The Accuracy of an Audio Interface Designed for Value Elicitation

Eliciting Personal Values from the Users to Build Responsible AI

Elvira Voorneveld[1]

Supervisor(s): Catholijn Jonker[1]**, Pei-Yu Chen**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Elvira Voorneveld
Final project course: CSE3000 Research Project
Thesis committee: Catholijn Jonker, Pei-Yu Chen, Stephanie Wehner

Abstract

Behavior support applications aim to provide personalized and flexible support to users in various domains. To achieve this, understanding users' preferences, values, and context is crucial. Creating user models that incorporate users' norms and values has been proposed as a solution to capture the relationship between desired behaviors and values. However, updating and modifying user models at run-time remains a challenge, as users' norms and values may change over time. This study investigates the accuracy of an audio interface designed to elicit values-related information using isolated questions. This involves designing an audio interface and evaluating its effectiveness through participant interactions, where they are presented with four scenarios. It was found that the audio interface performs above average in terms of usability, as indicated by the System Usability Scale score. The accuracy of the user models is evaluated through the Hamming distance and value differences between the base model and the participant-improved model. Most models required a small number of changes, and when changes were made, they were generally minimal. Additionally, feedback collected through open-ended interview questions lays down a basis for further development. The study contributes to the field by demonstrating the efficacy of the audio interface and its potential for updating user models in real-time. Overall, the research findings support the development of more effective and personalized behavior support applications that can adapt to its users.

1 Introduction

Behaviour support applications have become increasingly prevalent in various domains, such as healthcare, education, and productivity, providing support to its users in a flexible and personalised way [van Riemsdijk et al., 2015]. To be effective, this requires understanding the user's preferences, values, and context. The challenge is to capture these aspects explicitly in the agent's decision-making process when giving behavioural advice to its users, especially in unanticipated situations.

One solution to this challenge is to create user models that integrate users' norms and values (e.g. [Tielman et al., 2018], [Kließ et al., 2019], and [Cranefield et al., 2017]). User models capture the relationship between users' desired behaviours and their values, enabling the support agent to make its reasoning explicit and improve transparency and explainability. Several researchers have explored the creation of user models in the context of behaviour support applications (e.g. [Berka et al., 2022] and [Honka et al., 2022]).

However, updating and modifying the user models at runtime is still a challenge, as users' norms and values may change over time, so it is important to be able to elicit necessary information from them. Several researchers have explored different approaches to doing so (e.g. [Pasotti et al., 2016]), but there is still a knowledge gap in understanding the accuracy of an interface that can elicit values-related information in real-time and incorporate it into the user model.

In that context, this study investigates the accuracy of an audio interface that enables the agent to elicit values-related information from the users that can be used to update the user model in real-time. The objective is to contribute to the development of more effective and personalised behaviour support applications that can adapt to users' changing norms and values.

The study's approach involves designing an audio interface and evaluating its effectiveness. Participants will be asked to interact with the behaviour support agent through the audio interface and answer its questions. Data is collected on the participants' values and preferences, through their answers to the agent's questions. The study's results will provide insights on the accuracy of an audio interface that was created to elicit values-related information from the users for updating the user model in real-time.

In summary, the current study aims to address the gap in knowledge by investigating the efficacy of an interface that can effectively elicit values-related information from users and incorporate it into the user model. By doing so, the study will contribute to the development of more effective and personalised behaviour support applications that can adapt to users' changing norms and values in real-time.

The remainder of this paper is structured as follows. Section 2 describes the methodology, after which section 3 details the experiment setup and results. Section 4 reflects on responsible research and in section 5 the results are discussed. Finally, section 6 provides the conclusions and recommendations for future research.

2 Methodology

This study was conducted to investigate how a conversational agent can successfully elicit values-related information using an audio interface and use that to update the underlying user model in real-time. Participants engaged in conversations with the agent through the audio interface, where they were presented with multiple scenarios and in turn provided insights into their values and the influence of context. The interface aimed to minimize misunderstandings and resulting misalignments, but considering the recency of research in this field, the extent of its success could not be ascertained ahead of time. Therefore, both qualitative measures and data about the usability and final user model were gathered. The experiment received approval from the Ethics Committee of the Delft University of Technology, and participants provided informed consent.

2.1 Participants

Fifteen technologically literate individuals aged between 18 and 65, diverse in gender, took part in the study. None of the participants had any hearing impairments.

2.2 Measures

Various measures were employed to assess the system's usability and the accuracy of the resulting user model. Following the elicitation experiment, participants were presented

with the resulting user models, provided an explanation on interpretation, and given the opportunity to make improvements as they saw fit (further explained in Chapter 2.4). A baseline user model generated by the agent was compared to the participant-provided improved version. Additionally, participants were asked open-ended interview questions to provide feedback on the system's usability. To gauge general usability, the System Usability Scale (SUS) [Brooke, 1995] was employed, known for its reliability even with small sample sizes. The SUS results were compared with the baseline established by previous systems that utilized the SUS.

