Eliciting Personal Values through Isolation Questioning

A Graphical Interface Approach

Selena Mendez

Supervisor(s): Prof. Dr. Catholijn M. Jonker, Pei-Yu Chen

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: Selena Mendez
Final project course: CSE3000 Research Project
Thesis committee: Prof. Dr. Catholijn M. Jonker, Pei-Yu Chen

Abstract

The alignment of behavior support systems with our personal values becomes increasingly important as behavior support systems continue to influence our daily lives. The purpose of this paper was to explore the use of graphical interfaces and isolation questions to elicit personal values and build accurate user models. An experiment was conducted in two phases to assess the accuracy and usability of the created interface. A comparison was also made with other types of interfaces developed within the research group. This experiment provided valuable insights but also held some limitations. The findings form a valuable contribution for future research and development in building responsible AI and personalized assistance from behavior support agents.

1 Introduction

"If AI systems influence our day-to-day lives, should they not align with our personal values?" Asking such questions is inevitable when dealing with technology that directs our conduct and decisions. Behavioral support agents are systems that assist people in their daily lives. For instance, they provide guidance on how to maintain a healthy lifestyle or how to go about recurrent activities [Kola *et al.*, 2020]. These agents should support us in a flexible and personalized manner to be most successful [Riemsdijk *et al.*, 2015]. For this to be accomplished, the agent must understand the individual values that stand for each user to make tailored decisions.

The concept of a personal value can be defined as a belief or principle that a person holds as being important or desirable, which influences their attitudes and behaviors [Schwartz, 1992]. Each person has a different set of values (e.g., achievement, security, benevolence) and the relevance of a particular varies significantly from individual to individual [Schwartz, 2012]. In general, it is difficult to define the values that represent a person. This becomes even more challenging when delegating this task to an intelligent system without cognitive capability. For behavior support agents to understand their users, it is useful to create user models that take into consideration norms and values [Kließ *et al.*, 2019];[Kola *et al.*, 2020];[Cranefield *et al.*, 2017]. These user models document the connections between the users' desired actions and values, enabling the support agent to be transparent and explainable by making its justifications explicit.

User models must be updated in real-time to provide a personalized user experience. As part of achieving this, behavior support agents must be able to comprehend a user's priorities, current situation, and the impact of the surrounding context on their behavior [Tielman *et al.*, 2018]. It is, however, difficult to determine the values of a user in unexpected situations and obtain realistic and accurate responses. As a result of unexpected situations, the user model needs to be updated. In this study, these situations are referred to as misalignment scenarios. These scenarios act as simulations of real-life scenarios in which users may diverge from their intended course of action.

This project aims to measure the accuracy of a graphical interface to elicit value-related data from participants that are appropriate for personalizing user models. Moreover, the research focuses on exploring a technique for eliciting and modeling human values. Specifically, through a graphical interface that uses questions in isolation. The latter refers to questions that are asked independently of other questions, without any connection or relationship to each other. Therefore, the research question is: "How accurate is a graphical interface that uses questions in isolation, in eliciting personal values?"

The report is structured in the following way. Section 2 describes the methodology of this research, detailing how the research questions were answered. The material preparation, the design process of the graphical interface and the setup of the main experiment for this research can all be found in this section. Section 3 discusses the main findings of this study, here the results of the experiment are analyzed and presented. The ethical implications that were considered during the research are described in section 4. Section 5 focuses on a detailed discussion about the insights and limitations of this research as well as the project's future perspectives. The final section, section 6 presents the report's conclusion.

2 Methodology

To answer the main question of this research; *How to elicit personal values through a graphical interface that uses questions in isolation?* different approaches and methods were used. This section describes the methodology of the research in detail. First, a graphical interface was designed through which the user's values were elicited. Secondly, an experiment was conducted to test the efficiency and accuracy of the interface in creating a model that reflects the actual user's values. Lastly, a detailed data analysis of the experiment results was carried out. This section will be structured in the following way. The material preparation for the research will be discussed in subsection 2.1. In subsection 2.2 the development process of the graphical interface will be detailed. Lastly, the experimental set-up will be stated in subsection 2.3.

2.1 Material Preparation

This sub-section provides information about the preliminary steps conducted prior to the graphical interface implementation and evaluation. As stated before, this study investigated the elicitation of personal values using a graphical interface with isolated questions. As a result, this study focused on situations where a person's values could be expressed. The study identified these situations as misalignment scenarios. The term misalignment refers to the conflict or incongruity between an individual's established values as a result of new experiences or circumstances. If misalignment occurs, the existing user model may need to be updated, causing values to be reassessed or adjusted. It is therefore important to identify these situations and ask users about their values in such circumstances.

All misalignment scenarios for this project were based on the possible changes in behavior and decisions that people might experience when they try to live a healthy lifestyle. Each scenario describes a goal, a misalignment reason, and a set of

related values. For example, if someone wants to drink more water and avoid sugary drinks, a misalignment scenario could arise if the context of a party is introduced. There is a possibility that the user might consume a sugary drink instead of water, as the social acceptance of drinking water is lower than the acceptance of drinking sugary drinks. Such misalignment scenarios are formally described as in the following example:

Goal: drink more water

Misalignment reason: attending a party

Related values: health, enjoyment, social acceptance.

A visualization of misalignment scenarios was employed to model users' values. For this purpose, a behavior tree approach was used. A behavior tree is a way to structure switching between tasks of an autonomous agent or virtual entity [Colledanchise and Ogren, 2018]. This representation can also be used in personal value modeling. Here, the behavior tree describes the different decisions a user could make and how they relate to specific values. The behavior tree representation of the misalignment scenario described in the previous section will be shown next in Figure 1.

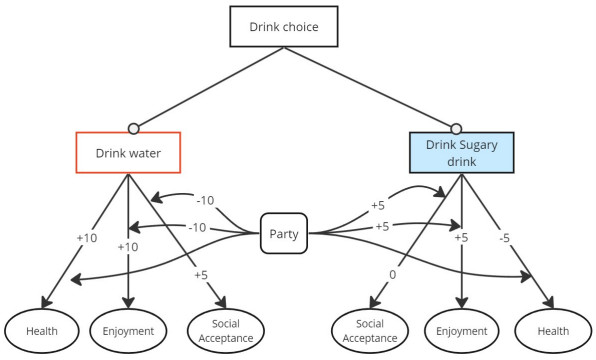

Figure 1: Behavior tree representing the modeling of personal values and the updates to the model. This tree represents the misalignment scenario of drinking water/sugary drinks at a party.

In this research, multiple brainstorming sessions were held to create the misalignment scenarios used in the elicitation process. There were originally fifteen scenarios created, divided into five goals. After analyzing all the scenarios, four scenarios were chosen. These scenarios were chosen after discussing within the research group which would be more appropriate for a diverse and more accurate elicitation process. A detailed description of the four scenarios can be found in Appendix A.

