Philip Weber*, Faisal Mahmood, Michael Ahmadi, Vanessa von Jan, Thomas Ludwig and Rainer Wieching

Fridolin: participatory design and evaluation of a nutrition chatbot for older adults

https://doi.org/10.1515/icom-2022-0042
Received November 25, 2022; accepted February 2, 2023;
published online March 14, 2023

Abstract: In recent years, emerging approaches to chatbot-guided food coaching and dietary management, while innovative and promising in nature, have often lacked long-term studies. Therefore, with this work, we pursued a participatory approach within a design case study to the co-design and development of a nutrition chatbot for elderly people. Overall, 15 participants were directly involved in the study, of which 12 participated in the initial co-design phase, seven in the first real-world evaluation study over four weeks, and three in the second evaluation study over seven weeks. We contribute to the fields of Human-Computer Interaction by showing how the long-term use of such a chatbot in the area of nutrition looks like, which design implications arise for the development of nutrition chatbots, and how a participatory design approach can be realized to design, evaluate and develop nutrition chatbots.

Keywords: chatbot; nutrition; participatory design.

1 Introduction

Over the last decades, global life expectancy has risen to 72 years [1]. While this is a positive development, the last years are often spent in poor health, unhappiness, or poverty [2]. Therefore, the World Health Organization coined the term “active aging”, which is defined as “the process of optimizing opportunities for health, participation, and security in order to enhance quality of life as people age” [2]. In connection to health, nutrition is an essential criterion for the elderly. With increasing age, the demands on nutrition change, which may lead to malnutrition if the diet is not adjusted accordingly. The consequences are unwanted weight loss/gain or deficiencies, resulting in an accelerated need for care and severely affecting the life quality of older adults. Malnutrition significantly increases the length of hospital stay, readmission rate, the risk of complications during surgery and leads to a higher mortality rate [3, 4]. Along with other determinants such as physical activity and non-smoking, a healthy diet can prevent chronic illness and increase the quality of life while decreasing frailty and risk of sarcopenia [2]. While the risk of malnutrition can be reduced by regular visits to nutritionists and medical experts, the availability and effort required for traditional on-site or telephone consultations is usually high. As a result, there have been recent research initiatives to support a healthier diet: From using gamified systems [5, 6], to understanding food journaling through social media [7], to using small “pleasurable troublemakers” [8] to get people to eat more mindfully [9, 10], to using conversational agents and chatbots [11–13]. The focus of this research is on chatbots due to their potential to provide older adults with convenient access to nutrition information and resources, as well as the ability to track their progress and receive personalized recommendations [14, 15]. The general concept of chatbots involves terms that are often used interchangeably, such as conversational agents (CAs), virtual assistants, and virtual agents. That said, the core concept behind these terms is essentially the same; Dale [16] defines it as “achiev[ing] some result by conversing with a machine in a dialogic fashion, using natural language”. In the context of this study, our understanding of chatbots (unless otherwise specified) means text-based, specialized digital assistants.

While most of chatbot interventions are usually evaluated with a focus on interaction modalities and within a short period of time (e.g. [12]), we are interested in how the interaction between older adults and the chatbot changes over time. In the first weeks of testing new technology, users’ engagement often is high, but often quickly fades

***Corresponding author: Philip Weber**, Cyber-Physical Systems, University of Siegen, Kohlbettstr, 15, 57072 Siegen, Germany, E-mail: philip.weber@uni-siegen.de. https://orcid.org/0000-0003-3537-5753
Faisal Mahmood and Thomas Ludwig, Cyber-Physical Systems, University of Siegen, Kohlbettstr 15, 57072 Siegen, Germany
Michael Ahmadi, **Vanessa von Jan and Rainer Wieching**, Information Systems and New Media, University of Siegen, Kohlbettstr 15, 57072 Siegen, Germany

in the following weeks [17, 18]. To examine the long-term perception and interaction with chatbots by elderly people, we set up a participatory design study following a design case study approach [19]. Older adult participants tested two functional prototypes for at least four weeks each, allowing us to answer the research question, *"How might a long-term use of a nutrition chatbot look like?" (RQ1)*.

Based on the findings of our study, we derive design recommendations to answer a second research question, *"What are design implications for the development of nutrition chatbots?" (RQ2)*. In particular, we address the areas of user control and proactivity, expectation management and onboarding, input and output modalities and the personification/anthropomorphization of nutrition chatbots.

In addition, we share insights and methods that proved to be effective during the design and evaluation process, thereby addressing the research question, *"How can a participatory design approach be effectively implemented to design, evaluate, and develop nutrition chatbots?" (RQ3)*.

The structure of this paper is as followed: First, we present related work to provide an overview of current research on voice- and text-based CAs in the context of well-being and dietary change for older adults and outline the research questions that guided this study. We then describe our methodological approach, laying out in detail the setup of our two user studies. After presenting our insights, we compare the initial with the redesigned prototype and interpret our findings in light of existing literature. Furthermore, we provide recommendations for other designers of chatbots and highlight the limitations of our study. Lastly, we summarize our findings and present further areas of research.

2 Related work

2.1 Application fields, opportunities and risks of conversational agents in healthcare for older adults

In recent years, there has been increased interest in researching the impact and the design of virtual coaches, e-coaching systems, and CAs in healthcare, which has led to a number of systematic literature reviews [14, 15, 20]. While Kocaballi et al. [15] focus on the personalization of CAs in healthcare, Laranjo et al. [20] provide a general overview of the use of CAs in healthcare contexts and place a great emphasis on the underlying scientific methods used in previous studies. Particularly noteworthy is the study by Kamali et al. [14], who examined virtual coaches and e-coaching systems as a means of improving the well-being of older adults. They promote 'personalization' as the primary intervention technique to promote behavior change. In the majority of studies, users could set personal goals and receive tailored messages. Furthermore, virtual coaches often enabled users to track their progress through feedback from the system or self-tracking. In many cases, reminders were utilized to remind users to enter their data regularly. Additionally, virtual coaches rewarded users with praise. Overall, virtual coaches were generally accepted by older adults. Initial findings show that e-coaching interventions can have a positive effect, although it often remains unclear whether these benefits can be maintained in the long term. While Laranjo et al. [20] reported that the most frequent issues with CAs were the agent's parsing of spoken language and dialogue structure. Wiratunga et al. [21] argue that the voice agent is ideal for older adults as they are less accustomed to texting on smartphones and Pradhan et al. [22] found that older adults who were more socially isolated were most likely to personify the voice assistant (VA) and find satisfaction in their social needs through surface-level interactions with *Alexa*.

2.2 Voice assistants to increase the well-being of the elderly

Most recent published research in the field [23–26] seem to pursue mostly voice and speech-based approaches. For instance, the research group around El Kamali et al. developed a smart speaker [24] based on a variety of participatory co-design methods together with older adults to improve their well-being in the long term [27]. Four different domains (physical activity, nutrition, social and cognitive) were designed, which were linked to different coaching plans and sub goals. As sub-goals from the nutrition domain, achieving a healthier diet and losing body weight in particular were most frequently during the evaluation of the system [28]. In an interview study, Gudala et al. [23] focused on the advantages and barriers of a VA to provide older adults with information about medication and remind them to take it regularly. Similarly, approaches to health information management systems were developed using participatory design, again with the participating older adults designing strongly in the direction of speech and voice-based systems [29]. Previously, Yaghoubzadeh et al. [30] showed that the use of participatory methods increases the acceptance of virtual agents among older adults. Using a privacy-by-design approach, Seiderer et al. [26] developed a speech assistant to make nutrient information for food more conveniently available to older adults. Razavi et al. [25] focus in their work on enabling realistic communication with a virtual avatar to improve the well-being of elderly people suffering

from social isolation or social anxiety. However, there are still many open research questions in voice assistant design, such as perceptions and barriers regarding the use of VAs, or the effect of anthropomorphized design on older adults [31].

2.3 Chatbots to promote healthy lifestyles

With regard to text-only chatbots or chatbots that run via messenger services and thus often also allow for at least some additional speech input and output, there is less research that focuses specifically on the well-being and health of older adults [13, 32, 33]. While studies that used participatory design to strengthen the health of adolescents [34] or to improve the mental well-being of individuals living in rural areas [35] do exist, they do so without specifically considering perspectives of older adults. Graf et al. [12] developed "Nombot" which runs on the instant messenger service Telegram. Users can monitor their food intake and track their weight by writing to the chatbot. Users receive reminders to enter their meals; the interval is based on their motivational type. In a user study, participants preferred the chatbot over a conventional food tracker application. Similarly, Casas et al. [11] tried to simplify the process of keeping a food diary by introducing "Rupert". When first interacting with this chatbot, users select one of two goals, namely reducing their meat consumption or increasing their fruit- and vegetable consumption. In a month-long user study, authors found that although only 11% reached their goals, 65% of participants still improved their consumption overall. When looking at chatbots specifically designed for older adults, many systems are designed to combat loneliness and isolation by providing a virtual companion. Ring et al. [32] developed a proactive CA consisting of a touchscreen with an avatar for this purpose. It assessed the user's well-being, engaged them in small talk, and provided motivational stories to encourage them to move more. Thus it led to a reduction in loneliness as measured with the UCLA loneliness scale [36]. The lonelier a participant, the more often they talked to the CA [32]. Also focused on mental well-being, Ryu et al. [33] developed a mental healthcare chatbot called "Yeonheebot" and showed that the use of buttons is preferred as many of the older adults struggled to use the keyboard. One of the few examples of a chatbot designed and evaluated with and for older adults, which also focuses on diet change, is the chatbot "Paola" [13]. Thus, Maher et al. [13] showed that older adults were encouraged to exercise more and follow a Mediterranean-style diet over a period of 12 weeks. *Paola* led participants through the introduction, answered participants' questions around the clock, and provided weekly check-ins. Paola did not proactively contact participants; instead, they received weekly emails reminding them to check in with the chatbot. Overall, the intervention was deemed a success. Not only did participants exercise almost 2 h more each week, but they also increased their adherence to the Mediterranean diet.

