Accuracy of Textual Interfaces using Comparative Questions to Elicit Personal Value-Related Information from the Users for Building Responsible AI

Beatrice Andreea Vizuroiu[1]

Supervisor(s): Catholijn M. 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 24, 2023

Name of the student: Beatrice Andreea Vizuroiu
Final project course: CSE3000 Research Project
Thesis committee: Catholijn M. Jonker, Pei-Yu Chen, Stephanie Wehner

Abstract

Individuals seeking a healthier lifestyle can benefit from behavior support agents. Customization and transparency are crucial for system effectiveness. This paper proposes using behavior trees as a user model, with a conversational agent extracting necessary information. The conversational interface enhances transparency, allowing users to understand how the system perceives them. Understanding comparative questions is vital to this approach's success. The objective is to investigate modeling personal values accurately using a conversational agent. Technologically literate participants engaged in iterative dialogue to elicit a personalized user model. Scenarios explored contextual factors' impact on value alignment. Results revealed decreased accuracy when more values were affected by contextual factors. Comparative questions were less effective than isolated questioning. System usability was rated poor but approaching acceptability. Larger sample sizes are needed for comprehensive conclusions. This research lays the foundation for conversational agents that model personal values within behavior trees, advancing behavior support systems.

1 Introduction

Behavior support applications have not only gained popularity, but have been introduced to our daily lives extensively [1]. They are software or digital tools designed to assist individuals in modifying their behaviors to achieve specific goals or improve their well-being. These applications typically employ various strategies such as goal setting, tracking progress, providing feedback, and offering personalized interventions or guidance. They aim to support users in adopting positive habits, breaking unhealthy patterns, managing stress, promoting physical activity, enhancing mental well-being, or addressing specific behavioral challenges. However, to be most effective, these agents should support their users differently based on personal values [2]. Values can be described as intangible drivers that influence the way we form opinions and carry out actions [3].

The focus of the behaviour support agent we are designing is supporting the user in building a healthier lifestyle by creating habits in their routine. Its goal is to suggest activities to be performed daily in the user's free time and keeping track of their performance status (i.e. 'go to the gym' or 'run in the park' instead of 'watch a movie'). As Dr. M. Maltz described in his book Psycho-Cybernetics [4], it takes 21 days to create a habit and 90 days to make it part of one's lifestyle. Moreover, context needs to be closely taken into consideration alongside one's personal values [5] as it would be unwise to suggest going for a run on a rainy day if i.e. the user values comfort more than exercising daily. Therefore, to gain access to the personal values of the current user, interaction is necessary. The two would interact through the means of a textual interface that will allow the agent to elicit personal values from the user through conversations.

One method employed by the agent to elicit these personal values is through the use of comparative questions. Comparative questioning involves presenting individuals with sets of options or values and asking them to rank or compare them based on their preferences or importance. It helps to gather information about relative preferences and priorities between different choices or attributes. This approach allows for a deeper understanding of individuals' subjective evaluations and helps in capturing nuanced differences in their preferences. Comparative questioning is commonly used in surveys, interviews, and user studies to elicit valuable insights and aid in decision-making processes.

The aim of this paper is to answer the following research question: *"How accurate is comparative questioning in eliciting personal value-related information through textual interfaces?"*. Therefore, it presents the accuracy of eliciting personal value-related information modeled in behaviour trees from users through comparative questions. Specifically for this paper, the interaction between the user and the agent was investigated through a textual interface. The comparative questions were used to identify which values change in different contexts compared to times at which no context is involved. Based on the answers to these questions, the agent could build an underlying user model represented by behaviour trees which could be used further. This paper also addresses a related research question: *"How usable is the system that employs comparative questioning through a textual interface?"*. Its objective is to assess the usability of the system.

To accomplish this task, a user study has been held with 15 participants as well as five separate participants who tested all five interfaces included in the parent pilot. The methodology of this user study is further explained in section 3

The report is structured as follows. In section 2, the background and motivation of this study are introduced. A detailed explanation of the methodology can be found in section 3, which includes further information about the textual interface, question types, and construction process of the underlying user model, as well as the evaluation methods used. Section 4 presents the main numerical results. Ethical implications are discussed in section 5, followed by a thorough examination of limitations and alternative interfaces included in the parent study as well as future work deemed necessary which can be found in section 6. Finally, section 7 provides the conclusion of this paper.

2 Background and Motivation

This section aims to provide background and motivation for the assumptions underlying this paper.

The first assumption made is that it is desirable to model the user's behavior and preferences into user models such as behavior trees. The leaves of those behavior trees are represented by values. Values can be described as intangible drivers that influence the way we form opinions and carry out actions [3], therefore, they constitute the standards employed by individuals to make decisions and assess both individuals and occurrences [6], [7]. Furthermore, values are now being explicitly embedded in technology itself. This can be observed in various ways, such as raising awareness about individuals' environmental actions [8], resolving conflicts between norms [9], or determining behavior plans for artificial agents [10].

For run-time personalization within a system, it is crucial to establish a connection between values and the available options for behavior [11].

Previous studies such as [12] and [13], have examined a framework that captures the hierarchical nature of activities as individuals conceptualize them and explores how values can be integrated into this framework. By linking values to activities in the hierarchy, the study by Pasotti et al. [13] determines the impact of specific choices on different values. Building upon this foundation, the formal user model for visually impaired travelers was developed in the study by Berkaa et al. [11]. The findings of the [11] study emphasize the importance of establishing a mutual understanding between the system and the user throughout the conversational process when eliciting knowledge for formal user models using a conversational interface. It is this motivation that drives the current research paper, aiming to involve the user more actively in the process to address the challenges identified in the [11] study.

3 Methodology

To assess the accuracy of eliciting personal values modelled in behavior trees with comparative questions through textual interfaces, a user study was conducted with 15 participants. This study aimed to investigate the interaction between the agent and the user when a scheduled activity could not be performed due to external factors (specifically, the misalignment reason within the scenario). Understanding the accuracy of this system is crucial, and the user study played a vital role in achieving this goal.

3.1 Scenarios Used

The scenarios used for this user study are represented by a goal, a misalignment reason (context and alternative choice) and positively affected values. In the context of behavior support applications or systems, a misalignment reason refers to a factor or cause that leads to a misalignment or discrepancy between the intended behavior or actions of the system and the desired behavior or goals of the user. It represents the underlying reason or explanation for why the system fails to accurately understand, interpret, or respond to the user's values, preferences, or objectives. For all scenarios it is assumed that by choosing the alternative option, the user is performing an unhealthier activity, therefore health (which is a value) is affected negatively.