This sub-section presented the outline of misalignment scenarios and their relation to this study. How the graphical interface was implemented, the design choices and, an overview of the questions it contained will be discussed in the next section.

2.2 Implementation of the Graphical Interface

A detailed discussion of the development of the graphical user interface will be presented in this sub-section. The interface played a significant role in the research project since it enabled the elicitation of the user's values. The graphical interface used illustrations, color schemes, and widgets to convey information to the user and receive input from them. Additionally, the interface used text to ask the user questions about the misalignment scenarios.

The focus of this research was on asking questions in isolation. The term "questions in isolation" describes an approach or method that presents individual questions without explicit linkage or context to other questions. Using this method, individuals are asked to provide responses or information without being influenced by questions prior to or following them. Consider, for instance, the example of the misalignment scenario described in 2.1. For this scenario, the "in isolation" approach would ask how the user agrees with the following statement: *It is socially acceptable to drink water*; *It is socially acceptable to drink sugary drinks.*

Since this project's objective was to elicit personal values and not to develop a fully functional system, a high-fidelity prototype was built to simulate the way in which a complete system would elicit personal values. This graphical interface was created with Figma[1] and can be accessed at [https://www.figma.com/file]. The prototype was designed to be interactive, simple to use, and easy to understand. Each screen displays a question in isolation about a possible action and a related value. A 5-point Likert scale [Joshi *et al.*, 2015], is used to rate the possible answers, ranging from strongly disagree to strongly agree. Figure 2 shows two screens of the graphical interface.

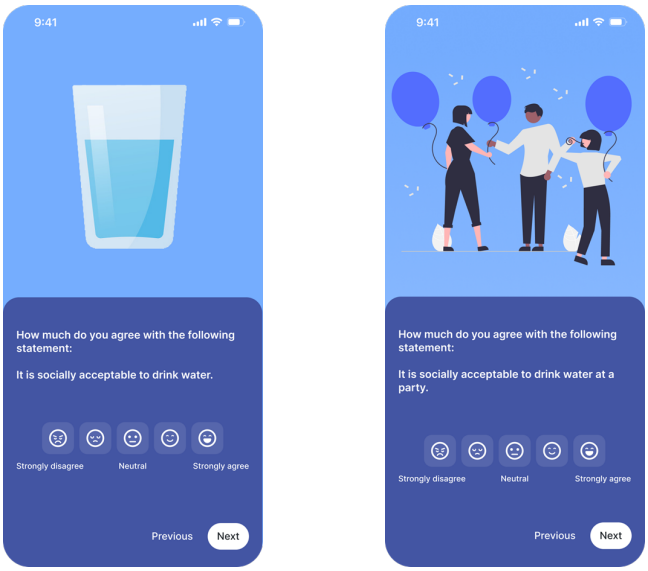

Figure 2: Screenshots of the graphical interface. Here, questions are asked about the perceived social acceptance of drinking water.

Figma prototypes cannot save user responses and interactions. For this reason, a Python script was developed to create

[1]Figma [Collaborative web application for interface design]. (2023). Retrieved from https://www.figma.com/

the user model and calculate the values and their updates after a misalignment scenario. The complete code can be found in a dedicated GitHub repository[2]. The script was used to fill in the responses of the user and to summarize each value's level of importance. Additionally, it detailed how the value-action relation changed when the context was added. The model was created using the responses to the questions described before. The possible responses to the questions were; strongly disagree, disagree, neutral, agree, and strongly agree. These responses got a value in the model of -10, -5, 0, 5, and 10 respectively. The Python script was particularly useful during the experimentation phase, where it enabled the user to review the obtained model and measure its accuracy.

Different Approaches within the Research group

It is important to note that other members of the research group focused on other types of interfaces. Two members studied graphical interfaces, two studied textual interfaces, and one studied audio interfaces. As a result, five interfaces were created, including the interface related to this project. Another crucial difference between the interfaces of the different team members was the way in which the questions were formulated. Two of the researchers used comparison questions, while the other three used isolation questions. The "in comparison" approach would ask how the user agrees with the statement: *Drinking water is more socially acceptable than drinking sugary drinks*.

The purpose of this sub-section was to describe the development process of the graphical interface for this study. In addition, it explained how the questions were formulated in the interface. Lastly, it included an explanation of how a Python script was used to model the values. The script was used as a tool to support the experimentation phase of this research. The next section will provide a detailed description of this experiment as well as all the information regarding its setup.

2.3 Experiment Set-Up

The experiment used for this research aimed to determine the accuracy of the created graphical interface to elicit personal values. Additionally, it studied the impact of the isolation questions in the elicitation process. Participants interacted with an interface and answered questions that represent misalignment scenarios in their behavior. Based on their responses, a user model was created that identified the personal values of the participants. Assessing the accuracy of the model was done by asking the user whether the obtained model reflect their personal values.

Experimental Design

The experiment of this research had two phases. The objective of the first phase was to assess the perceived accuracy of the graphical interface in capturing participants' values and decision-making processes based on misalignment scenarios. The objective of the second phase was to determine if questions in isolation and the type of interface had an impact on the user's responses. Both aims were addressed by one experiment. The experimental design to reach the first objective was quasi-experimental, with a focus on measuring the graphical interface performance rather than manipulating an independent variable. For the second aim, this experiment could be seen as part of a broader experiment where the type of question and the type of interface served as an independent variables in the analysis. Here, the research focused on questions in isolation and graphical components while other researchers in the group focus on other types of questions and types of interfaces.

Participants

For this study, a total of 19 participants were recruited from personal connections and social networks. The first phase of the experiment involved 15 participants. Furthermore, 4 participants were used in the second phase of the experiment to interact with all the created interfaces within the research group. The participants were computer literate and were comfortable using smartphones and other technological devices in their daily lives. The participants were aged 18 to 65 and were diverse in gender. Furthermore, none of the participants were visually impaired. Lastly, all participants declared no conflicts of interest.

Measures

First, participants were asked about their age and gender in a general questionnaire. Secondly, the users read and signed a consent form regarding the risks of the experiment. Thirdly, the participants interacted with the graphical interface, and their responses were recorded to create a user model. Afterwards, the users' opinions were gathered to evaluate if the selected values were accurate and if the obtained model represented them. Lastly, the participants were asked to fill in a standard system usability scale questionnaire [Brooke, 1995]. The complete system usability scale questionnaire can be found in Appendix B.