2.4 Navigating the challenges of novelty and Hawthorne effects in participatory design studies

In 1993 participatory design (PD) was coined as a *"rich diversity of theories, practices analyses, and actions, with the goal of working directly with users (and other stakeholders) in the design of social systems including computer systems […]"* [37]. In recent years, there has been an increased focus on involving older adults in the design process of technology aimed at improving their well-being. Participatory design and co-design methods provide valuable insights into the needs and preferences of older adults, leading to more effective and user-centered designs (e.g. [38–40]). However, there are challenges that need to be addressed to ensure that the findings of participatory design studies are accurate and reliable. Two of the main challenges are the novelty [17] and Hawthorne effects [41, 42].

Novelty effects refer to the phenomenon in which participants in a study typically show increased enthusiasm or positive behavior in the early stages of the study due to the novelty of the experience with the (IT) artifacts. However, the impact of novelty on the participant's experience can be unpredictable, as it can also lead to negative experiences [17]. To reduce the novelty effects, it is recommended to allow participants a period of familiarization with the technology, or at least to point out the weaknesses of the technology in an introductory session [17]. Additionally, most studies (e.g. [43, 44]) attempt to observe the novelty effect on their results by choosing a longer study period, such as three to four months. The *Hawthorne effect* refers to the phenomenon that participants in a study change their behavior simply because they know they are being observed [41, 42]. To address the Hawthorne effect, it is important to reduce the researchers involvement [44]. Since, however, researchers especially in longitudinal participatory studies run the risk of developing close relationships with participants that could increase the Hawthorne effect on their study, Winkle et al. [45] recommend regularly involving new/unbiased users in the design process.

3 Research approach and methodology

Within this chapter, we present our identified research gap and the overarching research and design framework. In the past, potential (future) users were often times neither questioned before the development nor were they involved in the design process. Previous work (e.g. [38–40]) has successfully demonstrated time and time again that participatory design and co-design with older adults provide interesting insights into the use of technology and its design of it. Similarly, we anticipated that integrating users throughout the design process can lead to a nutrition chatbot that fits their personal needs. Thus, our chatbot was developed using participatory methods [46]. The design case study as proposed by Wulf et al. [19] acted as our overarching framework. With the chosen approach we were particularly interested in examining how a nutrition chatbot for older adults should be designed to encourage regular interactions as well as what should be considered when developing a nutrition chatbot for older adults.

We recognize that we have a similar starting point to recently published work, such as the studies by Seiderer et al. [26], who used a privacy-by-design approach to develop a speech assistant to provide information to older adults based on food packaging, Angelini et al. [28], who used co-design methods to develop a chatbot, smart speaker, and an app to improve the well-being of older adults through various coaching activities, Martin-Hammond et al. [29], who used a participatory design study to elicit requirements from older adults for intelligent assistants and used them to design concepts for obtaining health-related information, and Maher et al. [13], who conducted a long-term medical evaluation over 12 weeks of a virtual health coach to increase physical activity and diet intervention of older adults. However, because these studies were all ongoing during our study period, we were unaware of the endeavors and outcomes at the time we designed and planned our study. Interestingly, many of these current approaches appear to focus purely, or at least primarily, on voice interaction [24, 26, 29]. In comparing recent studies with our findings, such and other intriguing similarities and differences are revealed and discussed in Chapter 9 in relation to design implications. Our work provides insights and perspectives into the design of CAs with a focus on nutritional support for the elderly, thereby contributing overall design implications for the design and use of CAs and virtual coaches for healthy aging and well-being.

3.1 Pre-study

Before developing the first chatbot prototype, we conducted an extensive pre-study to unravel users' needs and their current use of technology for monitoring their eating habits. We conducted a group interview in each social setting to capture any changes that occurred throughout the project. Furthermore, we observed four meetings of a cooking-course for older adults. This observation was intended to show what problems arise in preparing and eating meals for our older participants. Additionally, we conducted interviews with two nutritionists (N1 and N2). This resulted in one of these nutritionists providing feedback and domain advice throughout the chatbot development.

3.2 Design phase

We wanted to include participants' ideas and requirements in the design process and therefore conducted four consecutive workshops, throughout October to December 2019, followed by a two-month phase (until the end of February 2020) to complete the first prototype, resulting in a total of five months of initial design. The first workshop consisted of a group discussion primarily focused on the participants' expectations of the chatbot's functionality. In the second workshop, we connected a simple chatbot to the participants' devices, so they were able to interact with it. Afterward, we answered their questions and provided a more in-depth explanation of chatbots. The session ended with discussing potential functionalities a nutrition chatbot could have. Within the next workshop, we presented participants an initial paper prototype. Following a Wizard of Oz approach [47], one moderator took on the chatbot's role by reading out pre-formulated statements and reacting to participants' input. The last workshop of the design phase was focused on the content of the chatbot. Before the session, we asked participants to come up with five nutrition-related questions of their interest. Participants then presented their questions to the group, which naturally led to a discussion of nutrients, vitamins and their connection to common age-related problems (e.g., dementia, declining vision or diarrhea).

3.3 First evaluation and re-design

For the first prototype evaluation in early March 2020, we rolled out the nutrition chatbot to seven participants. Participants then started the conversation and completed the chatbot's tutorial independently; we only interfered when questions arose. Afterward, participants could try out other features. On average, this rollout meeting lasted 1 h. We

took notes throughout and finished them after each session. Afterward, the participants were allowed to use the chatbot for one month unassisted and without any further interaction with the research team. One month later, we reached out again and scheduled individual phone calls with all available participants. We conducted semi-structured interviews with four participants and received written feedback from two participants. Moreover, we collected quantitative data in the form of chat records from all seven participants. Additionally, based on the demonstration of the first prototype, the nutritionist also provided feedback and offered suggestions for further improvement. We analyzed the results of the first evaluation phase and subjected the prototype to an 8-month re-design phase.

3.4 Second evaluation

For the evaluation of the second prototype in late November 2021, we recruited three participants. After completing the tutorial, participants could try all the implemented features for seven weeks. Afterward, we again conducted a semi-structured interview to find out about their impressions of the chatbot. All of the final interviews were again administered over Zoom. Participants reported how well they could interact with the chatbot and how their perceptions changed over time. Analogous to the evaluation of the first prototype, we analyzed the chat logs quantitatively (see chapter 3.6).

3.5 Analysis of qualitative data

Through thematic analysis [48] of the transcribed interviews, the authors first constructed individual inductive codes and, following regular consultation with each other, merged these into a single coding scheme. To better understand how participants interacted with the chatbot during the evaluation period, we also analyzed the chatlogs through content analysis: When users sent short messages consisting of one or two words, directly related to a feature of the chatbot, we interpreted this as an indication that they regarded the chatbot merely as a machine or a tool. On the other hand, if they greeted the chatbot, phrased longer sentences with frills (which were unnecessary for the feature itself), or inquired after the chatbot, we saw this as signs of a more human-like style of dialogue.

3.6 Analysis of quantitative data

To evaluate chatbot performance, we extracted some quantitative metrics from the chat logs. During our literature review, the most commonly used metrics to measure user engagement were interaction time and message count. The latter is defined as the number of messages exchanged between a participant and the chatbot. Both metrics can be calculated for the entire interaction with the chatbot or for each conversation separately. Right now, there are no established criteria for the quantitative evaluation of a chatbot. While Zhou et al. [49] only report the longest interaction time of a conversation, Stephens et al. [50] additionally report the shortest and average duration of a conversation as well as the total interaction time. On the other hand, Jain et al. [51] only report the total number of messages and the interaction time. In addition, they used the character length of user statements as a measure of user engagement. We decided to analyze the average number of messages and total interaction time, as well as the average, minimum, and maximum number of messages and interaction time per conversation. To calculate the averages, we used the median as it is rather robust to outliers. For interaction time, we calculated it as follows: If an hour or more elapsed between the last message and the user's response, the messages belonged to different conversations. If more than 5 min passed between messages, we assumed that the user took a short break from the conversation. Interaction time is resumed as soon as the user resumes the conversation (and 1 h has not passed). In addition, we only considered interaction time after the rollout meeting for several reasons. First, we wanted to understand how engaging the chatbot would be on its own, without any human guidance. Second, the interaction time in the rollout meeting would be highly unreliable because there were many pauses during the conversation that were not captured in the logs.

We also wanted to quantify how well the chatbot understood the user's requests overall. We therefore calculated the appropriateness score, which is mentioned in a meta-study on chatbot evaluation by Maroengsit et al. [52]. To calculate this score, each response to a user request was rated by us as either "suitable", "neutral", or "unsuitable".