2.3 Procedure

The study lasted approximately one hour per participant, with the experiment itself taking up half of that time. Participants were welcomed and given concise instructions. They read and understood the consent form, and any queries were addressed by the experimenter. Consent was obtained through a name, date, and signature.

The first part of the session involved interaction between participants and the conversational agent via the audio interface. The experiment employed a Wizard of Oz setup, with the experimenter acting as a perfect speech-to-text system, transcribing the participant's responses to the conversational agent. This setup aimed to circumvent issues related to speech-to-text conversion, as it was not the focus of this study.

Participants were presented with four distinct scenarios, all concerning choices about health. The first asks about water and the participant's choice of an unhealthy beverage, then adds the context of a party. The second is about relaxing at home or doing an outdoor exercise of the participant's choice and asks what changes when doing either activity in bad weather. The third scenario concerns improving one's diet, choosing from eating nutritious foods or processed foods in the context of eating at a restaurant with friends. The final scenario chooses between sleeping early and staying up late, with the context of a work deadline. These scenarios are further detailed in Appendix A. Participants answered a series of questions concerning the values associated with each scenario. The questions were asked in an isolated manner, without any comparative element between the presented values and choices.

For each scenario, two sets of questions were asked. The first set aimed to establish a general model of participants' values, enabling the agent to understand their general priorities in the absence of specific context. The second set of questions aimed to identify any differences between the general ranking of values and the ranking assigned when contextual information was provided.

Four of the participants also did the experiments of four other, similar interfaces designed for value elicitation ([Kastelein, 2023], [Krupskis, 2023], [Mendez, 2023], and [Vizuroiu, 2023]). The other interfaces include graphical and textual ones, with two types of questioning. Two others ask questions in an isolated manner, like done in this study (further explained in Chapter 2.5), and the other two involve comparing values and choices. The setup of the other experiments is otherwise identical, so the results can be directly compared for further analysis.

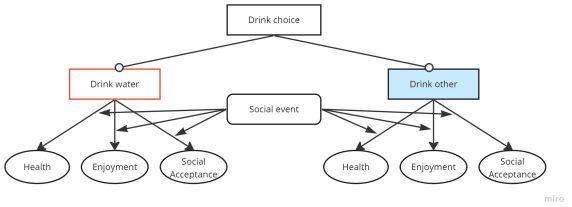


Figure 1: A behaviour tree that represents the template (weightless) user model for the first scenario. Here, *Drink choice* represents the choice in the scenario, *Drink water* and *Drink other* are the user behaviour options, the leaves of tree (e.g. *Health* and *Enjoyment*) are the values and *Social event* is the context.

In the second part of the session, participants completed the SUS and answered additional feedback questions. Based on participants' answers during the first part of the session, behavior trees representing their user models were constructed for each scenario. The experimenter explained the behavior trees to participants, allowing them to make changes to more accurately reflect their values within each scenario.

2.4 User Model

The user model (see Figure 1 for an example) comprises four components: the root of the behaviour tree represents the choice available in the scenario, followed by actions representing user behaviour options. Each option has a number of values (the leaves of the tree), which could be influenced by choosing that action. Additionally, a separate node represents additional context to the scenario, which can alter the relationship between an option and a value. A behaviour tree was selected to model the user due to its intuitiveness and established use for this purpose (e.g. [Tielman et al., 2022]). This aimed to ensure participants could understand and potentially modify the model to more accurately represent themselves, emphasizing the need for ease of interpretation.

The basic structure of the tree was the same for each participant, the only difference being the weights on the edges that relate to the values. Participants were allowed to answer similarly to a 5 point Likert scale, with the following range of answers: very positive (+10), slightly positive (+5), neutral (0), slightly negative (-5), and very negative (-10). The numbers in brackets represent the weights that were added to the edges of the tree.

Weights on the edges originating from the context are calculated by taking the difference between the answer given with and without contextual information. For example, if a participant answered that drinking water has a slightly positive effect on Enjoyment normally, but a very negative effect when at a social event, the edge leading from the context to the edge of the Enjoyment value would be -15. Adding this weight to the original +5 (corresponding to slightly positive) results in the -10 (corresponding to very negative) like they answered in context. This way it is easy to see the influence of the context on the base tree.