Procedure

The experiment was conducted with one participant at a time. Participants were welcomed first, followed by the reading and signing of a consent form. The consent form that was presented to the user can be found in Appendix C. Afterwards, participants interacted with the graphical interface, answering questions related to the misalignment scenarios selected before. The examiner recorded their responses and created their user model with the corresponding program in Python. Participants were shown the calculated values based on their answers. In the final step, the participants were asked how accurately the calculated model was. To do this, users were asked to change the rating of each value in the obtained model if they believed it was incorrect.

2.4 Data Collection and Analysis methods

After the experiment was conducted, different sets of data had to be collected. The first step was to collect responses to the graphical interface questions. Appendix D contains a list of all questions. In total, 56 answers were collected from each participant. Each answer corresponds to one of the four misalignment scenarios described in Appendix A. A second step involved collecting the age and gender of each participant and matching them with their responses. The third step was storing the Python script output summary. Thereafter,

[2]https://github.com/SelenaMendez2801/usermodel.

the responses to the system usability scale questionnaire were collected together with the changes each participant made to the summary and the obtained behavior trees. Lastly, the data from the experiments of the other members of the research team were used in this study and shared within it. All the data mentioned before was saved in a data set for storage, manipulation, and analysis.

A two-phase data analysis was conducted for this project. The first phase examined the perceived usability and accuracy of the graphical interface with isolated questions. The perceived accuracy was based on the users' responses and corrections to the model. The difference between the obtained model and what the user expected it to be served as the main indication of the overall accuracy of the interface in eliciting personal values. To calculate the difference, two distance measures were used. Specifically, hamming distance and difference measures were applied. Basic statistics, such as averages, standard deviations, and means, were also used to analyze the responses. Additionally, the responses to the system usability scale questionnaire were examined to determine whether the system was usable and if any improvement points were needed.

The second part of the data analysis compared the results of the created interface with those of the other interfaces created within the research group. Four participants experimented with five different interfaces. Then, a comparison was made between the different interface types (graphical, textual, audio) and between the types of questions (in isolation or comparison).

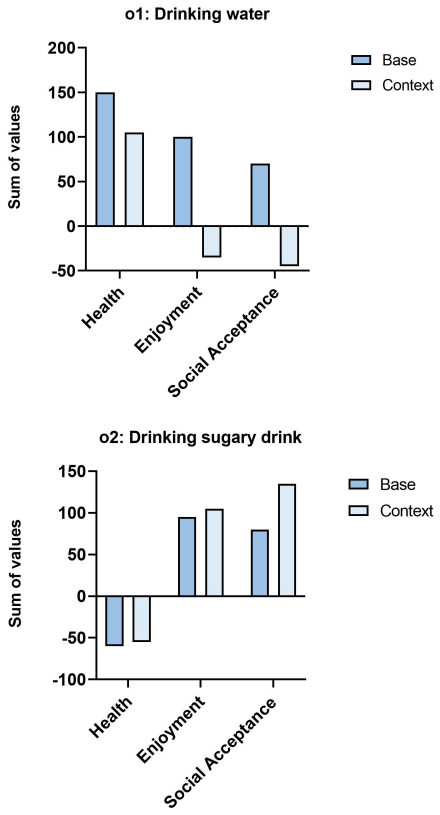

Figure 3: Aggregated responses for scenario 1.The top graph corresponds to option 1, the bottom corresponds to option 2.

3 Results

The aim of this project was to elicit personal values through a graphical interface that used isolation questioning. To evaluate the perceived accuracy and usability of the graphical interface in eliciting those values, an experiment was carried out. A detailed explanation of the experiment can be found in section 2.3. All the results of this experiment will be discussed in this section. The outcomes of the interaction from the users with the graphical interface will be shown in subsection 3.1. Additionally, sub-section 3.2 will demonstrate the outcomes from the comparison between the different interface types created within the research group. Finally, this section will conclude by detailing the results of the system usability scale questionnaire, in sub-section B

3.1 Elicitation Process

In the first phase of the experiment, the personal values of fifteen participants were elicited. Each participant answered the fifty-six questions that were displayed in the graphical interface. The tables in Appendix E summarize the answers of each participant. Two major things will be discussed in this subsection. The aggregated results of the experiment will be presented first. Then, the aggregated corrections to the model from each participant will be displayed. Figures 3 to 6 show the aggregated results of the experiment.

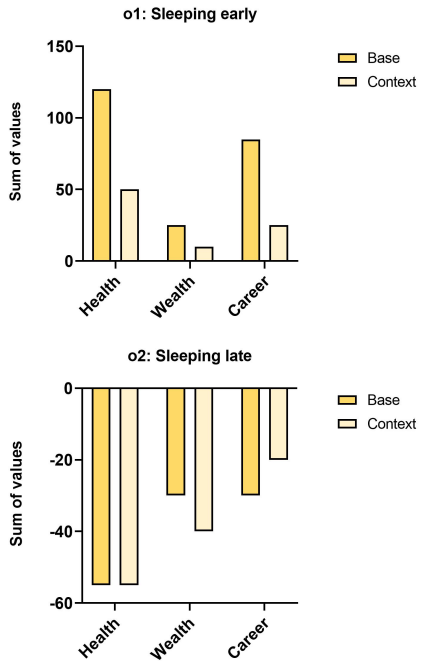

Figure 4: Aggregated responses for scenario 2. The top graph corresponds to option 1, the bottom corresponds to option 2

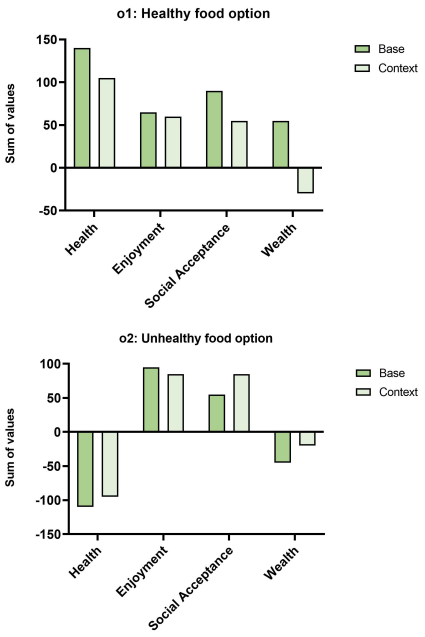

Figure 5: Aggregated responses for scenario 3. The top graph corresponds to option 1, the bottom corresponds to option 2