For the second prototype, we took some additional measurements due to the added capabilities of the chatbot. We calculated how many conversations a user started compared to the chatbot. When the chatbot proactively sent a message, we only counted it as a conversation starter if a user responded to the request. From this measurement, we wanted to see what the ratio of user-initiated conversations to chatbot-initiated conversations was and how it changed over time. We also wanted to see how users interacted with the buttons. We therefore looked at the ratio of typed messages to button presses.

3.7 Participants

The study at hand took place in Germany. Our participants (Table 1) have been at least 60 years old and, to our knowledge, identify as cisgender. The university's ethics board approved our research activities and participants received an explanation and signed a letter of consent at the start of the study. Twelve participants took part in the design phase and/or pre-study, while seven seniors took part in the evaluation of the first prototype version. The most common reasons for discontinuing the study were lack of time, illness or disinterest in the topic of nutrition. Most participants were already part of previous or existing research projects with our university and were thus recruited largely through well-established relationships with different rural and urban care facilities and through newspaper announcements. For the second prototype, we worked with three participants but analyzed their usage in detail. The reason for this is that recruiting more participants was infeasible because of the COVID-19 pandemic. During a time of turmoil, we faced severe challenges in reactivating previous participants in the first place but were also concerned about ensuring the safety of a potentially vulnerable user group. The participants for the second prototype phase were acquired from the researchers' network. We are aware that the small number of participants is a limitation of our study (see chapter 10).

4 Results from pre-study

Overall, more participants struggled with unwanted weight gains than losses. All participants were familiar with the Body Mass Index (BMI) and had calculated their measurements in the past. However, for many of them, this measurement was not very relevant. Most participants were willing to adjust their diet to lose weight or have more balanced meals. Overall, there seemed to be a trend of reducing alcohol intake over the years. For some, drinking alcohol left them feeling nauseous; others were hesitant to combine their medication with alcoholic drinks. However, we found that older adults were not drinking enough water and were not consuming enough vegetables and fruits. In general, most participants referred to online health information critically and often compared different sources or got a second opinion. That said, they were not actively trying to get health information.

When the nutritionists were asked about their counseling methods, they replied with traditional counseling methods (e.g., face-to-face meetings, food logs on paper). One of the most important tools in nutrition counseling is keeping a food diary. It allows clients to monitor their eating habits which in itself leads to awareness and change. One potential use case is reminding users to drink or eat regularly. In their opinion, it was imperative to use a friendly and encouraging tone; the chatbot should feel more like *"a friend"* (N1) or *"a daily companion"* (N2) than a commander. According to N2, prompts would be even more effective if they named advantages. Generally, N1 and N2 only hand out recipes when clients explicitly request them. In N1's experience, this usually happens when she suggests foods that clients have not prepared before. Otherwise, it is barely necessary as *"there are recipes in abundance"* (N1). For N1, it was essential that the chatbot should not provide incorrect nutritional advice. Any advice regarding the user's

Table 1: Basic data about the participants.

ID	Gender	Age	Living situation	Participation
PN1-1	Male	68	Lives alone	Pre-study, design phase, first evaluation
PN1-2	Male	78	Lives with partner	Pre-study, design phase, first evaluation
PN1-3	Female	74	Lives with partner	Pre-study, design phase, first evaluation
PN1-4	Male	68	Lives with partner	Pre-study, design phase, first evaluation
PN1-5	Male	68	Lives with partner	Pre-study, design phase, first evaluation
PN1-6	Male	74	Lives with partner	Pre-study, design phase, first evaluation
PN1-7	Female	73	Lives with partner	Pre-study, design phase
PN1-8	Male	76	Lives with partner	Pre-study, design phase
PN1-9	Female	81	Lives alone	Design phase
PN1-10	Male	72	Lives with partner	Design phase
PN1-12	Female	78	Lives with partner	Design phase
PN1-13	Male	63	Lives alone	Design phase, first evaluation
PN2-1	Female	60	Lives with partner	Second evaluation
PN2-2	Female	66	Lives alone	Second evaluation
PN2-3	Female	69	Lives with partner	Second evaluation

medications was completely off-limits. Furthermore, she requested that if any health indicators were out of the ordinary for an extended period (e.g., consistently high blood sugar levels), the chatbot should refer users to contact a professional. N1 added that the chatbot would be better suited for primary prevention, dealing with healthy people.

5 Design and implementation

5.1 Laying the foundation: from simple food tracking, healthy recipe suggestions, general nutrition advice to body weight tracking

The nutrition chatbot should assist older adults living at home. According to nutritionists, the chatbot should only be used for primary prevention and avoid giving medical advice. The chatbot should focus on achieving a balanced diet in everyday life, and should take users' preferences into account. Participants expected to be able to track their food intake with the chatbot, which was by far the most requested feature. Thus, they suggested *"to make [the logging process] as simple as possible – otherwise you quickly lose interest"* (PN1-1). It was suggested that users would need to enter their meals for two weeks so that the chatbot could learn their eating habits, thus eliminating the need for weighing ingredients in the future. After logging their meals in the chatbot, participants wanted to receive nutritional information. PN1-4 requested a comparison of the caloric intake and expenses during the day. Also, many participants wanted to find healthy recipes with the help of the chatbot. Apart from specific ingredients, participants wanted to filter search results by regionality, recipe difficulty, and preparation time. PN1-1 and PN1-3 (independently) explained that they did not feel like trying complicated recipes *"that take one or two hours in the kitchen"*.

A common pattern among users' nutrition questions showed that they were interested in getting recommendations according to their priorities, such as how much water or fruit they should consume, what they should eat to avoid increasing uric acid, or information about the nutritional value of food, e.g. vitamin C. Moreover, participants were interested in portion sizes. PN1-7 wanted to know how much dark chocolate and cookies she could eat without gaining weight.

For the first prototype, we focused on four main features to enable keeping a food diary, asking for nutrition information, searching for recipes, and tracking users' weight. Additionally, we added two features with a guided onboarding tutorial, where the chatbot introduces itself as "Fridolin" and a help feature, which can be accessed anytime through variations of the word 'help' to which the chatbot responds with a list of its features and how they can be triggered.

5.2 Language and persona specifications of the chatbot

In the design workshops, participants voiced many suggestions on how a chatbot should interact with them. They agreed that the nutrition chatbot should use colloquial language and a simple sentence structure, avoiding *"politicians' language"*, as PN1-7 put it. PN1-10 stressed the importance of using simple words: *"For us older people it is important that you omit foreign words. And if you do use them, that you also provide an explanation […] [or look up] if there is maybe also a German word for it"* (PN1-10). Participants wanted a *"kind but direct"* approach. For one participant, PN1-5, the element of humor played an important role. He said that *"it would definitely motivate me if [the chatbot] is funny."* Additionally, PN1-4 wished to receive praise. PN1-3, on the other hand, wanted to receive constructive criticism if she had overindulged in a particular food.

5.3 Platform, interaction modalities and implementation

Participants agreed that the messenger-like interface felt is easy to use, especially if people are already familiar with services like WhatsApp or Telegram. For instance, PN1-1 stated: "*Whoever understands WhatsApp – and that is not very difficult – will understand [this chatbot] as well*" (PN1-1). However, when we presented possible chatbot functionalities to the participants, most agreed that the chatbot should focus on text. Though PN1-13 added: *"I can imagine it better, comprehend it better than if it is read aloud to me"*. Additionally, pictures should be displayed where suitable, e.g., showing the completed dish when looking up a recipe. Throughout the design workshops, participants expressed interest in a secure application. Participants wanted to know where all their conversations with the chatbot would be stored. Overall, utilizing a messenger app seemed to be the most reasonable option for our participants as they were already familiar with Telegram or WhatsApp. Of these two options, Telegram was selected because it's easy creation and maintenance of chatbots. For content management and intent handling, we created a Dialogflow agent. To store user information long-term and provide a more personalized experience, we used several other components besides the Telegram chatbot (Figure 1). The middleware consists of a Nginx web server, a NodeJS application and a MongoDB

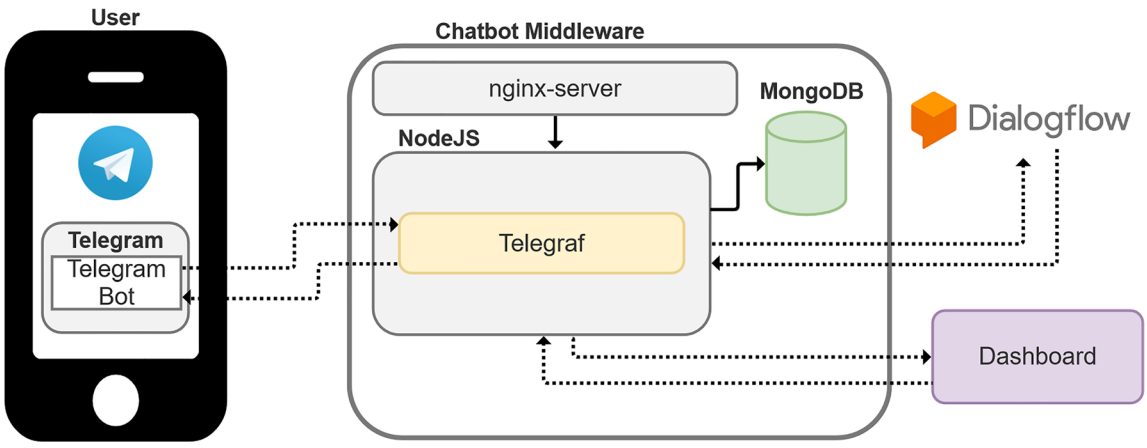

Figure 1: Chatbot middleware architecture and connected services (first iteration).

database. The latter is used for storing all user information, using the Telegram username as an identifier. Moreover, the Swiss nutrition database (https://naehrwertdaten.ch/en/) containing nutritional information is also saved in the MongoDB.