Scenario 1:

Goal: Increase water intake

Misalignment Reason:

Context: Party

Alternative Choice: Soda

Values Affected: Enjoyment, Social Acceptance

Scenario 2:

Goal: Run 3 km daily

Misalignment Reason:

Context: Bad weather (rain, snow, thunderstorm, canicular days)

Alternative Choice: Watch movie at home

Values Affected: Enjoyment, Safety

Scenario 3:

Goal: Maintain a more nutritious diet

Misalignment Reason:

Context: Dining in the company of others at a restaurant

Alternative Choice: Fast-food

Values Affected: Enjoyment, Social Acceptance, Wealth

Scenario 4:

Goal: Improve your sleep schedule

Misalignment Reason:

Context: A crucial business meeting is forthcoming

Alternative Choice: Work until the late hours

Values Affected: Career, Wealth

3.2 Textual Interface as Chatbot & Construction of the User Model

The textual interface used for this study was implemented through the Landbot[1] platform, using a chatbot. The choice to utilize this platform was based on its suitability for the program's objectives. The platform enables the creation of a text-based bot and facilitates immediate human intervention if required. Additionally, it allowed participants to interact with the agent in a genuine textual interface, as they were asked to provide their answers by typing them in. This setup aimed to simulate the expected behavior of the agent in the application and served as the focus of the testing. Since the interface used in this experiment was text-based, the agent was configured as a chatbot, using the Landbot platform, to facilitate the study.

The interface design incorporated comparative questions, an example of which is illustrated in Figure 1. These comparative questions were strategically crafted to provide insights into the values associated with each scenario and the two alternatives being considered. To construct the user model and assign importance to each value, the agent attributed a value ranging from -10 to 10, with a step size of 5, to each possible answer depicted in figure 1. This approach facilitated the incorporation of user preferences and allowed for a quantified assessment of the significance associated with each value. To account for the contextual aspect, the agent would first present the context to the user and then proceed to ask the same set of comparative questions while the context applied. The only distinction in these questions, as shown in Figure 1, would be the replacement of the phrase "in general" with "in the context of the _", where the blank space would be filled

[1]Landbot - Chatbot platform: https://landbot.io/chatbot-for-website

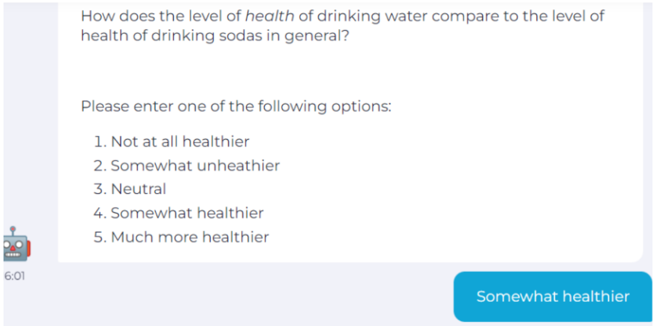

Figure 1: Textual interface used during experimental phase: Example of comparative question in general case

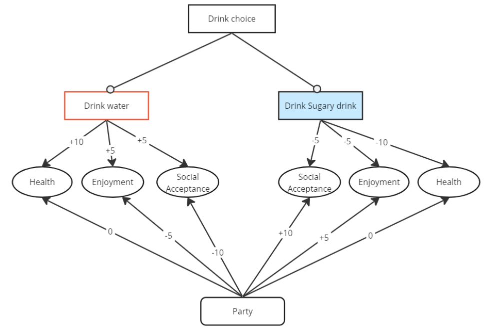

Figure 2: Example of Values Modelled in a Behaviour Tree

with one of the specific contexts outlined in Section 3.1. This approach allowed for a tailored exploration of values within different contexts. To ensure that the final weights of the values fall within the range of -10 to 10, with a step size of 5, the weights of the context edges in the behavior trees have been calculated as follows:
Let us consider the example where the answer to the question in Figure 1 was "3. Neural" and the answer to the same question while the context applied was "5. Extremely healthier". In this case, the difference in levels between the two answers is 2 (going from level 3 to level 5), with each level representing a step.
To calculate the context edge weight, we multiply the difference in levels (2 steps) by the weight per step (5), taking into account the direction of the transition. If the transition is from a lower level to a higher level, it indicates a positive weight. Conversely, if the transition is from a higher level to a lower level, it indicates a negative weight.
In the given example, the transition is going from 3 to 5, representing a positive weight. Thus, the context edge weight would be +10, as it is 2 steps above the general case. Conversely, if the transition went from 5 to 3, it would be 2 steps below, resulting in a context edge weight of -10.
Considering the scenario of the "health" value, the final weight is calculated by combining the weight associated with the general case (which is 0 in this example, representing the "Neutral" answer) with the context edge weight. In this specific case, the final weight of the "health" value would be +10. By following this approach, the weights on the context edges can be accurately calculated to ensure they fall within the desired range.

This approach allows for contextual weighting of values systematically which constructs a behaviour tree of the kind depicted in figure 2, facilitating a nuanced analysis. The primary objective of this study is to assess the perceived accuracy of the behavior tree by constructing the user model after engaging in dialogue with the agent. Specifically, the purpose of creating the behavior tree following the dialogue is to enhance the agent's understanding of the user's values and enable effective planning of activities to achieve the user's goals within a suitable timeframe.

3.3 Participants

A total of 15 participants were recruited through personal networks, ensuring that no positions of power were involved. The selected participants were technologically literate and represented a diverse range of genders, with ages ranging from 18 to 65.

3.4 Procedure

The participants were presented with four distinct scenarios and were asked to answer a series of comparative questions related to each scenario.

For each scenario, two sets of questions were administered. The first set aimed to establish a general ranking of the user's values, allowing the agent to understand what the user considers important in the absence of specific context. The second set of comparative questions was designed to identify any disparities between the general ranking of values and the ranking assigned when a context was provided. These two sets of questions provided valuable insights into the user's value hierarchy and how it may shift based on different contexts.

In this user study, a Wizard of Oz approach was employed to simulate the creation of the behavior tree by the agent. The behavior tree was manually constructed to investigate the accuracy of user modeling within this setup. To construct a behavior tree that represents the user's underlying model, the answers provided by the user to both sets of questions were used. To test the perceived accuracy of these constructed behavior trees, they were presented to the user. The researcher explained the behavior tree to the user and asked whether it accurately reflected their values in each given scenario. This step allowed for validation and verification of the behavior trees by directly involving the user in assessing their representation of their own values within each scenario.

3.5 Evaluation Method & Data Analysis

After conducting the experiment and collecting data from user interactions with the system, as well as manually assessing the perceived accuracy of the user model generated by the system, it is necessary to analyze the collected data and summarize the results. This analysis aims to evaluate the effectiveness of a system that employs a textual interface with comparative questions to model values in behavior trees.

This study focused on the accuracy and usability of a textual interface that uses comparative questioning. The alternative to this type of questioning is in isolation questions. In isolation questioning refers to a method or approach where individuals are asked to provide their opinions, preferences, or judgments on a particular topic or set of options without any comparative or contextual reference. It involves presenting choices or scenarios to individuals and asking them to evaluate or rank each option independently, without considering or comparing them to other alternatives. It is important to note that the alternative questioning type, in isolation questioning, can be valuable for data analysis purposes.