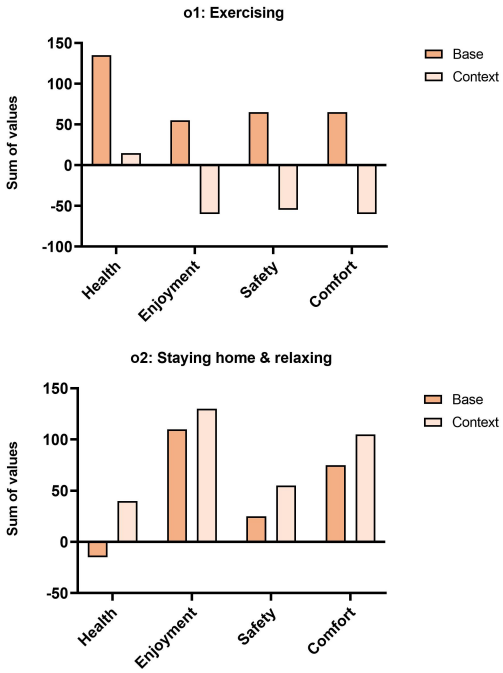

Figure 6: Aggregated responses for scenario 4. The top graph corresponds to option 1, the bottom corresponds to option 2

The figures represent the summed-up responses from the fifteen participants of the experiment. For each question, five possible responses were presented; strongly disagree (-10), disagree (-5), neutral (0), agree (5), and strongly agree (10). The answers from all the participants were added together to create the bars in the graphs. These graphs capture how the participants perceive the possible actions in a misalignment scenario. Additionally, it shows how the responses from the users vary when contextual factors are introduced. From the graphs, it is clear that introducing context changes the perception of the personal values related to an action. For "o1", the socially perceived "healthier" option, these changes are all negative. When it comes to "o2", the "unhealthier" option, the change is almost always positive, but not always.

Other crucial outcomes from this experiment were the corrections to the models that each participant provided. This indicates the perceived accuracy of the elicitation process. Two distance measures were used to analyze these corrections. First, the hamming distance was used to count the modifications made per scenario, per type (base or context). Table 1 presents the main statistics about the hamming distance of all results. The second distance measure was the magnitude of the change per scenario, per type (base or context). Table 2 details the main statistics about this measure.

Table 1: Standard statistics, hamming distance of all results.

Sample Size	Mean	Median	Standard Deviation
15	1,333	0	2,193

Table 2: Standard statistics, value difference of all results.

Sample Size	Mean	Median	Standard Deviation
15	8	0	13,065

The total error over all the participants was calculated by taking the dividend value of all magnitudes of the changes (per misalignment scenario, per type) and the hamming distance (per misalignment scenario, per type). See Formula 1. Based on this calculation, the aggregated error for all participants was 30,33. The average error per participant was 2,022. This means that each participant had total modifications of an average magnitude of 2,022.

$$\sum_{i=1}^{n} \frac{\sum value\ difference}{\sum hamming\ distance} \tag{1}$$

This sub-section presented the main results of the first part of the experiment. The aggregated data showed the perceived value-action relationship in general and context scenarios. Additionally, it detailed the modifications that participants made to the model. In the following sub-section, the second part of the experiment will be discussed; a comparison of interface types within the research group.

3.2 Comparison between Interface Types

For the second phase of the experiment, some participants interacted with all the different interfaces created within the research group. The aim was to compare the different types of questioning (in isolation and in comparison) and the interface type (graphical, audio, and textual). Table 3 shows the summed-up changes that each participant made to the obtained models.

Table 3: Comparison between interfaces. For this calculation, the same principle as in Formula 1 was used. In the table, a/b can be interpreted as the average magnitude of the correction a participant gave. Here, "a" is equal to the summed-up differences a user gave. Moreover, "b" represents the number of changes in all behavior trees (hamming distance).

Participant	**Graphical + Isolation**	**Graphical + Comparison**	**Textual + Isolation**	**Textual + Comparison**	**Audio + Isolation**
1	x	7/2	40/8	10/2	0/0
2	25/5	x	25/4	110/10	5/1
3	0/0	45/9	x	70/10	0/0
4	20/4	75/8	40/5	x	5/1
5	25/5	75/13	55/9	30/5	x
Average	3.75	5.911	6.340	7.25	2.5

A higher error was found in the different types of interfaces. According to this, the following interface types are ranked in ascending order: Audio + Isolation, Graphical + Isolation, Graphical + Comparison, Textual + Isolation, Textual + Comparison. This sub-section detailed the results from the second phase of the experiment. Results of the experiment related to usability will be presented in the next section.

3.3 System Usability Scale Questionnaire

In the two phases of the experiment, participants were asked to fill in the standard system usability scale questionnaire described in [Brooke, 1995]. The scores for each participant were calculated in the way the literature describes. Table 4 shows the resulting score per participant.

Table 4: Standard statistics system usability score questionnaire.

Sample Size	Mean	Median	Standard Deviation
19	86	87	8,185

To interpret the scores, the guidelines in [Brooke, 1995] were followed. Based on this information, fifteen participants rated the system as excellent, three participants rated it as good, and one participant rated it okay. Appendix F shows the rules that were followed for the analysis of the scores.

The data collected throughout the two phases of the experiment were presented in this section. It included detailed information on the participants' responses, their modifications to the model as well as their system usability scale (SUS) scores. In the next section, the results and main conclusions from the data will be discussed in greater detail.

4 Ethical & Responsible research

To ensure participant protection and well-being, some ethical considerations were made during the design and execution of this research study. The following section highlights the key ethical concerns that were addressed throughout the study.

Informed consent was adhered to in this study. Each participant was made aware of the purpose, procedures, risks, and benefits of the study before participating. It was emphasized that the participation was voluntary and that questions could be asked. The participants were informed that they could withdraw from the study at any time without repercussions upon signing the written consent form.

Participants' privacy and confidentiality were also protected in this study. All participants' data were kept strictly confidential and used only for research purposes. The participant's data was anonymized and pseudonymized to remove any personal information. In addition, any data shared or published from this study were aggregated and anonymized. Moreover, no invasive procedures or interventions were involved in the study, so participants were protected from potential harm or risks. The participants were not exposed to risks beyond those they might typically face every day. According to ethical guidelines, all participants were monitored for their well-being and safety throughout the study.

During the research process, measures were taken to ensure transparency and debriefing. Upon completion of the study, participants received a summary explaining the purpose, results, and use of their contributions. Consequently, participants learned more about the study's objectives, and their concerns and questions could be addressed. Furthermore, broader ethical considerations were considered in addition to ethical implications for participants. In the study, conflicts of interest were avoided to preserve the integrity of the study. Several steps were taken to ensure that the research was conducted independently and without undue influence from outside sources.

The study adhered to ethical standards, maintained confidentiality of data, and protected participants' rights and well-being. As a result of these measures, the research was conducted ethically and with the highest regard for the welfare of the participants.

5 Discussion

This section aims to provide an in-depth analysis and discussion of the main findings of the study. First, the insights of the experiment will be discussed in sub-section 5.1. Secondly, the principal limitations that were discovered will be pointed out in sub-section 5.2. Lastly, the recommendations for future studies will be stated in sub-section 5.3.

