '20 Companion) (pp. 483–492). Association for Computing Machinery.
- Potts, C., Bond, R., Mulvenna, M. D., Ennis, E., Bickerdike, A., Coughlan, E. K., Broderick, T., Burns, C., McTear, M. F., Kuosmanen, L., et al. (2021). *Insights and lessons learned from trialing a mental health chatbot in the wild* [Paper presentation]. 2021 IEEE Symposium on Computers and Communications (ISCC) (pp. 1–6). IEEE, Athens, Greece.
- Procter, M., Lin, F., & Heller, B. (2018). Intelligent intervention by conversational agent through chatlog analysis. *Smart Learning Environments*, 5(1), 30. <https://doi.org/10.1186/s40561-018-0079-5>
- Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). *Using the Jupyter notebook as a tool for open science: An empirical study* [Paper presentation]. Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL '17), Toronto, Ontario, Canada (pp. 338–339). IEEE Press.
- Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151, 102630. <https://doi.org/10.1016/j.ijhcs.2021.102630>
- Sanguinetti, M., Mazzei, A., Patti, V., Scalerandi, M., Mana, D., & Simeoni, R. (2020). Annotating errors and emotions in human-chatbot interactions in Italian. In Stefanie Dipper & Amir Zeldes (Eds.), *Proceedings of the 14th Linguistic Annotation Workshop* (pp. 148–159). Association for Computational Linguistics.
- Sezgin, E. (2024). Redefining virtual assistants in health care: The future with large language models. *Journal of Medical Internet Research*, 26, e53225. <https://doi.org/10.2196/53225>
- Shih, E., Aylward, B. S., Kunkle, S., & Graziani, G. (2022). Health-related quality of life among members using an on demand behavioral health platform: Pilot observational study. *JMIR Formative Research*, 6(7), e35352. <https://doi.org/10.2196/35352>
- Siddique, S., & Chow, J. C. L. (2021). Machine learning in healthcare communication. *Encyclopedia*, 1(1), 220–239. <https://doi.org/10.3390/encyclopedia1010021>
- Vaira, L., Boicchio, M. A., Conte, M., Casaluci, F. M., & Melpignano, A. (2018). *MamaBot: A system based on ML and NLP for supporting women and families during pregnancy* [Paper presentation]. Proceedings of the 22nd International Database Engineering & Applications Symposium, New York, NY, USA (IDEAS '18) (pp. 273–277). Association for Computing Machinery, Villa San Giovanni, Italy.
- Vanian, J. (2023). *ChatGPT and generative AI are booming, but the costs can be extraordinary*. <https://www.cnn.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html> (accessed February 21, 2024).
- Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X., & Wang, H. (2021). CASS: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–31. <https://doi.org/10.1145/3449083>
- Xu, L., Sanders, L., Li, K., & Chow, J. C. L. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer*, 7(4), e27850. <https://doi.org/10.2196/27850>
- Yadav, D., Malik, P., Dabas, K., & Singh, P. (2019). Feedpal: Understanding opportunities for chatbots in breastfeeding education of women in India. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–30. <https://doi.org/10.1145/3359272>
- Zhan, X., Xu, Y., & Sarkadi, S. (2023). Deceptive AI ecosystems: The case of ChatGPT. <https://doi.org/10.1145/3571884.3603754>
- Zhang, X., Zhu, R., Chen, L., Zhang, Z., & Chen, M. (2024). News from messenger? A cross-national comparative study of news media's audience engagement strategies via Facebook messenger chatbots. *Digital Journalism*, 12(3), 336–355. <https://doi.org/10.1080/21670811.2022.2145329>

About the authors

Diana Liebetrau is a Business Consultant at Mindera, where she specialises in identifying and resolving friction points in customer journeys. Her work focuses on aligning digital solutions with user needs through collaborative research, service design, and systems thinking. This work was conducted while Diana was a Master student in I.T. at the University of Cape Town, and a researcher at Fraunhofer Portugal AICOS.

Melissa Densmore is an Associate Professor in Computer Science at the University of Cape Town, where she leads the HCI Lab and the Hasso Plattner Institute Research School at UCT in ICT4D. Her research looks at community-based digital innovation for maternal and child health and community wireless networks.

Francisco Nunes is Senior Researcher at the Human-Centered Design department of Fraunhofer Portugal AICOS. His research is concerned with the user research, design, and evaluation of technologies for care, with special focus on self-care technologies.