When improving the user model, participants were allowed to change any weight, so long as it stayed within the range of

Agent	Great! Let's start. Can you tell me an unhealthy or sugary drink that is available to you, that you enjoy drinking? Examples include beer, cola, and juice.
User	DRINK_CHOICE
Agent	Did you say DRINK_CHOICE?
User	[Continue if 'yes' or ask again if 'no']
Agent	Now, imagine the following scenario. You have decided to drink more water and have been doing so every evening in the past week. The alternative to drinking water is to drink DRINK_CHOICE. If you ever need a reminder of this situation, feel free to ask me to repeat it. Shall we continue?

Agent	Now, let's talk about DRINK_CHOICE. How healthy is it to drink DRINK_CHOICE?
User	DRINK_HEALTH
Agent	How enjoyable is drinking DRINK_CHOICE?
User	DRINK_ENJOY

Figure 2: Part of the dialogue template of the first scenario. Here, *DRINK_CHOICE* is a variable that gets assigned a value when the participant answers, which is then used in subsequent descriptions and questions. Variables like *DRINK_HEALTH* and *DRINK_ENJOY* would contain the essential data.

-10 to +10. The exception being the edges originating from the context, which could be anywhere from -20 to +20, as long as the resulting edge towards the value stayed within normal range. For example, if drinking water has a very positive effect on social acceptance normally, but a very negative effect in context, the edge leading from the choice to the value would be +10 and the edge originating from the context towards that edge would be -20, resulting in a -10.

2.5 Dialogue Structure

In this study, questions were asked in an isolated manner as mentioned in Chapter 2.3. This means that for each behaviour option, the participants were asked how that option influences each value. So for a user model that has two options and three values per option, it would require six questions to establish the base model without context. Using the example in Figure 1, we would ask what the health, enjoyment, and social acceptance of drinking water is, and repeat these questions for the other option. After introducing the context, the same questions would be repeated, resulting in twelve questions for a user model with two options and three values per option.

For each scenario, the conversational agent requested input from participants, where possible, to personalize and streamline interactions. Concrete examples aided participants in envisioning specific scenarios more easily. For instance, prior to introducing the first scenario, participants were asked about their preferred unhealthy or sugary drink. This input was incorporated into the scenario description and subsequent questions, as shown in Figure 2 with the example of *DRINK_CHOICE*. While this input was not used in data processing and analysis, it enhanced usability for participants.

Variables such as *DRINK_HEALTH* and *DRINK_ENJOY* in Figure 2 contained the desired data. These variables were also used to provide participants with summaries of their answers, confirm the agent's understanding of their responses, and offer the opportunity to correct any inaccuracies. Additionally, these summaries assisted participants in maintaining an overview of their answers thus far.

Agent	Can you tell me a social event you enjoy going to? Examples include a birthday party, a movie night, and a wedding.
User	EVENT_CHOICE
Agent	Did you say EVENT_CHOICE?
User	[Continue if 'yes' or ask again if 'no']
Agent	Now, let's consider an addition to the previous scenario. There is EVENT_CHOICE coming up that you will attend. At EVENT_CHOICE both water and DRINK_CHOICE are available, but you have chosen to drink DRINK_CHOICE. Shall we continue?

Agent	How healthy is it to drink DRINK_CHOICE at EVENT_CHOICE?
User	EVENT_D_HEALTH
Agent	How enjoyable is drinking DRINK_CHOICE at EVENT_CHOICE?
User	EVENT_D_ENJOY

Figure 3: Part of the dialogue template of the context for the first scenario. Here, *EVENT_CHOICE* is a variable that gets assigned a value when the participant answers, which is then used in subsequent descriptions and questions. Variables like *EVENT_D_HEALTH* and *EVENT_D_ENJOY* would contain the essential data.

Following the introduction of the general scenario, the conversational agent asked participants questions to construct the base model. Since this research focused on updating a model with additional context, a base model was necessary.

Once the base tree was constructed, the agent introduced additional context that might alter the effect of a particular choice on the associated values. The same questions were posed as in the scenario, but with the incorporation of context, as depicted in Figure 3.

3 Analysis and Results

As described earlier, there were three measures employed to assess the efficacy of this method of value elicitation, the results to which are detailed below.

3.1 System Usability Scale

To analyse the results of the SUS, we decided to calculate both the final score as described by Brooke, as well as the scores for each individual item on the scale. SUS is a 5 point Likert scale, ranging from strongly disagree (1) to strongly agree (5). So for each of the ten items, a number between 1 and 5 was obtained. First we calculated the score contributions from each item according to the SUS instructions. Take the average score contributions from each item and multiply those by 25 to obtain the average SU value for each item. To calculate the overall SU value, take the average of the SU value for each item.

The overall SU value for this audio interface is 76.7. The average SUS score worldwide is 68, which means that this interface performs above average. To further interpret the score, it can be said to fall in between the good and excellent ratings as per the adjective rankings of SUS scores [Brooke, 2009].