5.1 Key Findings & Insights

This section discusses the main findings obtained during the experiment and some reasoning derived from them. First, the findings from the first phase of the experiment will be discussed. This will be followed by an overview of the results obtained from the comparison between interface types. Lastly, insights into the system usability scale questionnaire will be detailed.

Insights about the Experiment's First Phase

Based on the results of the experiment and the data collected, it is evident that introducing context to common activities can change perceptions of what certain decisions/actions mean. Sub-section 3.1 indicates that participants prioritize different personal values when certain situations occur. Therefore, it is crucial to continue research that will help behavior support technology adapt to these situations and meet the preferences of users.

The perceived accuracy of the graphical interface in eliciting personal values was calculated by asking experiment participants to modify the obtained user models in behavior tree representations. Based on Tables 1 and 2, most participants made no to minimal changes to the obtained model. This suggests that graphical interfaces are quite engaging and precise when gathering user information.

Following the experiment, the moderator answered participants' questions and sometimes held short discussions with them. From these discussions, some new insights were gained. First, it was evident that some participants put greater emphasis on long-term values than short-term values, which resulted in responses that didn't match actual behaviors. When updating the model, this must be taken into account. Specificity in the way of questioning is crucial to prevent dubious interpretations. Secondly, some participants thought there were too few answer possibilities. The use of the Likert scale as a response choice may have affected the accuracy of the process. Participants' opinions and preferences might have been better understood if the choice of answer options had been broader and more specific.

In the experiment, it was observed that users tend to overread some questions without fully comprehending what the question is asking for. The information provided before each question may not be properly retained by the user. In particular, this is problematic when there are many questions to answer. Different ordering of questions could improve elicitation reliability. As an example, asking about the base scenario immediately after asking about the context scenario may improve response accuracy. Furthermore, it may also be possible to mention the specific details of the misalignment scenarios on each screen of the graphical interface.

Insights about the Experiment's Second Phase

The study examined the ability of interface types to elicit users' values. A comparison of the types of questions (in isolation, as a comparison) was also used to evaluate whether particular design decisions were effective in capturing responses from participants. Each interface developed by the research group had its advantages, as demonstrated by experimental results. The accuracy, usability, and satisfaction of each interface must, however, be further analyzed and compared.

Based on Table 3, the interfaces can be ranked in ascending order of perceived accuracy as follows: Audio + Isolation, Graphical + Isolation, Graphical + Comparison, Textual + Isolation, and Textual + Comparison. This ranking indicates the perceived accuracy and effectiveness of each type of interface in eliciting values based on participant responses. However, this measure is subjective and possibly biased as it is dependent on participants' opinions.

The experiments found that interfaces with isolation questioning generally performed better than interfaces with comparison questioning since fewer modifications to the model were required. In addition, audio and graphic interface types outperformed textual-based interfaces in terms of capturing participants' values. All this information could suggest that focusing on these types of interfaces and questioning could be beneficial for eliciting information to create more personalized AI models.

Insights about the SUS questionnaire

On the system usability scale questionnaire, most participants rated the system as usable. In 15 of the surveys, participants rated the system as excellent, indicating that they were highly satisfied and found it easy to use. This positive feedback indicates that the system facilitates user interaction and achieves its intended goal. To improve usability, it is necessary to seek feedback from users who rated the system as "okay". Furthermore, other user evaluations could be conducted on the interface to enhance it even further.

5.2 Limitations

To understand the research findings comprehensively, it is crucial to acknowledge the study's limitations. The purpose of this section is to discuss and identify the limitations identified in the experiment, the graphical interface, and the questioning. Additionally, the limitations related to the comparison of the different interfaces will be detailed.

First, there were relatively few participants recruited, resulting in limited statistical significance and generalizability. Despite the insights gained from the study providing a starting point, a larger sample size would have strengthened the research outcomes. Secondly, participants' responses could also have been affected by fatigue, mood, and time of day. This may have introduced confounding variables that could have influenced the study's results. Additionally, sitting close to the participants and collecting their responses may have introduced biased responses. It is possible that participants were influenced to respond in ways that aligned with their expectations or fit in with perceived norms. This could have negatively affected their authenticity.

Another limitation of the experiment is the use of a system usability score questionnaire as the sole measure of system usability. Using this questionnaire might not capture nuanced feedback or overlook specific aspects of the user experience. For instance, user interviews or usability testing could have provided more insight into the system's strengths and weaknesses.

As part of the research, the experiment results were compared to results obtained from other types of interfaces. There were also limitations found in the comparison. Although consistency in questions and design was used to ensure comparability, it is important to recognize the inherent differences between textual, graphical, and audio interfaces. Individual

preferences and context can affect the usability and likability of a system. This makes direct comparisons difficult. In addition, subjective measures such as perceived accuracy and ease of use complicate evaluations. As a result, it was difficult to establish objective measures for comparison.

It is important to consider all the limitations mentioned above when interpreting the study findings as they may affect their generalizability, reliability, and validity. Future research should expand the sample size, consider additional participant factors, refine experimental procedures, and explore alternative measures to address these limitations. In the next sub-section, the future prospects for this research will be discussed.

5.3 Future Prospects

The findings from this study may provide opportunities for further exploration and development in the field. This subsection outlines future prospects to consider and how to take advantage of these insights. Future research should focus on deepening the understanding of the implications of the findings and exploring ways to apply them in the real world.

As a first step, longitudinal studies may be valuable for a better understanding of how user values evolve. Researchers can capture personal values dynamics and fluctuations by observing participants' values and behaviors over an extended period. As a result, more accurate and significant user models can be developed. In this study, all questions were asked simultaneously. This approach provides an initial snapshot of the user's values, but it overlooks the possibility of changes over time. To truly understand the complexity of personal values, longitudinal studies are necessary.

A second possibility to consider in the future is to empower users and engage them in the elicitation process. It may be viable to conduct future research that empowers individuals to have agency over their own value models and decision-making processes. This would allow individuals to directly engage with and shape behavior support systems.

Future research could also consider the combination of different technologies in the interface implementation. For instance, collaboration filtering techniques and recommendation systems technologies could be used to elicit personal values more effectively. These systems allow users to tailor recommendations based on their preferences and interests by leveraging the user model. They also analyze similarity patterns with other users. Another example could be expanding on the interface types. Virtual reality (VR) and augmented reality (AR) could be considered for eliciting and modeling user values in future studies. It could be very useful to explore a wide array of interface options to better understand how different modes impact the personalized user experience.