6 Results of the first study

6.1 Overview of the interactions: impact of COVID-19, misunderstandings and non-proactive chatbot design

The results show that a third of all messages sent to the chatbot was related to the recipe search. Of these 55 messages, 20 messages were unique requests for a recipe, whereas 28 messages were counted as repeated attempts, meaning that participants rephrased (or repeated) a previously stated recipe search. As this high number indicates, several attempts were often necessary to find the desired recipe. The second most used feature was the nutrition values feature with 28 messages, followed by nutrition questions. The weight feature was only used once after the kick-off meeting to update a participant's weight. None of the participants accessed the overview of weight development.

Another way to capture user engagement is the interaction time. In total, participants used the chatbot for approximately 2.5 h after the rollout. The evaluation of the first prototype of the nutrition chatbot provided many valuable insights into how older adults interact with such a chatbot. However, many shortcomings of the utilized chatbot and its implementation also became apparent. We noticed that many participants only interacted with the chatbot at the beginning and the end of the evaluation period, shortly before we interviewed them again. This usage pattern could be explained initially by the novelty effect [17], i.e. the curiosity to use new technology, and in the end by a form of the Hawthorne effect [41, 42], i.e. the influence on behavior because our participants were aware of being part of a research study. But also, as the beginning of March 2020 marked the outbreak of COVID-19, participants presumably had other things on their minds. While the ongoing pandemic might have heavily influenced their behavior, we believe additional factors have also contributed to less frequent interactions with the chatbot. One factor might be the limited understanding of the chatbot. Two participants, PN1-1 and PN1-5, entirely abandoned the chatbot after a short, unsuccessful message exchange. For the remaining participants, the number of messages and interaction time seemed to be less a sign of user engagement but more a mark of the user's persistence. If the chatbot could understand users better, they might have enjoyed the interaction more and would potentially return more often.

Looking at the increase in messages after we contacted participants to schedule the final interview, it seems one of the main reasons for less interaction is that they had forgotten about the chatbot's existence. The first prototype never initiated conversations because participants voiced concerns about being contacted too often. This passive behavior had additional downsides. By designing the chatbot so that users have complete control over the interaction, many features remained hidden if the user never actively explored the chatbot.

6.2 Onboarding experience

While the set-up of the chatbot in general went smoothly, the biggest obstacle during the installation for the participants was deciding which permissions should be set for the Telegram messenger. After the successful setup of the

Telegram Messenger platform, the chatbot's tutorial proved to be valuable for introducing users to the operation of the chatbot. All participants grasped the advanced functionality of buttons and the speech-to-text (STT) option. Participants immediately understood who was texting them and what the general purpose of the chatbot was. Aside from minor insecurities mostly based in vague wording, the tutorial seemed well-structured. Merely the ending seemed to be too abrupt as participants were unsure how to continue the conversation. While there was an overview of all the chatbot's abilities, it was hidden within the help feature. We anticipated that participants would open this feature directly as it was the only feature mentioned explicitly in the tutorial and would explore all the other features from there. However, the participants appeared to be hesitant to type anything when they did not know the full scope of what the chatbot would understand. In retrospect, users needed a better jumping-off point, which provided them with a short overview of all abilities and a possibility to get started quickly. All participants but one completed the tutorial without (almost) any assistance.

6.3 Conversation breaks and communication style

In the tutorial, the chatbot led the conversation and asked closed questions, but the roles were swapped after finishing it. Now, users phrased their own requests and the chatbot needed to react appropriately. If, however, the chatbot was unable to recognize any intent, the default fallback was triggered. When users then rephrased their statement without knowing the structure expected by the chatbot, there was a good chance that it would not recognize their request again. The closed questions in the tutorial invited short one-word answers from all participants that were understood most of the time. When participants tried the other features afterward, their requests grew longer and the chatbot understood them less well. Some participants noticed that shorter statements were better understood and began to use keywords over time. For others, this lack of understanding became demotivating. Participants often combined several intents into one message (e.g., greeting the chatbot, making a diary entry and then asking for the calories of said meal). The chosen chatbot framework only detects the most plausible intent. In the worst-case scenario, this combination of different intents then led to messages that cannot be assigned to any intent. Conversely, participants also sometimes added information to their previous request; thus, one cannot assume that an intent is completed after one message. Another difficulty were 'confirm' statements (i.e., some form of confirmation after a task is completed or information is given) or corrections after falsely identified intents. Both of these interactions are important elements of human conversation, indicating understanding or pointing out misunderstandings, therefore almost acting as 'conversational grease.' As the chatbot was unprepared to handle these statements, they almost always triggered the default fallback. Therefore, users that addressed the chatbot in a more human-like manner tended to be misunderstood more often. Over time, some participants thus began to treat the chatbot more like a machine than a conversation partner. This can also be seen in the tone that participants utilized when talking to *Fridolin*. Most started in a friendly tone, which then often turned into a neutral tone, reminiscent of web search phrasing. PN1-5, on the other hand, seemed to be testing the limits of the chatbot deliberately and did not hesitate to voice his opinions freely, even calling it *"stupid"* directly 'to its face'.

Most participants did not try all features, indicating that further proactive behavior from the chatbot is needed as encouragement. Moreover, participants often mistook the help overview for a selection menu and entered their request accordingly. While the examples helped some participants rephrase their requests, to others, it was often unclear why their request did not work.

6.4 Nutrition questions instead of food diary entries

In theory, the idea of keeping a food diary with a chatbot sounded very promising. Compared to other available nutrition apps, users could just write down what they have eaten without tediously searching every ingredient in a long list; the creation of a diary entry would only require little effort that would even decrease over time by saving past meals. In reality, the format of entries was very restrictive and, as none of the participants was able to make a completely valid entry, a further explanation was necessary. However, only PN1-2 and PN1-3 showed genuine interest in the functionality. The remaining participants only tried the food diary in the rollout meeting or not at all.

Similarly, the concept of answering nutrition questions basically suggested itself. A text-based chatbot should be able to answer some general questions. However, looking at the chat logs, it was not quite as simple as expected. Information was often falsely triggered and thus detrimental to recognizing other intents. When participants actually requested nutrition information, the value of the feature was still debatable because the information needed to be short to avoid sending users a wall of text. Furthermore, most follow-up questions and comments could not be answered so that a quick internet search provided more

new insights, as one participant reported. To participants, it probably was confusing why the chatbot did not understand such simple questions, which might have led to a decrease in their estimation of its intelligence. It might also have lessened the feeling of having a 'real' conversation.

7 Adjustments and redesign of the prototype

Initially, we had high expectations for our nutrition chatbot. Our vision was to create a tool to efficiently keep a nutrition diary, work on nutrition goals, search for recipes, and learn about nutrition by looking at nutrition values of specific ingredients and asking general questions. All these functionalities should lead to long-term behavior change. However, through the evaluation of the first prototype, we noticed that there were still fundamental issues with the chatbot, especially concerning its language understanding. For the second version of the prototype, we therefore mainly aimed to improve the recognition of user requests, the assistance to users and the general flow of conversation. Instead of further building upon a *'Swiss army knife'*, the chatbot's existing features should be revised and simplified whenever possible. The evaluation of the first prototype showed that the chatbot needed to be restructured fundamentally. As participants were best understood in the tutorial, we wanted the chatbot to guide the conversation even after users finished the tutorial. The design challenge for the second version was therefore to find a healthy mix between guiding users through the interaction while also allowing for free-text input, whenever possible. Moreover, some technical changes within the chatbot architecture were necessary to improve the understanding of the chatbot and ensure a higher level of privacy for users. In the second version of the prototype, the tutorial is updated, a new menu replaces the help feature (Figure 2), and the most important conversational elements are also taken into consideration.