The average SU value for each item can be seen in Figure 4, where the lowest average score of 43 was given to item 1 (questions if one would frequently use this system) and the

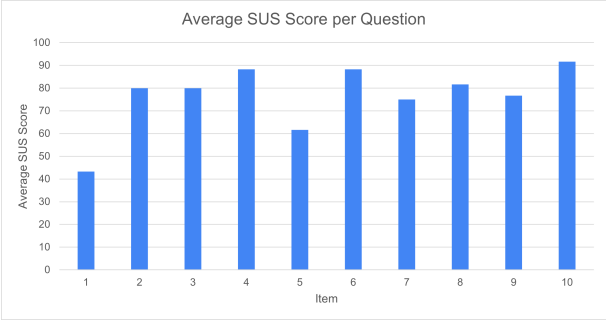
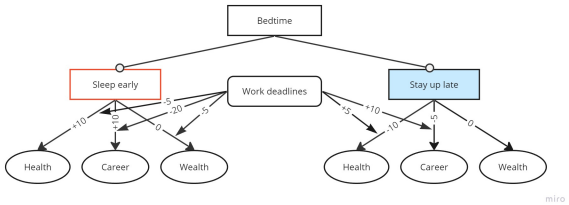
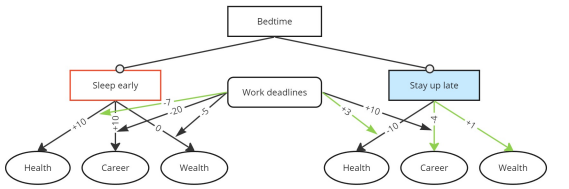


Figure 4: Results of the SUS score for each of the items.



(a)



(b)

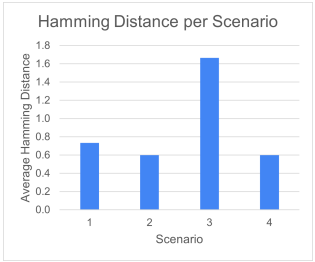
Figure 5: A participant's user model as generated from their answers and the improved model after letting the participant make changes, respectively. The coloured lines indicate where changes were made. The Hamming distance is 4 and the value difference is 6.

highest average score of 92 goes to item 10 (concerns the need to learn a lot before using the system). The only other item that performed below average was number 5 (whether the functions in this system were well integrated) with a score of 62. Appendix B shows the ten items of the SUS as used in this study.

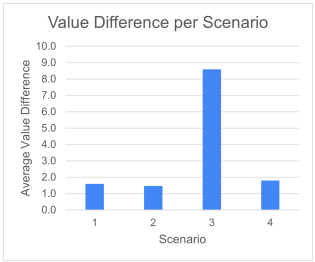
3.2 Accuracy Measure

The accuracy is calculated by taking the difference between the constructed user model and the one improved by the participants. Two measures were used to do so, the first is the Hamming distance between the trees and the second is the difference in values. Figure 5a shows an example of a model that was created from a participant's answers and Figure 5b shows their improved version. Here, the Hamming distance is equal to the amount of edges that were changed, and the value difference is equal to the sum of the absolute value of the changes per edge.

It can be seen from Figure 6a that on average for most scenarios, less than one edge was changed. From this it can



(a)



(b)

Figure 6: Average Hamming distance and average value difference of the behaviour trees per scenario, respectively.

be deduced that participants often thought those trees were mostly accurate. The only exception being scenario 3, with 1.7 changes on average.

Figure 6b shows the average value difference between the original tree and the improved version. Much like the Hamming distance, most scenarios have a small average difference in value of under 2. However, for scenario 3, there is an average difference of 8.6.

Dividing the value difference by the Hamming distance results in an average amount of change made per edge. For scenarios 1, 2, and 4 this is roughly 3.5, but considerably higher for scenario 3 with an average of 5.9. Not only does the latter scenario have most overall changes, they are also larger than in any of the other scenarios. The total average changes per edge is equal to 3.7. Further discussion and explanation of these results can be found in Chapter 5.2.

For both measures, we can also calculate the mean, median, and standard deviation to provide more insight into the spread of the data. The mean of the Hamming distance is 3.6 with a median of 1 and a standard deviation of 6.2. The value difference has a mean of 13.5 with a median of 5 and a standard deviation of 15.6. Since the mean is greater than the median in both cases, the data is skewed to the right. This does not necessarily indicate the presence of an outlier, but as can be seen from the experiment data in Appendix C, there is one outlier (participant 10) present, mostly with regards to the Hamming distance. Removing this entry would result in a mean of 2.1 and a standard deviation of 2.5 for the Hamming distance, and a mean of 11.2 and a standard deviation of 13.5 for the value difference.