Accuracy of the Behavior Tree

To measure the perceived accuracy of the behavior tree both Hamming Distance and Signed Changed Weight measures have been employed.

1. Hamming Distance (HD)

In this study, accurately measuring the system's performance is challenging since it deals with subjective aspects such as values and their perceived importance. As a result, the assessment of the system's accuracy is primarily based on the perceived accuracy rather than objective measurements. Therefore, to assess the perceived accuracy of the user model, the disparity between the constructed behavior trees and the user tweaked trees was measured. To evaluate the dissimilarity between the generated user model and the corrected version, the HD was employed. This choice was motivated by the HD's simplicity, intuitiveness, and ease of computation. Moreover, the HD has shown effectiveness in assessing the significance of individual variables in variable importance analysis. By leveraging its ability to measure the impact of variables on overall similarity or dissimilarity, the HD served as a suitable metric for evaluating the differences between objects or samples in this context [14]. By calculating the distance between these two behavior trees, the extent of accuracy in modeling the user's values and preferences could be determined. This analysis provided valuable insights into the effectiveness of the agent's user modeling capabilities due to the active involvement of the user in comparing the behavior trees. This aspect holds significant importance since there is no definitive ground truth available regarding the user's perceived reality. By actively engaging the user in the evaluation process, the analysis gains credibility and provides meaningful insights into the accuracy and alignment of the behavior trees with the user's subjective perception.

After a thorough analysis of the discrepancies in the collected data, it can be deduced that users with an odd HD would have benefited from isolation questioning. This observation arises from the fact that their changed value was limited to one side of the tree. On the other hand, users with an even HD might benefit from an alternative type of interface. The textual interface, being extensive and demanding the reader's attention for questions and details, may have contributed to this finding. However, it is also possible that the formulation of the questions themselves requires refinement.

2. Signed Changed Weight

To incorporate both the magnitude and direction of the change, the signed difference between the two values is calculated. This approach considers the algebraic sign of the difference, indicating whether the change is positive or negative. By considering the direction of the change, the assessment takes into account not only the magnitude but also the specific nature of the change. This allows for a more comprehensive understanding of the weight and impact of the change in the evaluation process.

Insights and Interpretation per Scenario Affected

Scenario 1 & 4: 1 subject manually changed the resulted behavior trees for these scenarios. This indicates that the majority of subjects were able to accurately navigate and comprehend the system's interface for these scenarios and the agent was able to correctly compute the user model.

Scenario 2: The system displayed the lowest effectiveness for this scenario, where 4 out of the 7 subjects made at least one mistake. Interestingly, due to the relatively small signed difference (0 or +5), this finding suggests that users with an even HD tend to exhibit impatience and require a more concise interface to avoid missing important details. Consequently, the accuracy of the user model creation for this sample size (15) in the second scenario stands at 73%.

Scenario 3: 3 subjects changed the behavior trees out of which 2 exhibited an odd HD with a low signed difference, indicating that isolation questioning could have been a more suitable approach for them. On the other hand, the remaining subject simply did not read or comprehend the agent's inquiries, as confirmed during the manual check session.

Usability

To assess whether the tested behavior of the agent aligns with user expectations in terms of a behavioral agent utilizing comparative questions and textual interfaces, a system usability survey (SUS) was employed. It consisted of 10 questions, which can be found in its entirety in Appendix B. This survey was chosen for its reliability even with small sample sizes and its ability to differentiate between usable and unusable systems [15]. The survey served as an effective tool to gauge user satisfaction and evaluate the overall usability of the system.

According to [16], the SUS provides a single numerical value that represents a comprehensive measure of the overall usability of the system under investigation. It's important to note that individual scores for each item are not meaningful by themselves. To calculate the SUS score, the score contributions from each item are summed. Each item's score contribution can range from 0 to 4. For items 1, 3, 5, 7, and 9, the score contribution is obtained by subtracting 5 from the scale position (referred to as Scale A). Conversely, for items 2, 4, 6, 8, and 10, the contribution is derived by subtracting the scale position from 25 (referred to as Scale B). The sum of the scores is then multiplied by 2.5 to obtain the overall SUS value. SUS scores can vary from 0 to 100, representing the full range of possible scores.

To evaluate the usability of the system based on all the data points, the average score has been considered. Taking the average score allows for a comprehensive assessment by considering the collective experience of all users. This approach goes beyond individual scores alone and provides a more holistic evaluation of the system's usability. Furthermore, using the average score enhances the reliability of the assessment. By averaging multiple scores, the impact of outliers or extreme scores is mitigated, resulting in a more reliable overall assessment of usability [17].

4 Results

This section delves into the intriguing details of the data gathered during the experiment, providing deeper insights into the findings.

4.1 Accuracy

Out of the 15 participants who participated in the experiment, it was observed that 7 of them had mistakes in the generated behavior tree. This suggests an initial accuracy rate of 53% in accurately constructing the user model.
Out of the 7 subjects identified with mistakes, only 2 of them exhibited errors in more than one scenario, specifically in 2 scenarios.
A similar pattern emerged for the third scenario, with 3 out of the 7 subjects found to have made mistakes. This suggests an accuracy rate of exactly 80% for this particular scenario.
In contrast, only one subject indicated misalignment for the remaining scenarios, resulting in a high accuracy rate of 93%.

Scenario 1	Scenario 2	Scenario 3	Scenario 4
93%	73%	80%	93%

Table 1: Accuracy of User Model per Scenario

<table>
<tr><td colspan="6">Scenario 1</td><td colspan="6">Scenario 2</td></tr>
<tr><td colspan="3">HD</td><td colspan="3">AWC</td><td colspan="3">HD</td><td colspan="3">AWC</td></tr>
<tr><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td></tr>
<tr><td>1</td><td>0.06</td><td>0.25</td><td>15</td><td>1</td><td>3.74</td><td>5</td><td>0.33</td><td>0.6</td><td>75</td><td>5</td><td>7.75</td></tr>
<tr><td colspan="6">Scenario 3</td><td colspan="6">Scenario 4</td></tr>
<tr><td colspan="3">HD</td><td colspan="3">AWC</td><td colspan="3">HD</td><td colspan="3">AWC</td></tr>
<tr><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td><td>SA</td><td>M</td><td>SD</td></tr>
<tr><td>5</td><td>0.33</td><td>0.7</td><td>45</td><td>3</td><td>4.76</td><td>1</td><td>0.06</td><td>0.25</td><td>10</td><td>0.66</td><td>2.49</td></tr>
</table>

Table 2: Summary of Scenarios
HD = Hamming Distance
AWC = Absolute Weight Change
SA = Sum of All
M = Mean
SD = Standard Deviation

All raw results can be found in Appendix D.