7.1 Quick reply buttons

To offer more helpful feedback when users are not understood (i.e. no intent could be assigned to their statement), we also implemented quick reply buttons into the default fallback message. This way, users can select the feature that they were trying to trigger. The chatbot can then reinterpret their previous statement with the appropriate feature, thus eliminating the need to reenter the same request with a different phrasing. This could be especially useful for older adults as some struggled with typing. Although most

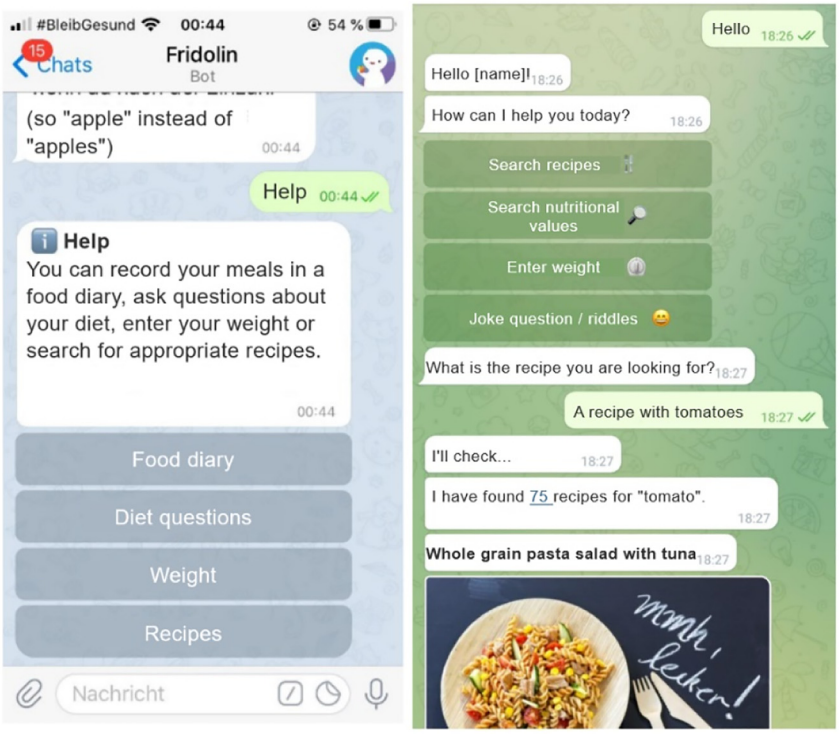

Figure 2: The help feature in the first version of the prototype (left) compared to the implementation in the second version of the prototype (right), which is also sent back within the default responses (Figure translated to English).

participants did not engage in small talk with the chatbot, they still adhered to some basic conversational structure. So, we decided that the chatbot should ask users whether they wanted to do anything else. This question was presented with the quick reply buttons to encourage users to explore other functionalities.

7.2 Riddles

During the design phase of the first prototype, some participants were interested in humor and suggested that the chatbot could tell them a joke once in a while. One participant compared it to looking at the Sunday newspaper, where comics and short jokes are printed at the end. Based on this analogy, the chatbot should present users with a humorous interaction once a week. Therefore, the following interaction flow was loosely based on the joke pattern by Moore and Arar [53]. Every Friday, the chatbot first asked users whether they were interested in hearing a riddle. If they agreed, the chatbot asked them a question. Users could then guess the answer. If their answer was correct, the chatbot praised them; otherwise, it sent them the correct answer. If users replied that it was funny, the chatbot asked if they want to hear another riddle. Furthermore, participants received options for asking for a riddle anytime.

7.3 Other adjustments

The recipe feature was a favorite amongst participants in the first prototype. Most problems were connected to users searching for specific dishes instead of ingredients or meal types. As the number of dishes is endless, it is impossible to make a custom entity consisting of all synonyms. Therefore, we decided to use a Part-of-Speech (POS) Tagger to label user requests and extract search terms that way to improve the recognition of ingredients and dishes. Participants wanted to have the option to receive recipes by email so that they can print them. We included an additional 'Send'-button beneath every recipe so that this feature is easily accessible. When a user presses the 'Send'-button beneath a recipe, the chatbot receives a shortened recipe URL and a keyword to identify it as an email request. From this callback message, the original recipe URL is reconstructed, and cheerio was then used to parse the recipe. The recipe is then sent via mail with the nodemailer package (Figure 3).

To increase the engagement with the weight feature, the chatbot was set up to ask users whether they are interested in entering their weight regularly. If they agreed, they could pick a day and time for a reminder to input their current weight. If they declined, the chatbot reminded them that they can enter their weight whenever they like. This way, undecided users might be persuaded to give it a try, while truly uninterested users are not forced. With regard to looking up nutritional values for specific ingredients, few changes were necessary. We just added a disclaimer and removed the daily recommended amount when users were looking up the sugar content of fruits.

We removed the food diary functionality from the second version of the prototype due to the low interest in the first evaluation and due to general questioning of the participants, whether they need this feature in the chatbot at all. When participants asked nutrition questions, they were often more interested in the chatbot's opinion of specific ingredients ('What do you think of figs?'). A database with an evaluation of ingredients would be necessary to answer these kinds of questions. Overall, the general nutrition questions created more misunderstandings than they did good; therefore, we decided to also exclude them from the second version of the prototype.

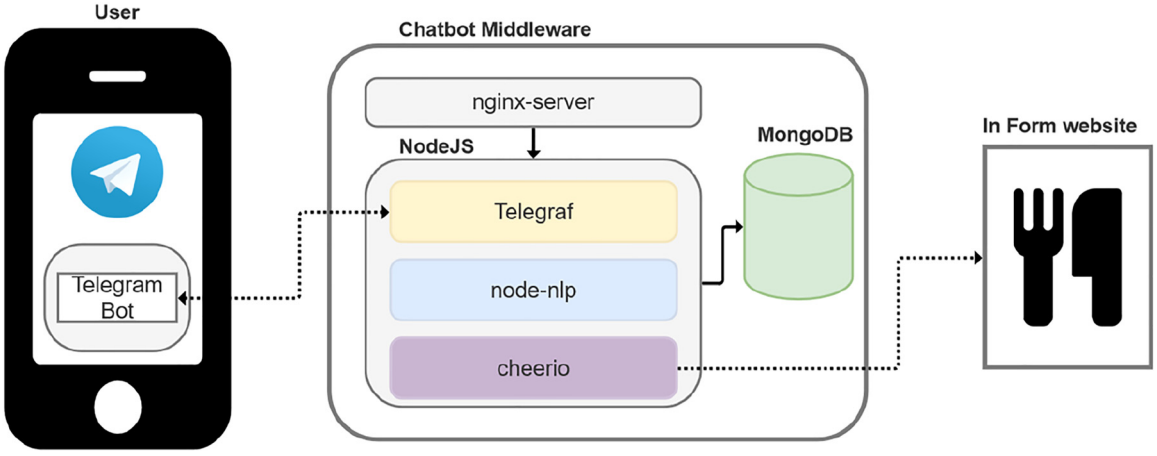

Figure 3: Overview of the chatbot architecture in the second version.

8 Results of the second study

8.1 Overview of the interactions

In the first four weeks, PN2-1 was by far the most active user with 514 messages (Table 2). This were more than twice as many messages as the second most active participant PN2-3 in that time span and even quadruple the number of messages for PN2-2. Over the remaining three weeks, the communication with the chatbot slowed down for both PN2-1 and PN2-2. PN2-1 exchanged 47 messages across three conversations and PN2-2 amassed 35 messages across four conversations. On the other hand, PN2-3 doubled the number of exchanged messages in the last two weeks, jumping to 421 messages from 210 messages. Looking at the number of messages per conversation, we can again observe large different patterns of use within the same participants. The minimal number of messages was two to three messages for all participants. The highest number of messages in a single conversation was 116 and 109 for PN2-1 and PN2-3 respectively; here, PN2-2 stood out with only a maximum number of 22 messages. Interestingly, the median number of messages was similar for PN2-2 and PN2-3, averaging at seven messages per conversation.

When looking at the distribution of messages, all participants sent many messages, particularly within the first days of using the chatbot. PN2-1 was especially active in the beginning: she interacted with the chatbot on eight out of the first ten days. But after that, she almost exclusively interacted with the chatbot on Fridays, when the chatbots wrote a message to all participants. In Figure 4, one can see that the latter half of PN2-1's and PN2-2's conversations were initiated by the chatbot. Overall, PN2-1 started 10 out of 19 conversations (approx. 53%), PN2-2 started 11 out of 18 conversations (approx. 61%) and PN2-3 started 16 out of 23 conversations (approx. 70%). But for PN2-3, the ratio of chatbot- and user-initiated conversations remained relatively steady throughout the whole evaluation.

In the tutorial, all participants wrote between 1.4 and 1.9 words per message. In the following stages, this number

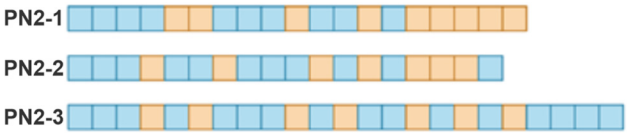

Figure 4: Overview of user-initiated (blue) and chatbot-initiated conversations (orange). One square represents one conversation.

was similar for PN2-1 and PN2-3. Only PN2-2 wrote considerably more words in the unassisted phase, arriving at almost three words per message on average. In comparison to the evaluation of the first prototype, the participants sent shorter messages. To evaluate the understanding of the chatbot, we again calculated the appropriateness score. Participants were very well understood in the tutorial as PN2-2 and PN2-3 only received suitable or neutral answers, whereas PN2-1 got two unsuitable answers. In the unassisted phase, we calculated an appropriateness score between 75% and 79% for all participants. All of them rated their experience with *Fridolin* as good, while acknowledging room for improvements. Two participants were even willing to continue using the chatbot beyond the evaluation period, which might be the biggest indicator of improvement over the first prototype.

8.2 Different initial expectations result in long-term changes in participants' interactions

Over time, PN2-2 and PN2-3 changed the way they interacted with the chatbot. Interestingly, their style of requests developed in opposite directions. In the beginning, PN2-2 wrote short key-word style sentences, which evolved into more elaborate, natural sentences. PN2-3, on the other hand, treated the chatbot as a real conversation partner and began to shorten her requests drastically when their initial sentences did not lead to the expected intents. According to PN2-2 herself, she saw the prototype rather as a tool and thus phrased requests as she would write them in a search

Table 2: Minimum, maximum and median number of messages per conversation.