As mentioned in Chapter 2.3, there were four other interfaces tested in tandem with the audio interface detailed here. The results of all experiments are presented in Table 1. It can be seen that graphical in isolation (GI) and textual in comparison (TC) perform best, with both the lowest overall Hamming distance and value difference. TC does have the highest value difference per change, with this interface (AI) having the lowest.

3.3 Interview

Simple, open-ended questions were asked after the participants completed the experiment, filled in the SUS, and improved their respective user models. These questions concerned the reasons they chose certain weights and assump-

	Hamming dist.	Value diff.	Value diff. per change
AI	3.6	13.5	3.8
GC	5.3	36.9	6.9
GI	1.3	8.0	6.2
TC	0.8	9.7	12.1
TI	5.1	30.9	6.1

Table 1: Hamming distance, value difference, and value difference per changed edge for each of the interface variants: **A**udio in **I**solation, **G**raphical in **C**omparison and **I**solation, and **T**extual in **C**omparison and **I**solation.

tions they made in order to do so, their likes and dislikes of the system, and any improvements they thought could be made to the system to improve it. Since these questions have no preconstructed answers to choose from, here we will summarise them.

Many participants complimented the questions, scenarios and contexts for being easy to understand and getting introduced in a logical manner. While it did not feel completely like a conversation, this was not necessarily considered a bad thing since they acknowledged that the nature of this system requires it to present information this way to keep it streamlined.

A number of participants mentioned that this type of system could benefit more from a graphical or textual interface. The reasons for this being that it is quicker to read, it is easier to see the possible answers rather than needing the system to read them aloud, and it is more comfortable to take things at your own pace. In addition, it would be appreciated if they could have the scenario in front of them while thinking about answering the question.

On the other hand, a few answered that they actually appreciated this audio format. This is because for people with reading disabilities such as dyslexia, this type of interface makes it faster and easier. In addition, it was mentioned that this conversational audio style is more natural and comfortable for older individuals that are not as used to technology. Finally, some participants appreciate being able to simply ask the system for more information where necessary, rather than having to navigate through an app to get the same result.

Over half of the participants indicated that while there wasn't any confusion on how to proceed or what to answer, the scenarios and contexts were very broad and could be interpreted in many ways. This required that the participants make assumptions and keep those consistent throughout the experiment, causing delay with answering at times. Some participants also attempted to ask clarifying questions to the interface or experimenter.

The scenarios used in this experiment were rather simplistic, resulting in simple user models. However, when trying to encompass more nuanced situations and contexts as drawn from the real world, this will also become more difficult to model. As one individual suggested, the model might want to add a modifier to show the importance of some values relative to others. In addition to showing the impact of a choice on a value (using weights), this would add a modifier to determine how important a user finds a value.

Additionally, it was commented that some of the values do not seem to make much sense for a specific scenario, highlighting the importance of choosing the correct values.

4 Responsible Research

In conducting this study, several considerations were taken into account to ensure responsible research practices. The following aspects were addressed: data misconduct, reproducibility, data bias in data analysis, and potential risks associated with the research. Appropriate mitigations were implemented to minimize these risks.

4.1 Data Misconduct

To mitigate data misconduct, ethical guidelines were followed throughout the study. The research received approval from the Ethics Committee of the Delft University of Technology, ensuring that the experiment adhered to ethical standards. Participants were provided with informed consent forms, clearly explaining the purpose of the study and their rights as participants. Any queries or concerns from participants were promptly addressed by the experimenter.

To prevent misconduct, steps were taken to maintain the integrity of the data collection and analysis process. The experimenter received appropriate lecturing on research ethics, confidentiality, and responsible data handling. Data collection procedures were carefully documented to ensure consistency and transparency.

To minimize the risk of bias or manipulation, the experimenter followed predetermined guidelines during any interaction with participants. This included avoiding leading questions or influencing participants' responses in any way. Any potential conflicts of interest or personal biases that could affect the data collection or analysis were identified and managed appropriately.

Measures were also taken to protect the privacy and confidentiality of participant data. All data were anonymized and securely stored to prevent unauthorized access or disclosure. Data integrity was ensured and unauthorized modifications were prevented.

By upholding ethical standards and implementing safeguards against data misconduct from both participants and researchers, the study aimed to maintain the credibility and reliability of the research findings.

4.2 Reproducibility

To promote reproducibility, the methodology of the study was documented in detail. This included a clear description of the experimental setup, participant recruitment criteria, and measures employed to assess system usability and the resulting user model. The dialogue templates and behavior trees used to capture participant responses were provided as examples to facilitate replication of the study. By transparently documenting the research process, other researchers can replicate and validate the findings, contributing to the scientific rigor of the field.