4.2 Usability

Table 3 presents the 15 data points along with their corresponding scales mentioned above and the computed usability scores. The table demonstrates the variability of opinions regarding the system's usability. It is important to note that among the 15 participants, the highest reported usability score is 100%, while the lowest is 52.5%.

The average score of 78%, indicates that the system used in this study is of good usability, under the acceptable threshold, which with little improvement would be above the excellent threshold. Therefore in the current state, this system would be acceptable as a deployed system according to [18].

Scale A	Scale B	Usability = (Scale A + Scale B) * 2.5
8	18	65%
9	21	75%
6	17	57.5%
8	19	67.5%
8	22	75%
12	25	92.5%
15	25	100%
9	21	75%
13	25	95%
12	25	92.5%
12	23	87.5%
5	16	52.5%
12	22	85%
10	22	80%
8	20	70%

Table 3: Individual Usability Scores

4.3 Common Participants

As the parent study involves five conditions, it is important to recruit a handful of participants to test all conditions as well. The following five conditions are included in the parent study:

A) Textual Interface and Comparative Questioning Context (this study)
B) Textual Interface and in Isolation Questioning Context
C) Graphical Interface and Comparative Questioning Context
D) Graphical Interface and in Isolation Questioning Context
E) Audio Interface and in Isolation Questioning Context

Five additional participants were recruited to test all five interfaces included in the pilot. These five participants were presented with five behavior trees per scenario. As each interface utilized the same scenarios (refer to subsection 3.1) and featured similar computed questions (as depicted in Figure 1), it was anticipated that the trees would exhibit considerable similarity. To compare the results, we agreed to measure the total HD (sum of HDs per scenario per participant) and the total weight change (absolute sum of manual changes and computed weight in the behavior trees) for this set of participants.
However, figure 3 clearly demonstrates that the interfaces utilizing a comparative questioning context required more manual changes in the behavior trees by the user compared to the interfaces using an isolated context. This observation is further supported by table 4, which highlights that the mean HD of the interfaces employing a comparative questioning context is higher than that of the interfaces using an isolated context. The standard deviation (SD) reinforces this finding and provides additional evidence of the same trend.

Additionally, figure 4 reveals that the absolute difference between the calculated value weight and the manually entered value is greater for interfaces employing the comparative questioning context. This is because, when constructing the user model using comparative questions, the agent assumes that the weight of an alternative activity is equal to the expected activity's weight but with reversed sign (for example, if the value of drinking water is +10 in scenario 1 from 3.1, the value of drinking soda is automatically set to -10). This trend is prominently evident in table 5, where the stark contrast between the mean and SD of the comparative interfaces and the isolation interfaces becomes apparent. The significantly higher values of the mean and SD for the comparative interfaces indicate that they necessitate a greater number of changes and require more careful consideration in terms of how the agent elicits values from the user, while ensuring that negative weights are not assigned to the alternative activity.

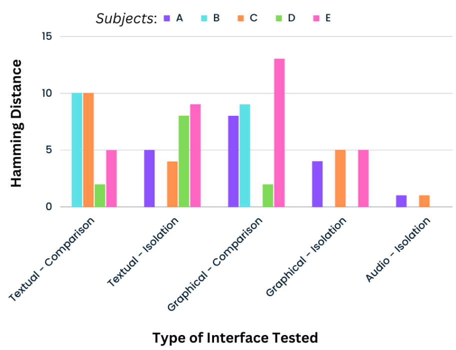

Figure 3: Hamming Distance

Comparison			Isolation		
Sample Size	Mean	SD	Sample Size	Mean	SD
10	5.9	4.46	15	2.8	3.01

Table 4: Hamming Distance Statistics in terms of Comparison and Isolation Interfaces

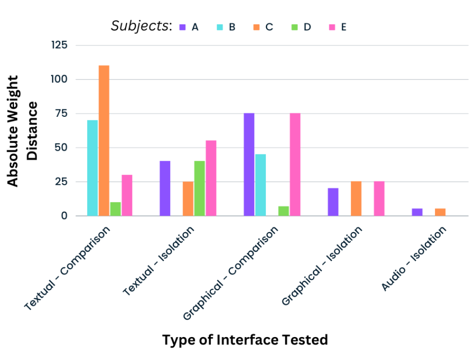

Figure 4: Absolute Weight

Comparison			Isolation		
Sample Size	Mean	SD	Sample Size	Mean	SD
10	42.2	36.76	15	16	17.72

Table 5: Absolute Weight Statistics in terms of in Comparison and in Isolation Interfaces

It is important to note that there is a difference in sample size between the two types of interfaces. The isolation interfaces were studied by three researchers, whereas the comparison interfaces were studied by only two researchers. This discrepancy in sample size should be taken into consideration when interpreting the findings.

These findings provide valuable insights for comparing all interfaces and understanding the necessary trade-offs.

4.4 Alternative Interfaces

This section aims to discuss the individual results of each condition and compare them to the findings of this study.

It is important to provide a brief analysis about the four alternative interfaces that were explored alongside the agent utilizing the comparative questioning context, along with their respective results. These interfaces were examined to gain a comprehensive understanding of the different approaches and their potential implications.

All five experiments were conducted with an individual sample size of 15. As observed from tables 6 and 7, the interface studied in this paper required the least amount of manual changes as indicated by the mean HD. However, when considering the results provided by the SD of Absolute Weight Changes (AWC), it suggests that the interface involved in condition B would perform better than the one presented in this study (as observed from the results presented in table 7). This trend is observed for the graphical interfaces (C, D) as well, where D outperforms C by a large threshold. Therefore, a hypothesis can be formed that interfaces using in isolation questioning would be more accurate than those utilizing comparative questions.

Condition D analyzes a graphical interface using an isolated questioning context, indicating that this interface is the most effective and accurate compared to the other alternatives when looking at the SDs of AWC, but not as accurate from the first try as the mean of HD is larger than for condition A. When plotting both tables 6 and 7, as in figure 5, it is clearly indicated that the previously mentioned hypothesis is wrong as it only applies for the graphical interfaces (D is better than C). However, for the textual interfaces this is not the case. Overall, this interface (A) is better than the one under condition B. This figure also indicates that the best interface overall is the one under condition D. The interface studied in this paper is very close to the one of condition D in terms of results indicating that with little improvement it could outperform condition D, but is the best one in terms of textual interfaces (better than B).

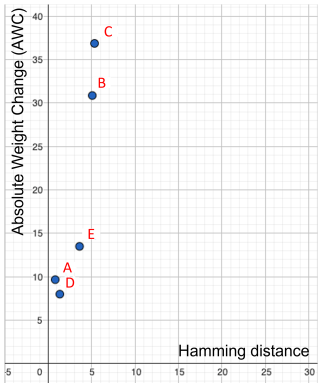

Figure 5: Plot of AWC and HD per interface

Moreover, this is sustained by the small difference in the SDs between conditions A and B, which also suggests that with further investigation, the interface presented in this study could be improved to surpass the performance of the other interfaces. This is supported by the fact that the mean number of changes for this interface is the lowest among all the alternatives, indicating that only a few values would require isolated questioning (such as health-related factors).