ID	PN2-1		PN2-2		PN2-3	
	First 4 weeks	Full period	First 4 weeks	Full period	First 4 weeks	Full period
# Conversations	16	19	14	18	16	23
Min. messages	3	3	2	2	2	2
Max. messages	116	116	22	22	37	109
Median messages	23	19	7	7	7	7
Total messages	514	561	121	157	210	421

bar. When she realized that the chatbot could also understand her longer answers, they naturally grew more elaborate. PN2-3 seemed to expect actual conversation from the chatbot, and when it could not react to her questions, she was disappointed. Although both participants went through the same tutorial, it is noteworthy that they had different expectations.

8.3 Recipe searches and the puzzles were enjoyed widely, while questions on nutrition values were rarely asked

Overall, the recipe search was the most requested feature across all participants. However, for PN2-1 and PN2-3, the majority of messages in that category were just requests to show more recipe results (92 and 83 messages, respectively), whereas PN2-2 only used the recipe search three times. In general, most participants were happy (or voiced no complaints) about the presentation of recipe results. PN2-1 however requested that instead of only displaying one result at a time, the titles of recipes should be listed instead. Therefore, it might make sense to reconsider how recipes are displayed, especially if there are many search results. Alternatively, it needs to be explored how to enable users to switch between these two different display options. This way, users with more particular ideas on dishes could scan the results more efficiently, while others could browse through them leisurely. When talking to the participants about their needs, it sometimes seemed that they were looking for curated recipe recommendations that give them ideas on what they should cook instead of a recipe search. These recipes should be sent weekly, be healthy and contain seasonal ingredients. Messages related to the riddle feature came in second place with 73 messages in total. Participants always agreed to these requests and actively guessed the solution, so they showed more engagement than required. The riddle functionality also shows that not all features need to be useful or informative. While the older adults did not engage in small talk, they were open to playing a short game with the chatbot.

PN2-2 and PN2-3 did not use the nutrition value feature at all, and PN2-1 stopped using the feature later on. The issue seemed to be conceptual instead of technical, as the functionality was not accessed at all or abandoned after the participant had worked out how to phrase her requests. Within the final interviews participants explained to us that they were not interested in looking up values for specific ingredients because they were already knowledgeable about nutrition and felt that they were already cooking quite healthily. Still, PN2-1 and PN2-2 wanted to learn more about nutrition; however, they preferred general knowledge instead of concrete numbers, which the nutritional value feature offered.

8.4 Proactivity and notifications: a fine line between perceived annoyance and social familiarity

All three participants liked that the chatbot contacted them proactively. Several times, the analogy of an ‘old friend checking in’ was mentioned. Even though they made this comparison, participants wanted these proactive messages to have a purpose. Besides the current messages, participants suggested sending out recipe recommendations or alert users when a new recipe was uploaded to the database. Moreover, the chatbot could prompt functionalities that the user had not explored before. Initially, we planned two interactions with users per week, but as the scheduled weight check was never triggered, participants only received one weekly interaction. While all of them would have also tolerated a second proactive message per week, they stressed that they would not like to be contacted daily. The importance of not notifying users too often was especially apparent in PN2-2, who named notifications as one of the main reasons for not using chatbots in the future.

8.5 The effect of quick reply buttons on the input styles

With the addition of quick reply buttons in this version, we also investigated the impact such buttons had on the interaction: Almost half of PN2-1’s messages were button presses. PN2-2, on the other hand, mostly typed her messages to the chatbot, only using the buttons in 14 messages. For PN2-3, it was the other way around; only a third of her messages were typed. None of the participant used the STT option. It should be noted that the participants in the second evaluation were content with typing their messages, as they were already used to this input method from their previous messenger experiences. Even if participants did not use the provided buttons all the time, they were not useless because they reminded them what features were available. In the interviews, all participants were aware of what the chatbot could do.

9 Design implications and discussion

Overall, we identified four overarching design spaces with numerous interesting observed and reported user behaviors and contextualized these with related work (Table 3).

Table 3: Design spaces and observed/reported user behavior/expectations in our study with references to related work.

Design space	Observed/reported user behavior/expectations	Related work
User control and proactivity	Hesitations and low number of self-initiated uses of the CA	[30, 54]
	Older adults enjoyed and used the proactive CA more	[32]
	Entering a "Help" command provided some guidance	
	Better exploration of all features was achieved through a menu	
Expectation management & onboarding	The tutorial provided comfort and guidance	[55]
	Low interest to ask questions about nutrition to the CA	[13]
	High unrealistic expectations for language understanding and interactions after the tutorial negatively affected the user experience	
Input and output modality	No clear preferences for voice or text input	[21]
	Text output was preferred	[21]
	Overall preference of free text input over quick-reply buttons	[33]
	The use of buttons/menus helped older adults better navigate	
Personification/Anthropomorphization	Low interest in small talk with the CA	[22]
	Openness to playful/gameful approaches	[56]
	Short confirmation messages on user requests were desired	
	Jokes by the CA were well received	
	Simple language and avoidance of foreign words were requested	

9.1 User control and proactivity

Similar to findings from Yaghoubzadeh et al. [30], the older participants were initially not keen on the system reaching out proactively. Thus, in the first prototype, users had complete control over the system, apart from the tutorial. Information about the chatbot's capabilities was provided in the "Help" feature. While this feature provided guidance to some participants, it was insufficient in resolving other older adults' uncertainties. The success of the first prototype relied on the assumption that the older adults would, based on the help feature function, independently explore all features of the chatbot. However, as the evaluation showed, this was not the case. None of the participants explored all functionalities; many participants were even unaware what features were available. This is in line with findings from Trajkova and Martin-Hammond [54], where older adults did not find valuable use cases for interacting with *Alexa* but also made no effort to find new features. The proactive messages and the implementation of a menu in the second version led to more frequent interactions and exploration of features throughout the evaluation period. This is in some contrast to the study conducted by Ring et al. [32], who indeed reported that participants felt less lonely, were happier, and also felt more comfortable while interacting with a proactive CA, but unlike in our study, proactivity of the CA did not influence the duration of use or the frequency of use in their study.

9.2 Expectation management and onboarding

The success of the tutorial indicated that older adults interacted more successfully with a chatbot when it guided them through the conversation. Though the primary downside of starting the chatbot interaction with the tutorial might be that it raised unrealistic expectations for many participants. As user statements could be better anticipated, the older adults were understood best in the tutorial. When they used the chatbot after the rollout, its recognition of their messages was much lower, leading to frustration and uncertainty. This was exacerbated by the non-descript fallback messages of the first prototype. In the first prototype, the chatbot could also answer general nutrition questions. But similar to the study by Maher et al. [13], the older adults did not think to pose these kinds of questions to the nutrition chatbot. The few questions that were asked were taken directly from the catalog of examples provided in the help feature. In our study, we therefore understood this less as disinterest of participants in most unused functionalities, but rather, similar to Alnefaie et al. [57], see adequate onboarding (beyond the usual design elements of greetings and menu-driven questions) with chatbots as an open research question. In addition to visual approaches – such as the use of horizontally scrollable images (carousels) [58] or the integration of short videos [55] – we see promising approaches in regular (e.g. daily, weekly) reminders of an

as yet unused or randomly selected feature in the form of a proactive "quick tip". Such approaches could soften the transition from the initial onboarding phase to everyday use and thus avoiding the hard transition observed in our studies with the associated lower user experience.

9.3 Text as the preferred input and output modality

In their work, Wiratunga et al. [21] stressed the value of spoken natural language as the primary input method for older adults. The majority of CAs for older adults used speech as the sole method to communicate with the agent [20, 24, 26]. However, not all older adults are comfortable with speaking out loud to a machine [21, 33, 54]. In our research, we uncovered no clear preference for verbal or written input. In the evaluation of the first prototype, half of the participants used the STT-functionality. In the second evaluation, all users preferred typing their messages. But overall, providing different modes of input seems advisable as the preference might even change over time (e.g., due to Parkinson's or arthritis). Therefore, we agree with the requirements for CAs by Weber and Ludwig [55] who also suggested making the input type selectable before interacting with a CA. In contrast to the study conducted by Ryu et al. [33], our participants preferred free-text entries over quick-reply buttons. However, using buttons strategically also helped older adults navigate the chatbot. Regarding the chatbot's output, the feedback was unambiguous; all participants preferred text to other media as they wanted the ability to re-read previous messages. Interestingly, other CAs often used speech as the sole output [20, 24].

9.4 The chatbot should use a friendly tone, but not act too human

In the evaluation of the second version of the prototype, participants' first impressions were heavily influenced by the chatbot's profile picture. While the appearance of CAs has already been studied for embodied CAs [59], we did not find a study where this effect was described for chatbots running on a messenger platform. In creating the character of *Fridolin*, we found that the inclusion of a simple avatar was a simple way to shape users' perceptions of the chatbot without spending too much time on developing unique speech patterns. Although the users said the chatbot should be friendly, it should not pretend to be a human friend as the older adults found it inappropriate. Most older adults did not request small talk in a chatbot in the design workshops, which turned out to be (mostly) true. Participants also thought that the conversation with the chatbot felt more 'real' when the chatbot touched upon their previous statements, either by explicitly confirming the information or reacting generically (e.g., 'Okay', 'Got it'). Especially considering that only 4 of the 15 participants were living alone, the results of Pradhan et al. [22] provide some context, as they proposed, that older adults who feel alone may be more inclined to personify CAs to satisfy the need for social interaction. Hence, the reason why our participants were less likely to personalize the CA could be because they did not feel that lonely. More attention needs to be paid to this relationship in future studies.