4.3 Data Bias in Data Analysis

Data bias in data analysis is a potential concern in any research involving human participants. To mitigate this risk,

diverse participants were recruited, encompassing individuals of different genders and age groups. The aim was to ensure a representative sample that reduces potential biases in the resulting user models. Additionally, measures were implemented to minimize bias during data analysis, such as using an established methodology for constructing behavior trees and allowing participants to modify the models to better reflect their values. The use of open-ended interview questions provided an opportunity to capture qualitative feedback, further mitigating potential bias.

4.4 Other Risks and Mitigations

While the study aimed to minimize misunderstandings and misalignments in the audio interface, there was inherent uncertainty due to the recency of research in this field. To address this, both qualitative measures and quantitative data were collected to assess the usability and accuracy of the system. Participant feedback, the System Usability Scale (SUS), and comparison with established baselines were used to evaluate the effectiveness of the conversational agent.

Additionally, a Wizard of Oz setup was employed to ensure accurate transcription of participant responses, reducing potential issues related to speech-to-text conversion. However, the limitations of this setup, including the reliance on the experimenter's transcription accuracy, were acknowledged and considered in the interpretation of the results.

Overall, responsible research practices were followed to mitigate potential risks and enhance the reliability and validity of the study. By addressing data misconduct, reproducibility, data bias in analysis, and other associated risks, the research aimed to contribute to the advancement of knowledge in the field of conversational agents and user modeling.

5 Discussion

The objective of this study was to investigate the accuracy of an audio interface for value elicitation and real-time user model updating in behavior support applications. By employing speech-based interaction, the aim was to provide a natural and intuitive way for users to express their values and preferences, while dynamically updating the user model to enhance personalization.

5.1 Usability

The results of this study demonstrated that the audio interface performed well above average in terms of usability, as indicated by the System Usability Scale (SUS) score. The audio interface was designed with a strong emphasis on ease-of-use and intuitiveness, considering that it would be part of a larger application. To improve usability, personalization was added to the interface, such as asking for participant input and using it in subsequent questions. Additionally, information was presented in steps rather than all at once, and examples were provided where possible.

While the SUS scores were generally positive, there were two areas in which the audio interface showed lower usability. The first concerns whether or not a user would like to use this system frequently. It is difficult to analyse why this score was low based on the SUS alone, but the interview questions give more context for this. Many participants expressed a preference for textual or graphical modalities over audio, as they were more accustomed to those formats. This experiment concerned behaviour change with regards to health and it could also be that a majority of the participants had no interest in changing that behaviour.

The second item is about the integration of the various functions in this system. As far as the participants were aware, they could answer questions from a given set of answers, they could ask the system to repeat the scenario, context, question, or options. From their perspective, it might not have seemed as though there were any functions in this system to be integrated, so it is possible they answered neutral (3) on the SUS as per its instructions. Participants frequently asked questions about this item while filling in the SUS, which further reinforces that hypothesis. Enhancing the clarity of functions and their integration could improve usability further.

5.2 Accuracy

In most cases, minimal modifications were required to the user models. Participants had the freedom to make changes in any integer increment, but some chose to stick to the 5-point increments associated with the answers provided during the experiment. This suggests that the scenarios might have been too simplistic for some participants to feel the need for smaller adjustments. Alternatively, it could indicate their satisfaction with the initial user models, as they did not make any changes despite having the opportunity.

There is a high deviation in user model accuracy for both measures, which might be because of a similar reason as mentioned in the previous paragraph. Another possibility is the differential impact of visualization on individuals. While speculative and not backed by any of the interview questions, it may be worthwhile to look into. In addition, this could be a good reason for combining audio with some kind of visuals, to create a hybrid interface that uses multiple modalities.

Among the scenarios, scenario 3 stands out with a higher number of changes and larger differences in values. Participants frequently mentioned this scenario during the interview, highlighting its vague nature and the need to keep track of additional information while answering questions. Some participants found it easier to imagine specific healthy and unhealthy foods rather than relying on the vague terms 'nutritious' and 'processed'.

Furthermore, participants tended to compare the two choices subconsciously. For example, when dining out at a particular restaurant, all food options are in the same price range, resulting in a weight of 0 in the user model. However, when seeing them in isolation they would both have a weight of -10 on the edges, because they are both expensive. They realized that they had made this comparison when looking at the resulting user model, causing them to adjust the weights accordingly.

The setup of the experiment might have worsened this effect. The first two scenarios involved asking the participant for input that would be used in descriptions and questions later, while the third and fourth scenario did not because of their perceived simplicity. Since the fourth scenario did not

have a significantly lower accuracy than the first and second, it is possible that the order of scenarios has an impact on accuracy.

Considering the results of all modalities and questioning types as presented in Table 1, we can compare the interfaces that have either the same modality or questioning type to gain more insights into the results. One graphical modality interface (GI) was highly accurate, while another (GC) was the least accurate. However, in textual interfaces, the in comparison questioning type (TC) outperformed the in isolation (TI) version.