Furthermore, increasing the study's objectivity is another future prospect. To gain a deeper understanding of how user interfaces affect value modeling, a larger sample size would lead to statistically significant results. Moreover, qualitative methods such as interviews and focus groups could be used in conjunction with quantitative data analysis. This would allow us to gain a better understanding of user experience and value modeling processes.

Finally, future research should refine evaluation measures to make interfaces more consistent and comparable. To establish more concrete comparisons between interfaces, objective measures like task completion time, value elicitation accuracy, and system usability metrics can be used. To ensure greater comparability and reliability when assessing user interface effectiveness, researchers should use standardized evaluation criteria.

These future prospects will contribute to the development of more personalized, adaptable, user-centric technology by advancing behavior support systems. Future behavior support systems will be more effective and responsible if various interface options are included, advanced technologies are integrated, longitudinal studies are conducted, contextual factors are taken into account, and ethical considerations are emphasized.

6 Conclusion

The use of behavior support agents is on the rise in all aspects of life. This makes it imperative that user models be developed and updated in a manner that aligns with users' personal values. This research project aimed to develop a graphical interface to elicit value-related data from participants that were appropriate for personalizing user models. Moreover, the research focused on exploring questions in isolation as a technique for gathering information

The two-phase experiment provided different insights in this research. First, participants' responses revealed that the particular graphical interface used for this study required relatively few modifications to its generated user model. Secondly, participants expressed different perceptions of their actions when the context was introduced to a scenario. Thirdly, the comparison between types of interfaces indicated that graphical and audio interfaces were perceived as having greater accuracy. Moreover, questions in isolation were perceived to be more accurate than questions in comparison. Finally, the majority of respondents rated the system as excellent using the System Usability Scale (SUS) questionnaire.

Despite the insights gained, this research also revealed some limitations. The experiment was conducted with only a couple of participants, which affected the study results' significance. Additionally, confounding variables such as participants' moods, explanations of the procedure, etc could have introduced some biased answers to the experiment. In addition, using the system usability scale questionnaire (SUS) as the only measure of usability lessened the significance of the results. Finally, interface types differ in many ways, making comparison difficult.

The implementation of several recommendations could make behavior support systems more effective at eliciting personal values. For instance, longitudinal studies and the use of advanced interface technologies could help make behavior support systems more effective. Additionally, allowing individuals to directly engage with and shape behavior support sys-

tems could improve them. Finally, making these types of experiments more objective could make the findings more significant. This can be accomplished by increasing consistency and recruiting more participants.

As a result of this study, valuable insights into value elicitation and decision-making processes have been gained, which is of great benefit to the field of behavior support systems. A critical aspect of eliciting individual values through graphic interfaces is accuracy, usability, and feedback from the user. The findings of this study will assist in future research and development to support value elicitation and personalized behavior support in the future.

References

[Brooke, 1995] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[Colledanchise and Ogren, 2018] Michele Colledanchise and Petter Ogren. *Behavior Trees in Robotics and AI: An Introduction*. 07 2018.

[Cranefield *et al.*, 2017] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. pages 178–184, 08 2017.

[Joshi *et al.*, 2015] Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Pal. Likert scale: Explored and explained. *British Journal of Applied Science Technology*, 7:396–403, 01 2015.

[Kließ *et al.*, 2019] Malte Kließ, Mariëlle Stoelinga, and M. Riemsdijk. *From Good Intentions to Behaviour Change: Probabilistic Feature Diagrams for Behaviour Support Agents*, pages 354–369. 10 2019.

[Kola *et al.*, 2020] Ilir Kola, Catholijn M. Jonker, and M. Birna van Riemsdijk. Who's that? - social situation awareness for behaviour support agents. In Louise A. Dennis, Rafael H. Bordini, and Yves Lespérance, editors, *Engineering Multi-Agent Systems*, pages 127–151, Cham, 2020. Springer International Publishing.

[Riemsdijk *et al.*, 2015] M.B. Riemsdijk, Catholijn Jonker, and Victor Lesser. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. 2:1201–1206, 01 2015.

[Schwartz, 1992] Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press, 1992.

[Schwartz, 2012] Shalom Schwartz. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2, 12 2012.

[Tielman *et al.*, 2018] Myrthe L. Tielman, Catholijn M. Jonker, and M. Birna van Riemsdijk. What should i do? deriving norms from actions, values and context. In *MRC@IJCAI*, 2018.

A Misalignment scenarios

Misalignment scenarios are situations where there is a conflict or incongruity between an individual's established values. For this project, four scenarios were identified to base the questions of the elicitation process. The four misalignment used in this research are described in detail in this Appendix.

Misalignment 1

Goal: Drinking more water

Misalignment reason: You are at a party

Value: Enjoyment / Social acceptance

Misalignment 2

Goal: Exercising

Misalignment reason: Bad weather (too hot, too cold, raining, snowing, etc.)

Value: Enjoyment / Safety / Comfort/ Health

Misalignment 3

Goal: Eating healthier

Misalignment reason: You are eating out with other people

Value: Enjoyment / Social acceptance / Wealth / Health

Misalignment 4

Goal: Better sleep schedule

Misalignment reason: Work till midnight

Value: Wealth / Career / Health / Enjoyment

B System Usability Scale Questionnaire

Please rate the following statements based on your experience using the system. Use a scale from 1 to 5, where 1 represents "Strongly Disagree" and 5 represents "Strongly Agree."

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need assistance to be able to use this system
- I found the various functions in this system well integrated
- I thought there was too much inconsistency in this system.

C Consent Form

You are being invited to participate in a research study titled Eliciting personal values from the users to build responsible AI. This study is being done by bachelor students Pien Kastelein , Martynas Krupskis, Selena Mendez, Beatrice

Vizuroiu, and Elvira Voorneveld from the Techincal University of Delft. The responsible supervisor is PhD candidate Pei-Yu Chen.

The purpose of this research study is to investigate different modalities and ways to elicit the information on conflicting values in daily scenarios. The study will take you approximately 30 minutes to complete. The data will be used for the course Research Project. If the results are interesting, we might publish a joint paper.

Our study is regarding eliciting necessary information to update and modify the user model in support agents that aim to help user adopt healthier lifestyles. In the study, you will read several scenarios where a user makes a decision that in not in line with their goals of being healthier, but they do so because they value something else as well (e.g., choose to drink beer instead of water at a party). We ask you to imagine being in the scenarios and make such choices. You will be asked a series of questions about why you would make the choice in terms of the relationship of the conflicting values. Afterwards, you will be asked to answer a questionnaire regarding the usability and correctness, and open questions with regard to your opinions of the interfaces.