Moreover, older adults especially valued simple language and avoiding foreign words. Based on the success of the joke feature, the gamification of nutrition information seems especially promising (e.g., true or false facts, a quiz in the style of 'Who wants to be a millionaire'). In the past, there have already been some first attempts to teach healthy diets and food waste management through a playful chatbot-based approach [56] – however, as far as we know, only for children and not for older adults.

9.5 Reflecting on the participatory design approach

In the first workshop, participants also interacted with a rudimentary chatbot. Letting participants directly interact with a functional chatbot provided the most valuable insights in the design phase. Not only could we gain a first understanding of how older adults interact with a chatbot, but it also allowed participants to form an opinion about the technology. Other researchers (e.g., Wiratunga et al. [21]) used the Wizard of Oz methodology to expose participants to chatbots. While this method can be effective for eliciting ideas for features, we believe that using a basic functional chatbot has several advantages. Firstly, one can observe what challenges participants face with the deployment platform, answering questions such as, "How well can older adults navigate the interface?", "What common mistakes occur when they use the interface?" and "How do they typically enter input?". We observed that participants phrased their responses differently when they replied in another medium (analog vs. digital). To test the first contact with the chatbot, we developed a paper prototype where participants responded by writing on index cards. Compared to their answers in the first prototype, participants wrote much longer messages in the paper prototype. Thus, using the final deployment platform (Telegram) elicited more realistic responses from participants. When developing ideas for a nutrition chatbot, many older adults were influenced by their previous experiences with nutrition applications. Although these might serve as a starting point for

discussions, suggestions based on a conventional application cannot easily be transferred to CAs. While we already excluded inhabitants of a care home, the categorization of "older adults living at home" was even too broad. Based on our experiences, this group can be further divided according to their nutrition knowledge, nutrition goals, and cooking ability, which should be considered respectively when designing future nutrition chatbots. In general, we found intensive usage behavior in the first five to ten days for most participants in both studies (caused by a novelty effect [17]), which flattened out significantly in the following days and increased again significantly in the last one to two days before the evaluation interview (caused by a Hawthorne effect [41, 42]). We therefore recommend evaluation periods of at least three to four weeks, as usage patterns varied widely, differing significantly at the beginning, middle, and end of the study period.

10 Limitations

As our target group was especially at risk due to the start of the COVID-19 pandemic, our study design needed to be adjusted halfway through the evaluation of the first prototype. To continue, we decided to replace all in-person meetings with digital equivalents. Therefore, we were unfortunately unable to lead the planned group discussion and had to schedule one-on-one phone interviews instead, which in some cases turned out short and contained rather superficial information. Other participants were unable to find the time for a phone call. In the second evaluation, we faced the additional challenge of installing and setting up the chatbot remotely. The participants of the first evaluation might have been more preoccupied with COVID-19 than in the second evaluation because they were facing a completely new situation. A certain level of uncertainty was still present in December 2020, and frustration was even growing. The pandemic affected every participant in some way, but there were large individual differences. Eventually, it is hard to determine or even estimate what impact the pandemic exactly had on the user studies and their results. Furthermore, as we only interacted with a small group of participants, our results should only be generalized with caution and must be verified with a larger group of users beforehand. However, we believe that we have identified exciting areas for future research such as integrating a wider variety of stakeholders into participatory design research. For instance, the rather marginal role of nutritionists in our case could be extended in the future. In addition, long-term studies could further investigate the effectiveness of nutrition chatbots.

11 Conclusions

We investigated how a nutrition chatbot for older adults should be designed to motivate users to interact regularly. To achieve this goal, we developed a Telegram chatbot called *Fridolin* using participatory design methods. Older adults and nutritionists were involved in the design and evaluation of two iterative versions of the chatbot. This allowed us to show how older adults interacted with a nutrition chatbot over an extended period of time. Overall, we showed that older adults were able to interact well with our messenger-based chatbot. During our two long-term studies of four and seven weeks, we saw that the engagement with the chatbot was much higher in the first five to ten days, after which use tapered off, only to spike again in the last day or two prior to the final evaluation interviews. While there was no clear preference for verbal or written input, text output (rather than voice output) was preferred by the vast majority of our participants. We did not observe any change in the type of input and output modalities over the course of the studies.

Based on our findings, we identified four important design spaces regarding the design of (nutrition) chatbots for older adults. For example, in terms of user *control and proactivity*, we recommend some proactivity on the part of CAs to keep older adults engaged. Regarding *expectation management & onboarding*, we believe it is important not to raise unrealistic expectations for actual use through tutorials and to design a smooth transition from the onboarding phase to everyday use. Regarding *input and output modalities*, we recommend to make the type of input and output selectable at best, but at least provide text input and output, and specifically use buttons in addition to free text input to facilitate navigation between functions. In the area of *personification/anthropomorphization*, a certain degree of 'humanlikeness' (in the use of jokes, playful approaches, or confirmation of user input) is advisable, but not an excessive ability to engage in small talk. Using simple language and avoiding foreign words is also recommended. Thus our study contributes to the design and use of CAs and virtual coaches for healthy aging and well-being along with a strong focus on nutrition. Despite the rather small number of participants, we believe that the design implications could be similarly considered for other domains of CAs with older adults, although this remains to be examined.

Furthermore, we have shown how a participatory design approach can be implemented to design, develop and evaluate nutrition chatbots. As highlights, we would like to reiterate the importance of long-term studies (with a duration of at least three to four weeks) to observe the

technology not only under the influence of temporary, short-term effects (such as novelty or Hawthorne effects), but also in everyday life. In particular, the approach of presenting potential technological interfaces and technical prototypes relatively early in the co-design process has proven suitable for generating realistic requirements and features based on realistic usage patterns. Although we included two nutrition experts in the co-design at the beginning of our design case study (one of whom provided regular guidance and ongoing feedback throughout the co-design process), we greatly appreciated their feedback and would suggest involving additional stakeholders or experts (e.g., medical experts, culinary experts, health coaches) in the co-design process.

Overall, with our study we have contributed to the field of Human–Computer Interaction by showing what the long-term use of a nutrition chatbot for older adults might look like, what general design implications arise for the development of such chatbots, and how a participatory design approach can be effectively realized for the design, evaluation and development of nutrition chatbots.

Acknowledgments: We would like to thank all participants who took part in our studies.
Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: The research and study is part of the e-VITA project and has received funding by the "Strategic Information and Communications R&D Promotion Programme (SCOPE)" of Ministry of Internal Affairs and Communications in Japan (JPJ000595), and the European Union H2020 Programme (grant agreement 101016453).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. World Health Organization. 2018. World health statistics 2018: monitoring health for the SDGs, sustainable development goals.
2. World Health Organization. 2002. Active ageing: a policy framework.
3. Ahmed T., Haboubi N. Assessment and management of nutrition in older people and its importance to health. *Clin. Interv. Aging* 2010, *5*, 207—216;
4. Shpata V., Prendushi X., Kreka M., Kola I., Kurti F., Ohri I. Malnutrition at the time of surgery affects negatively the clinical outcome of critically ill patients with gastrointestinal cancer. *Med. Arch.* 2014, *68*, 263.
5. Bomfim M. C. C., Kirkpatrick S. I., Nacke L. E., Wallace J. R. Food literacy while shopping: motivating informed food purchasing behaviour with a situated gameful app. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1—13.
6. Ren P. P., Qian Z. C., Sohn J. J. Learn to cook for yourself: employing gamification in a recipe app design to promote a healthy living experience to young generation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer International Publishing, 2020, pp. 458—470.
7. Chung C.-F., Elena A., Schroeder J., Mishra S., Fogarty J., Munson S. A. When personal tracking becomes social: examining the use of Instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1674—1687.
8. Hassenzahl M., Laschke M. Pleasurable troublemakers. In *The Gameful World: Approaches, Issues, Applications*; Walz S. P., Deterding S., Eds. The MIT Press, 2015, pp. 167—195.
9. Khot R. A., Aggarwal D., Yi J.-Y., Prohasky D. Guardian of the Snacks : toward designing a companion for mindful snacking. *Multimodality & Society* 2021, *1*, 153—173.
10. Khot R. A., Yi J.-Y., Aggarwal D. SWAN: designing a companion spoon for mindful eating. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2020, pp. 743—756.
11. Casas J., Mugellini E., Khaled O. A. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1676—1680.
12. Graf B., Krüger M., Müller F., Ruhland A., Zech A. Nombot — simplify food tracking. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, 2015, pp. 360—363.
13. Maher C. A., Davis C. R., Curtis R. G., Short C. E., Murphy K. J. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. *JMIR mHealth and uHealth* 2020, *8*, e17558. https://doi.org/10.2196/17558.
14. El Kamali M., Angelini L., Caon M., Carrino F., Rocke C., Guye S., Rizzo G., Mastropietro A., Martin S., Elayan S., Kniestedt I., Ziylan C., Lettieri E., Khaled O. A., Mugellini E. Virtual coaches for older adults' wellbeing: a systematic review. *IEEE Access* 2020, *8*, 101884—101902.
15. Kocaballi A. B., Berkovsky S., Quiroz J. C., Laranjo L., Tong H. L., Dana R., Briatore A., Coiera E. The personalization of conversational agents in health care: systematic review. *J. Med. Internet Res.* 2019, *21*, 11.
16. Dale R. The return of the chatbots. *Nat. Lang. Eng.* 2016, *22*, 811—817.
17. Mirnig A. G., Gärtner M., Meschtscherjakov A., Tscheligi M. Blinded by novelty: a reflection on participant curiosity and novelty in automated vehicle studies based on experiences. In *Proceedings of the Conference on Mensch und Computer*; ACM: New York, NY, USA, 2020; pp. 373—381.
18. Tsay C. H.-H., Kofinas A. K., Trivedi S. K., Yang Y. Overcoming the novelty effect in online gamified learning systems: an empirical evaluation of student engagement and performance. *J. Comput. Assist. Learn.* 2020, *36*, 128—146.
19. Wulf V., Rohde M., Pipek V., Stevens G. Engaging with practices: design case studies as a research framework in CSCW. In