These results directly contradict one another, so the reason for these large differences may lie elsewhere. While the scenarios and number of questions used are identical, the phrasing of said scenarios and questions was different. As was discovered from analyzing the third scenario for the interface presented here (AI), it is probable that the phrasing of a scenario and its subsequent questions significantly influence the resulting user model accuracy. Additionally, the order in which the scenarios were presented also differs per study. Finally, since the studies were conducted separately, it is likely that each experimenter presented the generated user models to the participants in a different manner. One might have presented them all at once and given them free range, while another might walk them through the models step by step. This can also influence the participants' tendency to change the given models.

6 Conclusions and Future Work

In this study, we aimed to determine the accuracy of an audio interface designed to elicit values-related information using isolated questions. Our findings provide insights into the usability and effectiveness of the audio interface in behavior support applications.

The usability assessment, measured by the System Usability Scale score, revealed that the audio interface performed above average. However, some participants expressed a preference for visual elements in addition to the audio setup. The added personalization in the interface helped mitigate the absence of visuals to some extent, potentially offering an overall advantage. It is worth noting that the audio interface required more time to elicit values compared to other interface types. Nonetheless, when changes were made, they were generally minimal and the smallest out of all interface variants, indicating the importance of nuanced scenario presentation and the benefits of personalization.

While the experiment yielded promising results, there are several areas for future research and improvement. First, further exploration is needed regarding the phrasing of scenarios and their presentation to the users. The scenarios used in this study were simplified versions of real-world events and need to become more detailed to fully encapsulate the user's preferences. This expansion necessitates the development of a more sophisticated user model capable of incorporating additional details, such as the approach described in [Cranefield et al., 2017].

Additionally, future studies should consider increasing the sample size and ensuring consistent procedures across all interface variants to enhance the reliability and generalizability of the findings. The current experiment employed a Wizard-of-Oz setup for Speech-to-Text, which can be integrated into the interface as an automated system component, thus eliminating the need for manual intervention.

Lastly, to enhance the ecological validity of the scenarios, it is crucial to establish a scientifically tested set of values that aligns with the specific contexts. This will provide a more robust foundation for value elicitation and facilitate a deeper understanding of users' preferences.

In conclusion, this study demonstrates the above-average usability of the audio interface for value elicitation in behavior support applications. The results highlight the importance of addressing user preferences, refining scenario phrasing, and further integrating automated speech recognition. By advancing these areas and conducting future research, we can unlock the full potential of audio interfaces for value elicitation and user modeling, thus contributing to the development of personalized behavior support systems.

A Scenarios

Detailed here are the four scenarios used in the study, where each scenario has its own choice, behaviour options, related values, and context.

A.1 Scenario 1

This scenario concerns the participant's choice of beverage, where their goal is to drink more water instead of an unhealthier alternative. The participant has an opinion on how each choice influences certain values, these being health, enjoyment, and social acceptance. The context in this scenario is attending a social event of the participant's choice and the corresponding user model can be seen in Figure 1. This scenario is noted down as follows:

Goal. Drink more water

Ideal choice. Drink water

Alternative. Drink beverage of the user's choice

Context. Social event of the user's choice

Values. Health, Enjoyment, Social Acceptance

A.2 Scenario 2

Goal. Exercise more

Ideal choice. Outdoor exercise of the user's choice

Alternative. Relax at home

Context. Bad weather of the user's choice

Values. Health, Enjoyment, Safety, Comfort

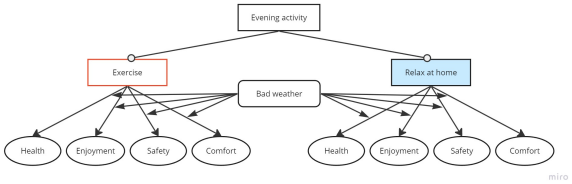


Figure 7: Template (weightless) user model for the second scenario. Here, *Evening activity* represents the choice, *Exercise* and *Relax at home* are the behaviour options, and *Bad weather* is the context.

A.3 Scenario 3

Goal. Improve diet

Ideal choice. Eat nutritious foods

Alternative. Eat processed foods

Context. Eating at a restaurant with friends

Values. Health, Enjoyment, Social Acceptance, Wealth

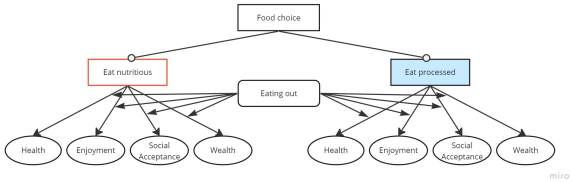


Figure 8: Template (weightless) user model for the third scenario. Here, *Food choice* represents the choice, *Eat nutritious* and *Eat processed* are the behaviour options, and *Eating out* is the context.