A			
HD		AWC	
Mean	SD	Mean	SD
0.8	0.98	9.66	14.2

Table 6: Statistics Summary of Own Condition

B				C			
HD		AWC		HD		AWC	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
5.07	1.81	30.87	13.43	5.33	3.53	36.87	21.92
D				E			
HD		AWC		HD		AWC	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
1.33	2.19	8	13.1	3.6	6.2	13.5	15.6

Table 7: Summary of Statistics for Alternative Conditions presented in section 4.3

5 Responsible Research

Responsible research practices are paramount when conducting an experiment [19]. My research involved 15 participants who were asked to test a system (agent via textual interface) and evaluate its usability. The research team ensured ethical considerations by obtaining informed consent from each participant prior to the start of the experiment. Consent forms were signed, outlining the purpose of the study, the nature of their involvement, and any potential risks or benefits. The participants were made aware of their rights, including the option to withdraw from the study at any time without consequences. To assess the usability of the system, participants were asked to complete a system usability form anonymously. Respecting participant privacy and confidentiality, the research team ensured that the collected data was securely stored and handled in accordance with relevant data protection guidelines. By adhering to these responsible research practices, the study promotes transparency, participant autonomy, and the ethical principles necessary to generate reliable and meaningful results.

In addition to the fundamental practices, conducting this pilot study requires careful consideration of the system being researched and its ethical implications as this pilot motivates the development of an AI system [20]. Privacy and data security play a critical role [21] as the agent would collect and store personal health-related data, including user responses to value-based questions. Protecting this sensitive information is essential to safeguard user privacy and prevent unauthorized access or misuse. Moreover, while the agent aims to assist users in improving their health, it is vital to respect their autonomy and preserve their ability to make independent decisions regarding their well-being. Users should have the freedom to define and prioritize their personal values without undue influence or coercion from the agent. To achieve this, it is crucial to ensure the accuracy of the behavior tree representing the user model or explore ways to allow manual adjustments. Additionally, there is a risk of imposing certain values onto users through the agent's reliance on personal value-based questions, particularly when starting with a predefined list of values. Care must be taken to maintain neutrality, respect diverse perspectives, and avoid enforcing specific values or biases unless users explicitly opt for the agent's predefined settings instead of presenting their personal initial views.

6 Discussion

The main objective of this study was to investigate the accuracy of an agent utilizing a comparative questioning context over a textual interface in modeling values in behavior trees. This section aims to discuss the limitations encountered during the study and provide a brief overview of the four alternative interfaces that were studied in parallel with this agent.

6.1 Limitations

Throughout the study, certain limitations were encountered that affected the overall findings. These limitations will be discussed in detail, shedding light on the potential impact they may have had on the results.

Lack of Ground Truth: Given the absence of a definitive ground truth and the subjective nature of 'personal values', assessing the accuracy of the user model becomes challenging. Objective measures for evaluating the creation of the user model are not readily available. Consequently, this study relies on assessing the perceived accuracy instead, considering the subjective perspectives and judgments of the participants.

Small Sample Size: Due to the time constraint, the sample size of participants for this experiment was 15, plus 5 extra participants to test all interfaces. A small sample size increases the risk of sampling error and decreases the ability to detect small but meaningful effects. It may also limit the reliability and precision of the results.

Time Constraints, Mood Changes & Bias: One limitation of this study is the time constraints that restricted the sample size to the researchers themselves for the testing of all 5 interfaces. As a result, there is a potential for bias in the results, as each researcher may have a personal stake in the system and may perceive their own interface as superior. Additionally, conducting multiple experiments in a single day may have affected participant fatigue, leading to reduced concentration and potentially influencing their responses. Moreover, conducting experiments in different days could have led to increased risk of discrepancy between two interfaces utilizing the same set of questions. The difference in mood could have influenced the resulted behavior tree and it could have suffered major modifications that would have not happened if the researcher tested the interface a day before. To mitigate these limitations, future studies should allocate more time for user studies and involve a larger and more diverse participant pool to ensure a more objective evaluation of the system's performance.

Lack of Real-world Deployment: The system was evaluated in a controlled experimental setting employing a Wizard of Oz approach, which may not reflect the complexities and challenges of real-world deployment. The system's performance and usability may differ when used in actual operational environments.

6.2 Future Work

This study aimed to assess the effectiveness of a textual interface using comparative questions to elicit personal values and model them in behavior trees. The conversational agent engaged users in two sets of comparative questions per scenario (refer to Appendix A) to construct the user model. Four scenarios were developed and presented to each participant by the agent.

It's important to note that the agent's focus in this study was on multiple activities simultaneously, rather than solely prioritizing a single main goal and investing more time in understanding the user for that particular goal.

To further enhance the efficiency and usability of such a system, it is crucial to conduct additional research that specifically investigates the agent's performance when it concentrates on a single goal instead of multiple goals. This focused investigation can provide valuable insights into the necessary improvements for the system.

The agent utilized in this study did not employ a predefined ranking of values, but instead relied on the values assigned to each scenario (refer to section 3.1). As a result, the initial phase of determining the values linked to each activity was excluded from this study. In order to correctly start identifying the user's values and classify additional ones, without overwhelming the user at the beginning, the agent could benefit from using a pre-defined list of values split according to the 10 categories S.H. Schwartz identified (figure 6) [3]. The behavior support agent would ideally rely on a set of initial values before personalization is accessible. For this matter, the list of 57 values shown in [22] could be used as it "is based on decades of evidence-based research from social psychologist, Shalom H. Schwartz who pioneered the Theory of Basic Human Values". This step would greatly assist the agent in establishing a foundational set of values for the user, without overwhelming them with excessive questions from the beginning. By utilizing the 10 categories identified by S.H. Schwartz, the agent can assign a less frequently expressed value to a category and prioritize other values within that category. This approach allows the agent to gradually construct the user's profile while considering a diverse range of values.

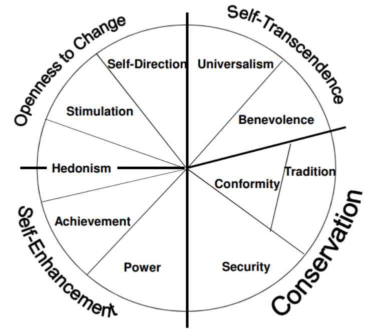

Figure 6: "Theoretical model of relations among ten motivational types of value" (Image source: Shalom H. Schwartz., An Overview of the Schwartz Theory of Basic Values., 2012.)