Proceedings of the ACM 2011 conference on Computer supported cooperative work — CSCW '11, 2011, pp. 505—512.
20. 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. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inf. Assoc.* 2018, *25*, 1248—1258.
21. Wiratunga N., Cooper K., Wijekoon A., Palihawadana C., Mendham V., Reiter E., Martin K. FitChat: conversational artificial intelligence interventions for encouraging physical activity in older adults, 2020. Available at: http://arxiv.org/abs/2004.14067.
22. Pradhan A., Findlater L., Lazar A. "Phantom friend" or "just a box with information": personification and ontological categorization of smart speaker-based voice assistants by older adults. In *Proceedings of the ACM on Human-Computer Interaction, CSCW*, vol. *3*, 2019, pp. 1—21.
23. Gudala M., Ross M. E. T., Mogalla S., Lyons M., Ramaswamy P., Roberts K. Benefits of, barriers to, and needs for an artificial intelligence—powered medication information voice chatbot for older adults: interview study with geriatrics experts. *JMIR Aging* 2022, *5*, e32169. https://doi.org/10.2196/32169.
24. El Kamali M., Angelini L., Caon M., Andreoni G., Dulake N., Paul C., Khaled O. A., Mugellini E. Building trust and companionship in e-coaching through embodiment. In *Research for Development*; Springer International Publishing, 2021, pp. 195—204.
25. Razavi S. Z., Schubert L. K., van Orden K., Ali M. R., Kane B., Hoque E. Discourse behavior of older adults interacting with a dialogue agent competent in multiple topics. *ACM Trans. Interact. Intell. Syst.* 2022, *12*, 1—21;
26. Seiderer A., Ritschel H., André E. Development of a privacy-by-design speech assistant providing nutrient information for German seniors. In *ACM International Conference Proceeding Series* 2020, pp. 114—119.
27. Chamberlain P., Craig C., Dulake N. Found in translation: innovative methods of co-design in the development of digital systems for promoting healthy aging. In *Digital Health Technology for Better Aging. Research for Development*; Andreoni G., Mambretti C., Eds.; Springer: Cham, 2021; pp. 29—52.
28. Angelini L., El Kamali M., Mugellini E., Khaled O. A., Röcke C., Porcelli S., Mastropietro A., Rizzo G., Boqué N., del Bas J. M., Palumbo F., Girolami M., Crivello A., Ziylan C., Subías-Beltrán P., Orte S., Standoli C. E., Maldonado L. F., Caon M., Martin S., Elayan S., Guye S., Andreoni G. The NESTORE e-coach: designing a multi-domain pathway to well-being in older age. *Technologies* 2022, *10*, 50.
29. Martin-Hammond A., Vemireddy S., Rao K. Exploring older adults' beliefs about the use of intelligent assistants for consumer health information management: a participatory design study. *JMIR Aging* 2019, *2*, e15381. https://doi.org/10.2196/15381.
30. Yaghoubzadeh R., Kramer M., Pitsch K., Kopp S. Virtual agents as daily assistants for elderly or cognitively impaired people. In *Intelligent Virtual Agents. IVA 2013*; Aylett R., Krenn B., Pelachaud C., Shimodaira H., Eds. Springer: Berlin, Heidelberg, 2013.
31. Sayago S., Neves B. B., Cowan B. R. Voice assistants and older people: some open issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces — CUI '19*, 2019, pp. 1—3.
32. Ring L., Barry B., Totzke K., Bickmore T. Addressing loneliness and isolation in older adults: proactive affective agents provide better support. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 61—66.
33. Ryu H., Kim S., Kim D., Han S., Lee K., Kang Y. Simple and steady interactions win the healthy mentality. In *Proceedings of the ACM on Human-Computer Interaction; CSCW2*; vol. *4*, 2020, pp. 1—25.
34. Maenhout L., Peuters C., Cardon G., Compernolle S., Crombez G., DeSmet A. Participatory development and pilot testing of an adolescent health promotion chatbot. *Front. Public Health* 2021, *9*, 1—16.
35. Potts C., Ennis E., Bond R. B., Mulvenna M. D., McTear M. F., Boyd K., Broderick T., Malcolm M., Kuosmanen L., Nieminen H., Vartiainen A. K., Kostenius C., Cahill B., Vakaloudis A., McConvey G., O'Neill S. Chatbots to support mental wellbeing of people living in rural areas: can user groups contribute to co-design? *J. Technol. Behav. Sci.* 2021, *6*, 652—665.
36. Russell D., Peplau L. A., Ferguson M. L. Developing a measure of loneliness. *J. Pers. Assess.* 1978, *42*, 290—294.
37. Muller M. J., Kuhn S. Participatory design. *Commun. ACM* 1993, *36*, 24—28.
38. Carros F., Schwaninger I., Preussner A., Randall D., Wieching R., Fitzpatrick G., Wulf V. Care workers making use of robots: results of a three-month study on human-robot interaction within a care home. In *CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2022; pp. 1—15.
39. Lindsay S., Jackson D., Schofield G., Olivier P. Engaging older people using participatory design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1199—1208.
40. Unbehaun D., Aal K., Vaziri D. D., Wieching R., Tolmie P., Wulf V. Facilitating collaboration and social experiences with videogames in dementia: results and implications from a participatory design case study. In *Proceedings of the ACM on Human-Computer Interaction, CSCW*; vol. *2*, 2018.
41. McCambridge J., Witton J., Elbourne D. R. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J. Clin. Epidemiol.* 2014, *67*, 267—277.
42. Parsons H. M. What happened at Hawthorne? *Science* 1974, *183*, 922—932;
43. Isomursu M., Ervasti M., Kinnula M., Isomursu P. Understanding human values in adopting new technology-a case study and methodological discussion. *Int. J. Hum. Comput. Stud.* 2011, *69*, 183—200.
44. Kim S., Paulos E., Mankoff J. inAir: a longitudinal study of indoor air quality measurements and visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2745—2754.
45. Winkle K., Caleb-Solly P., Turton A., Bremner P. Mutual shaping in the design of socially assistive robots: a case study on social robots for therapy. *Int. J. Soc. Robot.* 2020, *12*, 847—866;
46. Wagner I. Critical reflections on participation in design. In *Socio-Informatics: A Practice-Based Perspective on the Design and Use of IT Artifacts*; Wulf V., Pipek V., Randall D., Rohde M., Schmidt K., Stevens G., Eds. Oxford University Press, 2018, pp. 243—278.
47. Snyder C. *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2003.
48. Alhojailan M. I. Thematic analysis : a critical review of its process and evaluation. *W. East J. Soc. Sci.* 2012, *1*, 39—47.

49. Zhou L., Gao J., Li D., Shum H.-Y. The design and implementation of XiaoIce, an empathetic social chatbot. *Comput. Ling.* 2020, *46*, 53—93.
50. Stephens T. N., Joerin A., Rauws M., Werk L. N. Feasibility of pediatric obesity and prediabetes treatment support through tess, the AI behavioral coaching chatbot. *Transl. Behav. Med.* 2019, *9*, 440—447.
51. Jain M., Kumar P., Kota R., Patel S. N. Evaluating and informing the design of chatbots. In *DIS 2018 — Proceedings of the 2018 Designing Interactive Systems Conference*; ACM: New York, NY, USA, 2018; pp. 895—906.
52. Maroengsit W., Piyakulpinyo T., Phonyiam K., Pongnumkul S., Chaovalit P., Theeramunkong T. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology — ICIET 2019*, 2019, pp. 111—119.
53. Moore R. J., Arar R., Ren G.-J., Szymanski M. H. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*; Morgan & Claypool: New York, NY, USA, 2019.
54. Trajkova M., Martin-Hammond A. "Alexa is a toy": exploring older adults' reasons for using, limiting, and abandoning echo. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1—13.
55. Weber P., Ludwig T. (Non-)Interacting with conversational agents: perceptions and motivations of using chatbots and voice assistants. In *Proceedings of Mensch und Computer (Magdeburg 2020)*, 2020, pp. 321—331.
56. Fadhil A., Villafiorita A. An adaptive learning with gamification & conversational UIs: the rise of CiboPoliBot. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization — UMAP '17*; ACM: New York, NY, USA, 2017; pp. 408—412.
57. Alnefaie A., Singh S., Kocaballi B., Prasad M. An overview of conversational agent: applications, challenges and future directions. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies*, 2021, pp. 388—396.
58. Khan R., Das A. *Build Better Chatbots*; Apress: Berkeley, CA, 2018.
59. ter Stal S., Broekhuis M., van Velsen L., Hermens H., Tabak M. Embodied conversational agent appearance for health assessment of older adults: explorative study. *JMIR Human Factors* 2020, *7*, e19987.