A.4 Scenario 4

Goal. Improve sleep schedule

Ideal choice. Sleep early

Alternative. Stay up late

Context. Work deadline coming up

Values. Health, Career, Wealth

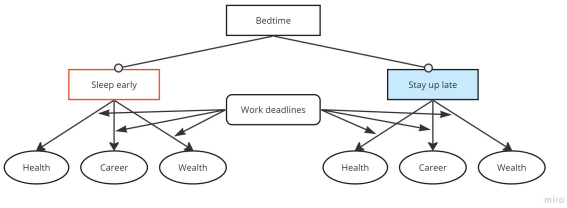


Figure 9: Template (weightless) user model for the fourth scenario. Here, *Bedtime* represents the choice, *Sleep early* and *Stay up late* are the behaviour options, and *Work deadlines* is the context.

B SUS Items

The items of the SUS [Brooke, 1995] used in this study.

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this systems were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system

C Accuracy Results

Results of the accuracy measure. Each participant had four trees (one for each scenario) and could make changes to them. The Hamming distance is the number of edges changed and the value difference is the total difference between the values of the generated and improved trees.

Participant	Scenario	Hamming Dist.	Value Diff.
1	1	1	5
	3	3	12
	4	1	5
2	1	2	4
	2	1	3
	3	3	11
	4	2	6
3	2	1	5
	3	3	21
	4	2	10
4	3	4	40
5	-	-	-
6	1	1	5
	3	1	10
7	-	-	-
8	3	2	10
9	-	-	-
10	1	7	10
	2	6	9
	3	8	20
	4	4	6
11	-	-	-
12	3	1	5
13	-	-	-
14	-	-	-
15	3	1	5

Table 2: Hamming distance and value difference calculated from the changes each participant made to their trees. Scenarios with no changes have been left out.

References

[Berka et al., 2022] Berka, J., Jonker, C. M., Mikovec, Z., van Riemsdijk, M. B., and Tielman, M. L. (2022). Misalignment in Semantic User Model Elicitation via Conversational Agents: A Case Study in Navigation Support for Visually Impaired People. *International Journal of Human-Computer Interaction*.

[Brooke, 1995] Brooke, J. (1995). SUS: A quick and dirty usability scale. *Usability Eval. Ind.*, 189.

[Brooke, 2009] Brooke, J. (2009). Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4:114–123.

[Cranefield et al., 2017] Cranefield, S., Winikoff, M., Dignum, V., and Dignum, F. (2017). No Pizza for You: Value-based Plan Selection in BDI Agents. *IJCAI 17: Proceedings of the 26th International Joint Conference on AI*.

[Honka et al., 2022] Honka, A. M., Nieminen, H., Similä, H., Kaartinen, J., and van Gils, M. (2022). A Comprehensive User Modeling Framework and a Recommender System for Personalizing Well-Being Related Behavior Change Interventions: Development and Evaluation. *IEEE Access*, 10.

[Kastelein, 2023] Kastelein, P. (2023). Evaluating the accuracy of user values elicited through a textual interface.

[Kließ et al., 2019] Kließ, M. S., Stoelinga, M., and van Riemsdijk, M. B. (2019). From Good Intentions to Behaviour Change: Probabilistic Feature Diagrams for Behaviour Support Agents. *PRIMA 2019: Principles and Practice of Multi-Agent Systems*.

[Krupskis, 2023] Krupskis, M. (2023). Designing graphical user interface to elicit personal values.

[Mendez, 2023] Mendez, S. (2023). Eliciting personal values through isolation questioning: A graphical interface approach.

[Pasotti et al., 2016] Pasotti, P., van Riemsdijk, M. B., and Jonker, C. M. (2016). Representing human habits: towards a habit support agent. *European Conference on Artificial Intelligence and Applications*.

[Tielman et al., 2022] Tielman, M., Jonker, C. M., and van Riemsdijk, M. B. (2022). Telling a computer about your habits and values: Interactively building Action Identification Hierarchies for personalized support.

[Tielman et al., 2018] Tielman, M. L., Jonker, C. M., and van Riemsdijk, M. B. (2018). What Should I Do? Deriving Norms from Actions, Values and Context. *Modelling and Reasoning in Context*.

[van Riemsdijk et al., 2015] van Riemsdijk, M. B., Jonker, C. M., and Lesser, V. (2015). Creating Socially Adaptive Electronic Partners: Interaction, Reasoning, and Ethical Challenges. *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*.

[Vizuroiu, 2023] Vizuroiu, B. (2023). Accuracy of textual interfaces using comparative questions to elicit personal value-related information.