As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. Personal information that could identify you as an individual will not be asked. If you inadvertently provide personal information while answering open-ended questions, it will be removed to ensure anonymity. Anonymized answers to open questions may be quoted in research publications. Your anonymized answers to both closed and open-ended questions will be archived in the repository of the four technical universities in the Netherlands, 4TU.ResearchData. The data will be made available to the public for non-commercial use, allowing for future research and education.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions. If you choose not to participate or withdraw at any time, there will be no consequences, and none of your information will be saved or stored.

If you have any questions about the research study, you can contact the lead researcher Pei-Yu Chen (p.y.chen@tudelft.nl).

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated [*DD/MM/YYYY*], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves:	☐	☐
• *Reading and answering value-related questions in a given scenario with the interface prototype.* • *Answering questionnaires regarding usability and correctness.* • *Answering open questions regarding your opinions on the given interface.*		
4. I understand that the experiment will take approximately 30 minutes. The study will end by July 2023 as the course is finished.		
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study involves the risks of 1) impersonating scenarios where I'm not complying with the goals, and 2) being asked about personal values as explanations . I understand that these will be mitigated by being able to withdraw at any time.	☐	☐
9. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach	☐	☐
• *Anonymous data collection and aggregation* • *Secure data storage on 4TU.ResearchData.*		
11. I understand that the personal data I provide (for administrative purpose) will be destroyed as soon as they study ends (July 2023).	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
12. I understand that after the research study the de-identified information I provide will be used for the final course report and possibly a publication.	☐	☐
13. I agree that my responses, views or other input can be quoted anonymously in research outputs	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
16. I give permission for the de-identified data to the questionnaire that I provide to be archived in 4TU.ResearchData repository so it can be used for future research and learning.	☐	☐
17. I understand that access to this repository will be made available to the public for non-commercial use.	☐	☐

D List of Questions

This appendix contains a listing of all the questions used in the graphical interface to elicit the personal values from the user. The questions are divided between the four misalignment scenarios described in Appendix A

Misalignment 1

Base Questions: Without context

Imagine the following setting:

You have decided that you should drink more water and have been doing so every evening in the past week. The alternative to drinking water is to drink a sugary drink (e.g. beer, cola, juice). This beverage is unhealthier than water, but you enjoy drinking it more.

Answer the following questions.

- It is healthy to drink water.
- It is enjoyable to drink water.
- It is socially acceptable to drink water.
- It is healthy to drink sugary drinks.
- It is enjoyable to drink sugary drinks.
- It is socially acceptable to sugary drinks.

Context Questions: With context

Imagine the following setting:

There is a party coming up which you are going to attend. At

the party there is both your sugary drink of choice (e.g. beer, cocktail, juice) and water available.

Answer the following questions.

- It is healthy to drink water at a party.
- It is enjoyable to drink water at a party.
- It is socially acceptable to drink water at a party.
- It is healthy to drink sugary drinks at a party.
- It is enjoyable to drink sugary drinks at a party.
- It is socially acceptable to sugary drinks at a party.

Misalignment 2

Base Questions: Without context

Imagine the following setting:

You have decided that you should sleep early and have been doing so every night in the past week. The alternative to sleeping early is to stay up late and have time left over to do other things (e.g. work, hobbies).

Answer the following questions.

- It is healthy to sleep early.
- Sleeping early increases your wealth.
- Sleeping early positively influences your career.
- It is healthy to stay up late.
- Staying up late increases your wealth.
- Staying up late positively influences your career.

Context Questions: With context

Imagine the following setting:

There is a work deadline coming up which you need to make sure to meet. To get the work done in time, you can choose to stay up late and work or to sleep early and try to finish it the next day, risking missing the deadline.

Answer the following questions.

- It is healthy to sleep early when you have a work deadline.
- Sleeping early increases your wealth when you have a work deadline.
- Sleeping early positively influences your career when you have a work deadline.
- It is healthy to stay up late when you have a work deadline.
- Staying up late increases your wealth when you have a work deadline.
- Staying up late positively influences your career when you have a work deadline.

Misalignment 3

Base Questions: Without context

Imagine the following setting:

You have decided that you should maintain a more nutritious diet and have been eating healthy food options in the past week. The alternative to eating healthy is to eat more processed and high-fat foods (e.g. hamburger, fries, pizza). These types of food are unhealthier, but you enjoy eating it more.

Answer the following questions.

- It is healthy to eat a healthy food option.
- It is enjoyable to eat a healthy food option.
- It is socially acceptable to eat a healthy food option.
- Eating a healthy food option positively influences my wealth.
- It is healthy to eat processed / high-fat foods.
- It is enjoyable to eat processed / high fat foods.
- It is socially acceptable to eat processed / high fat foods.
- Eating processed / high fat foods positively influences my wealth.

Context Questions: With context

Imagine the following setting:

This evening you and your friends are going to dine at a restaurant that serves both fast food and fine dining meals. Because the healthy alternative is extremely expensive, you decide to order fast food. So do more than half of your friends that are at the restaurant with you.

Answer the following questions.

- It is healthy to eat a healthy food option when dining at a restaurant with friends.
- It is enjoyable to eat a healthy food option when dining at a restaurant with friends.
- It is socially acceptable to eat a healthy food option when dining at a restaurant with friends.
- Eating a healthy food option positively influences my wealth when dining at a restaurant with friends.
- It is healthy to eat processed / high-fat foods when dining at a restaurant with friends.
- It is enjoyable to eat processed / high fat foods when dining at a restaurant with friends.
- It is socially acceptable to eat processed / high fat foods when dining at a restaurant with friends.
- Eating processed / high fat foods positively influences my wealth when dining at a restaurant with friends.

Misalignment 4

Base Questions: Without context
Imagine the following setting:

You have decided to start running 3 km daily to improve your health and strength. Before making this decision, you did not have a clear activity defined and were simply scrolling through social media/watching a movie. Consider the alternative to running 3 km daily to be watching a movie.

Answer the following questions.

- It is healthy to work out.
- It is enjoyable to work out.
- Exercising positively influences your safety.
- Exercising positively influences your comfort.
- It is healthy to stay home and watch a movie.
- It is enjoyable to stay home and watch a movie
- Staying home and watching a movie positively influences your safety.
- Staying home and watching a movie positively influences your comfort.

Context Questions: With context
Imagine the following setting:

The alternative to running 3 km daily is to watch a movie. Today the weather has been very bad. It rained the whole day and the temperatures fell, therefore, you have decided to stay inside and watch a movie today.

Answer the following questions.

- It is healthy to work out in the rain.
- It is enjoyable to work out in the rain.
- Exercising in the rain positively influences your safety.
- Exercising in the rain positively influences your comfort.
- It is healthy to stay home and watch a movie when it is raining.
- It is enjoyable to stay home and watch a movie when it is raining.
- Staying home and watching a movie positively influences your safety when it is raining.
- Staying home and watching a movie positively influences your comfort when it is raining.