Additionally, it is recommended to explore the impact of a larger sample size for both the single-goal and multiple-goal conditions. A larger sample size increases the statistical power of the analysis, reducing the likelihood of Type II errors and increasing the chances of identifying significant results. This exploration can significantly contribute to enhancing the usability of the system and improving its overall effectiveness.

7 Conclusion

The primary objective of this research paper was to investigate the accuracy of comparative questioning in eliciting personal values through textual interfaces, addressing the research question: *"How accurate is comparative questioning in eliciting personal value-related information through textual interfaces?"*.

To achieve this goal, four misalignment scenarios were created and presented through the conversational agent, as detailed in Section 3.1. Subsequently, a textual interface incorporating comparative questions was developed as the system under test. The experiment involved 15 technologically literate participants aged 18-65, who engaged with the interface to assess its performance.

Upon analyzing the data, as discussed in Section 4, it was found that the accuracy of the agent in constructing user models is influenced by the number of values involved. Greater impact on values resulted in decreased accuracy in constructing the user model. However, when considering the interface in the context of multiple activities, the interface presented in this study demonstrated one of the highest accuracy among the five interfaces examined in the parent study (as presented in Section 4.4). This interface required fewer manual changes, and the observed changes were not significantly deviated from the calculated values.

In terms of usability, the system exhibited good usability. However, further investigation with a larger sample size is advised to avoid a possible Type I error. Additionally, future studies should consider the division of the agent's focus to test for single or multiple activities, as in this study the focus was on multiple activities only.

The ethical implications of such a system were outlined in section 5, emphasizing the importance of respecting user autonomy and preserving their ability to independently make decisions related to their well-being. The system should allow users the freedom to define and prioritize their personal values without undue influence or coercion from the agent.

Overall, this research provides valuable insights into the accuracy and usability of a textual interface utilizing comparative questioning for eliciting personal values. It serves as a foundation for further investigation and development of ethical systems that empower individuals in making informed decisions regarding their lifestyles and well-being.

To conclude, this research paper successfully addressed the research question of *"How accurate is comparative questioning in eliciting personal value-related information through textual interfaces?"* through a well-designed experiment and a thorough analysis of its results.

References

[1] I. Kola, "Enabling social situation awareness in support agents," 2022.

[2] M. B. Van Riemsdijk, C. M. Jonker, and V. Lesser, "Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges." *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, pp. 1201–1206, 2015.

[3] H. S. Shalom, "An overview of the schwartz theory of basic values." *Online readings in Psychology and Culture 2, 1 (2012)*, pp. 1–20, 2012.

[4] M. Maltz, *Psycho-Cybernetics*. Prentice-Hall, 1960.

[5] M. B. Tielman, C. M. Jonker, and M. B. van Riemsdijk, "What should i do? deriving norms from actions, values and context," *MRC@ IJCAI*, pp. 35–44, 2018.

[6] B. Friedman, P. H. K. Jr., and A. Borning, "Human-computer interaction and management information systems: Foundations advances in management information systems, volume 5 (advances in management information systems)," in *Proceedings of the 5th Advances in Management Information Systems*, P. Zhang and D. Galletta, Eds. M.E. Sharpe, 2006, pp. 348–372.

[7] I. van de Poel, "Handbook of ethics, values and technological design," in *Proceedings of the Handbook of Ethics, Values and Technological Design*, J. e. a. van den Hoven, Ed. Springer, 2015, pp. 89–115.

[8] H. Haller, V.-B. Nguyen, G. Debizet, Y. Laurillau, J. Coutaz, and G. Calvary, "Energy consumption in smarthome: Persuasive interaction respecting user's values," in *Proceedings of the 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 2017.

[9] A. Kayal, W.-P. Brinkman, M. A. Neerincx, and M. B. v. Riemsdijk, "Automatic resolution of normative conflicts in supportive technology based on user values," *ACM Transactions on Internet Technology*, vol. 18, no. 4, pp. 41:1–41:21, May 2018.

[10] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum, "No pizza for you: Value-based plan selection in bdi agents," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

[11] J. Berka, J. Balata, C. M. Jonker, Z. Mikovec, M. B. van Riemsdijk, and M. L. Tielman, "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*, vol. 38, no. 18-20, pp. 1909–1925, 2022.

[12] P. Pasotti, M. B. van Riemsdijk, and C. M. Jonker, "Representing human habits: towards a habit support agent," in *European Conference on Artificial Intelligence*, 2016.

[13] P. Pasotti, C. M. Jonker, and M. B. van Riemsdijk, "Action identification hierarchies for behaviour support agents," in *Workshop on Cognitive Knowledge Acquisition and Applications*, 2017.

[14] "7 important distance metrics every data scientist should know," accessed on June 4, 2023. [Online]. Available: https://tinyurl.com/Distance-Metrics

[15] "Benefits of using a sus," may 12, 2023. [Online]. Available: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

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

[17] "Chat gpt," june 13, 2023. [Online]. Available: https://chat.openai.com/

[18] "System usability scale," june 13, 2023. [Online]. Available: https://blog.hubspot.com/service/system-usability-scale-sus

[19] "Responsible conduct of research training," june 4, 2023. [Online]. Available: https://oir.nih.gov/sourcebook/ethical-conduct/responsible-conduct-research-training

[20] "Ai ethics: What it is and why it matters." june 4, 2023. [Online]. Available: https://www.coursera.org/articles/ai-ethics

[21] "What is data privacy and why is it important?" june 4, 2023. [Online]. Available: https://www.integrate.io/blog/what-is-data-privacy-why-is-it-important/

[22] "The theory of basic human values: the list of personal values," may 4, 2023. [Online]. Available: https://www.discoveryourvalues.com/list-of-personal-values

A All Questions Used

For all of the questions listed below, the following answer options were available:

1. Much less *value* (e.g. enjoyable)
2. Somewhat less *value* (e.g. enjoyable)
3. Neutral
4. Somewhat more *value* (e.g. enjoyable)
5. Much more *value* (e.g. enjoyable)

Scenario 1

Imagine the following scenario:

You have decided that you should drink more water and have been doing so every evening in the past week. Before making this decision, you were not hydrating enough and when you got something to drink, it was usually a soda instead.

Questions:

1. How healthier is drinking water compared to drinking sodas in general?
2. How enjoyable is drinking water compared to drinking sodas in general?
3. How socially acceptable is drinking water compared to drinking sodas in general?

Context:

Imagine the following setting for the rest of the questions:

The alternative to drinking water is to drink soda. There is a party coming up which you are going to attend. At the party there is both soda and water available. You are a huge fan of soda, therefore you choose to drink soda for the rest of the night.

1. How healthier is drinking water compared to drinking sodas in the context of the party?
2. How enjoyable is drinking water compared to drinking sodas in the context of the party?
3. How socially acceptable is drinking water compared to drinking sodas in the context of the party?