E Complete Experimental Results

This appendix shows the complete data gathered from the experiment conducted in this research. The data corresponds to the responses from the users to the questions asked in the graphical interface. Each row is identified with an ID and corresponds to a participant of the experiment. The tables are divided and colored to differentiate the responses to the questions of each scenario.

General Questions

ID	Scenario 1						Scenario 2						Scenario 3								Scenario 4							
	Health		Enjoyment		Social Acceptance		Health		Wealth		Career		Health		Enjoyment		Social Acceptance		Wealth		Health		Enjoyment		Safety		Comfort	
	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2
1	10	-10	10	10	0	5	10	0	5	-5	10	0	10	-10	5	5	10	10	-10		10	-5	10	5	10	5	5	10
2	10	-5	0	10	10	5	0	-5	-5	0	0	0	10	-5	5	10	10	-5	10	0	10	5	-5	10	-5	10	5	10
3	10	-5	5	10	10	10	10	0	5	0	0	0	10	-5	5	10	10	5	0	0	10	10	5	10	5	5	0	10
4	10	-5	5	5	10	5	5	-5	0	5	5	0	10	-10	5	5	5	5	5	0	10	-5	5	10	0	0	5	5
5	10	-10	5	0	0	0	5	-5	5	0	5	0	10	-5	5	5	5	0	5	-5	5	-5	0	5	5	0	5	5
6	10	5	10	5	10	10	5	-5	0	5	10	-5	10	-5	5	0	5	5	0	-10	10	5	5	10	10	0	10	5
7	10	0	10	10	0	5	10	-5	-5	-10	0	-5	10	-10	10	5	10	5	0	0	10	-5	0	10	5	5	0	10
8	10	-5	0	5	0	10	10	-5	5	0	10	0	10	-5	10	5	10	5	0	-5	10	0	5	5	0	5	5	5
9	10	-10	10	10	5	0	10	-5	5	0	5	0	10	-10	0	10	5	-5	10	-10	10	0	10	5	10	0	10	-5
10	10	-5	5	5	10	5	10	-5	0	-5	5	-5	5	-5	5	5	5	5	0	-5	5	0	10	5	5	0	5	0
11	10	-5	10	0	-5	10	10	-5	10	-5	10	-10	10	-10	0	5	-5	10	10	5	10	-10	0	10	5	0	5	-5
12	10	0	5	10	10	5	10	0	-10	-5	5	0	10	-10	0	10	10	5	-5	-5	10	0	5	5	5	-5	10	10
13	10	0	10	5	5	0	5	-5	0	0	0	0	5	-5	0	5	5	5	0	0	5	0	0	5	5	0	0	5
14	10	0	5	0	0	10	10	-5	10	-5	10	-5	10	-5	0	5	5	0	5	0	10	-10	5	5	0	-5	5	0
15	10	-5	10	10	5	0	10	0	0	-5	10	0	10	-10	10	10	5	5	0	0	10	5	0	10	5	5	0	10
Totaal	150	-60	100	95	70	80	120	-55	25	-30	85	-30	140	-110	65	95	90	55	55	-45	135	-15	55	110	65	25	65	75

Context Questions

ID	Scenario 1						Scenario 2						Scenario 3								Scenario 4							
	Health		Enjoyment		Social Acceptance		Health		Wealth		Career		Health		Enjoyment		Social Acceptance		Wealth		Health		Enjoyment		Safety		Comfort	
	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2
1	10	-10	0	5	-10	10	10	-10	0	-5	5	0	10	-10	10	0	-5	10	5	-10	10	0	-5	10	-5	5	-10	10
2	10	-10	-5	10	5	10	-5	0	-5	0	-5	-5	10	-5	5	10	10	0	-10	10	0	0	-5	10	-5	10	-5	10
3	0	10	0	10	0	10	-5	0	-5	0	0	5	5	0	5	10	5	10	-5	5	-5	10	-10	10	-10	5	-10	10
4	5	-5	0	10	0	10	0	-5	5	0	5	0	5	-10	5	5	10	5	0	0	5	0	5	10	0	0	0	5
5	0	-5	-5	5	-5	5	5	-5	0	0	-5	0	5	-5	5	5	0	0	-5	-5	-5	5	-10	10	-5	5	-5	5
6	5	5	-5	5	-5	10	5	-5	-10	0	-5	5	10	-5	5	5	5	5	0	-10	0	5	-5	10	-10	5	-5	5
7	10	-5	-5	10	-5	10	10	-10	5	-5	0	-10	10	-10	10	5	5	10	0	0	0	0	-10	10	-10	5	-5	10
8	10	-10	0	5	-5	10	0	0	0	-5	-5	0	10	-5	5	5	5	5	-5	-5	0	5	0	5	0	0	0	5
9	5	-10	-5	10	0	10	-5	-5	5	0	10	0	5	-5	0	10	0	5	10	-5	10	-5	10	5	10	0	10	0
10	10	-5	0	5	5	5	5	-5	5	-5	5	-5	5	-5	0	5	5	-5	-5	-5	0	0	-10	10	-5	0	-5	5
11	10	-5	-5	5	-10	10	5	-5	5	-5	5	-10	10	-10	5	5	0	10	-5	-5	-5	5	0	10	0	5	-10	10
12	0	5	-5	10	-5	10	5	5	0	-5	10	5	0	-10	-5	10	0	5	-5	5	10	0	-5	10	5	-5	-5	5
13	10	0	5	5	5	5	5	-5	0	0	-5	0	5	0	0	5	5	5	0	5	0	0	-5	5	-5	0	0	5
14	10	-5	-5	0	-10	10	10	-5	5	-10	5	-10	5	-5	0	5	5	5	0	-5	0	5	0	5	-5	10	0	10
15	10	-5	0	10	-5	10	5	0	0	0	5	5	10	-10	10	0	5	5	-5	5	-5	10	-10	10	-10	10	-10	10
Totaal	105	-55	-35	105	-45	135	50	-55	10	-40	25	-20	105	-95	60	85	55	85	-30	-20	15	40	-60	130	-55	55	-60	105

= Scenario 1: Drinking choice
= Scenario 2: Bedtime
= Scenario 3: Food option
= Scenario 4: Activity
o1 = Socially perceived "healthier" option
o2 = Socially perceived "unhealthier" option

Figure 7: Raw data from the experiment. Each colored column represents one of the misalignment scenarios considered for the experiment. The top table contains the answers given to the base questions. The bottom table contains the answers given to the context questions

F SUS Interpretation Rules

Table 5: Guideline on the interpretation of the SUS scores.

SUS Score	Grade	Adjective Rating
>80.3	A	Excellent
68 – 80.3	B	Good
68	C	Okay
51 – 68	D	Poor
<51	F	Awful