Scenario 2

Imagine the following scenario:

You have decided to start running 3 km daily in order to improve your health and strength. Before making this decision, you didn't 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.

Questions:

1. How healthier is exercising (running) daily compared to watching a movie in general?
2. How enjoyable is exercising (running) daily compared to watching a movie in general?
3. How safer is exercising (running) daily compared to watching a movie in general?
4. How comfortable is exercising (running) daily compared to watching a movie in general?

Context:

Imagine the following setting for the rest of the questions in this section:

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 dropped by a few degrees, therefore, you have decided to stay inside and watch a movie today.

1. How healthier is exercising (running) daily compared to watching a movie in the context of bad weather?
2. How enjoyable is exercising (running) daily compared to watching a movie in the context of bad weather?
3. How safer is exercising (running) daily compared to watching a movie in the context of bad weather?
4. How comfortable is exercising (running) daily compared to watching a movie in in the context of bad weather?

Scenario 3

Imagine the following scenario:

You have decided to maintain a more nutritious diet and cut off heavily processed foods such as fast food. Please remember this scenario.

Questions:

1. How healthier is maintaining a more nutritious diet compare to eating fast food in general?
2. How enjoyable is maintaining a more nutritious diet compare to eating fast food in general?
3. How socially acceptable is maintaining a more nutritious diet compare to eating fast food in general?
4. How expensive is maintaining a more nutritious diet compare to eating fast food in general?

Context:

Imagine the following setting for the rest of the questions in this section:

The alternative to maintaining a more nutritious diet is eating fast food. 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.

1. How healthier is maintaining a more nutritious diet compare to eating fast food in the context of dining out with your friends??
2. How enjoyable is maintaining a more nutritious diet compare to eating fast food in the context of dining out with your friends??
3. How socially acceptable is maintaining a more nutritious diet compare to eating fast food in the context of dining out with your friends??
4. How expensive is maintaining a more nutritious diet compare to eating fast food in the context of dining out with your friends??

Scenario 4

Imagine the following scenario:

You have decided to improve your sleeping schedule and for the past 2 weeks have been going to bed before 10:30PM. Before making this decision, you used to scroll through social media/work until 2AM. Please remember this scenario.

Questions:

1. How healthier is sleeping early compared to staying up late in general?
2. How impactful is sleeping early compared to the staying up late on your wealth in general? By wealth I mean whether you think your bedtime has an effect on your performance at work, therefore on your salary.
3. How impactful is sleeping early compared to the staying up late on your career in general?

Context:

Imagine the following context:

The alternative to sleeping early is working late. There is a very important business meeting approaching which you need to make sure to prepare. In order to get the work done in time, you could stay up late and work or go to sleep early and try to finish it the next day, risking missing the deadline. You decide to work late today.

1. How healthier is sleeping early compared to staying up late in the context of the important business meeting?
2. How impactful is sleeping early compared to the staying up late on your wealth in the context of the important business meeting? By wealth I mean whether you think your bedtime has an effect on your performance at work, therefore on your salary.
3. How impactful is sleeping early compared to the staying up late on your career in the context of the important business meeting?

B System Usability Survey

System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

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

Figure 7: System Usability Survey Used in the Study

C Chat GPT Prompts & Answers

Question sent: Why is it a good idea to take average of usability scores to determine whether system is usable or not?

Answer: Taking the average of usability scores is a common practice to determine the overall usability of a system. Here are some reasons why it is a good idea:
Comprehensive assessment: By calculating the average usability score, you consider the collective experience of multiple users. It provides a more comprehensive assessment of the system's usability rather than relying on individual scores alone.
Smoothing out individual variations: Each user may have unique preferences, expectations, and experiences, which can introduce variability in their usability scores. Averaging the scores helps to smooth out these individual variations and provides a more generalized perspective.
Increased reliability: Averaging multiple scores increases the reliability and stability of the measurement. It reduces the impact of outliers or extreme scores that may skew the overall assessment. By considering a larger sample of scores, the average value becomes a more reliable indicator of the system's usability.
Simplified interpretation: A single average usability score provides a concise and easily interpretable summary of the overall usability. It simplifies the process of understanding and communicating the usability evaluation results, making it easier to compare different systems or track changes over time.
Basis for comparison: The average usability score serves as a basis for comparing different systems or versions of the same system. It allows you to assess whether a particular system's usability meets a predefined threshold or to compare the usability of different systems to identify which one performs better.
While the average usability score is a useful metric, it is important to consider other factors and conduct a comprehensive usability evaluation, including qualitative assessments, user feedback, and task-based analysis, to gain a holistic understanding of the system's usability.

D Raw Data from Results

D.1 Behavior Trees

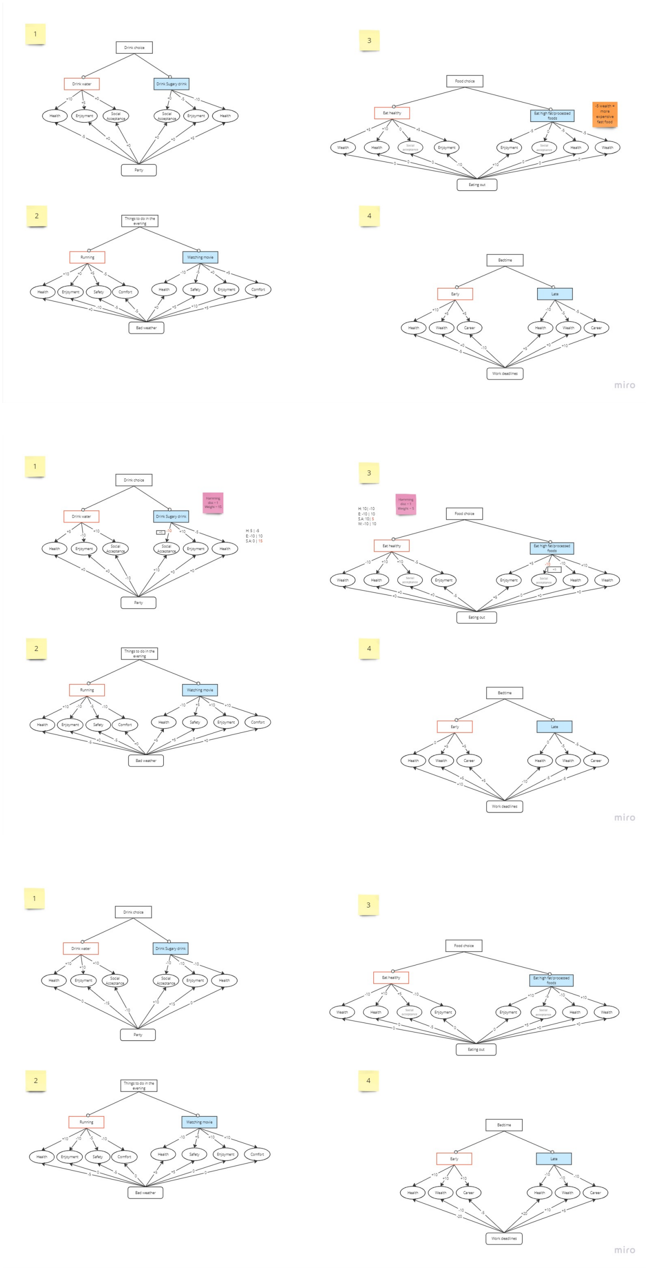

Figure 8: Calculated Behavior Trees (Part 1)

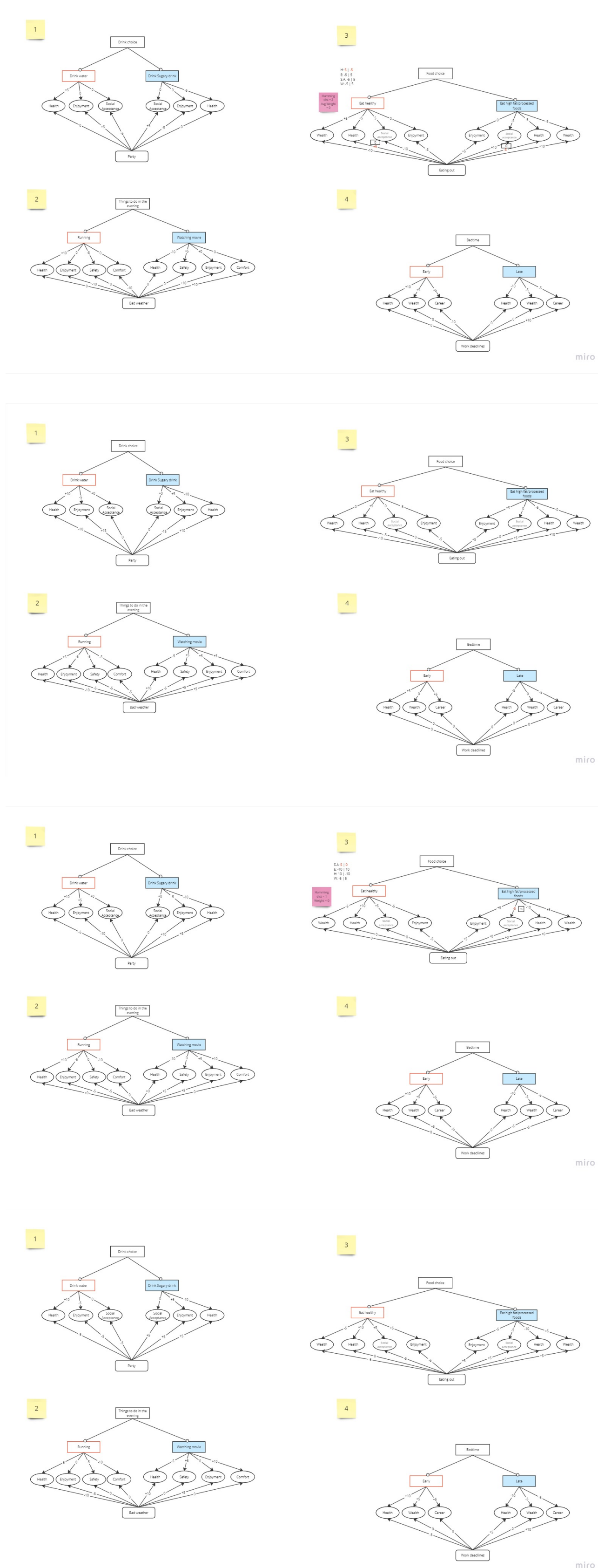

Figure 9: Calculated Behavior Trees (Part 2)

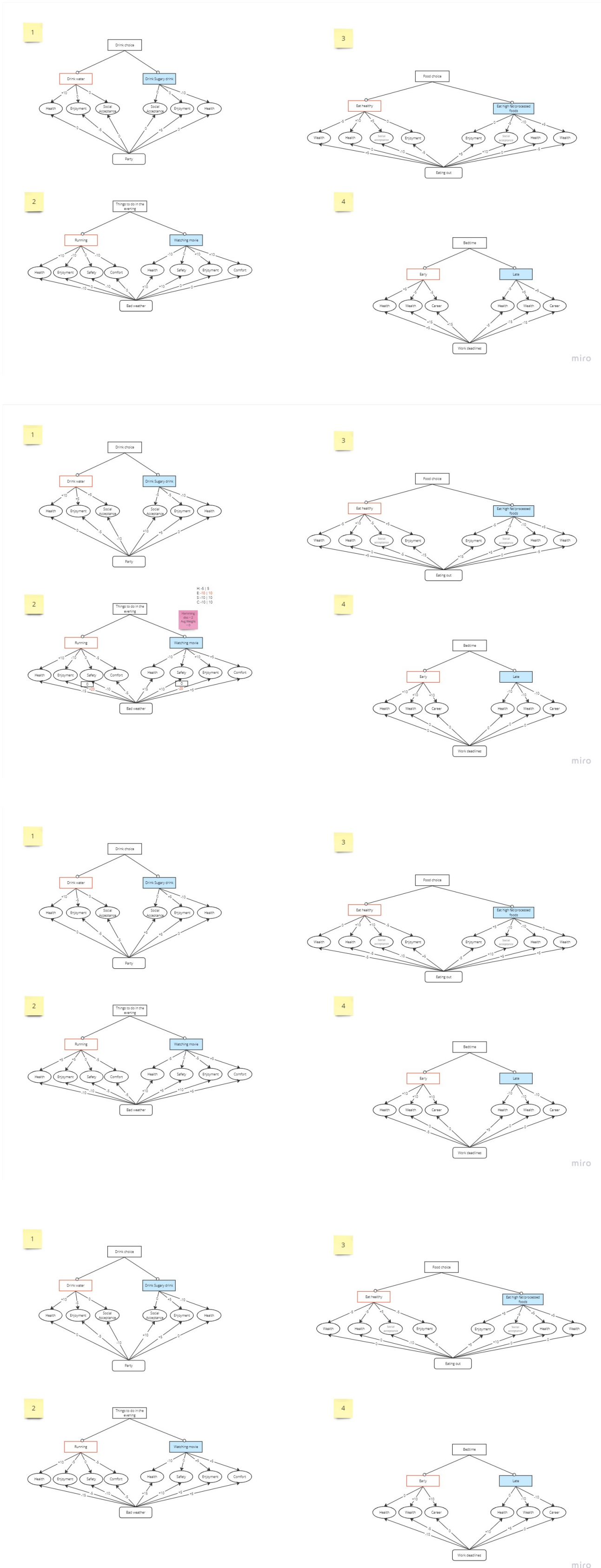

Figure 10: Calculated Behavior Trees (Part 3)

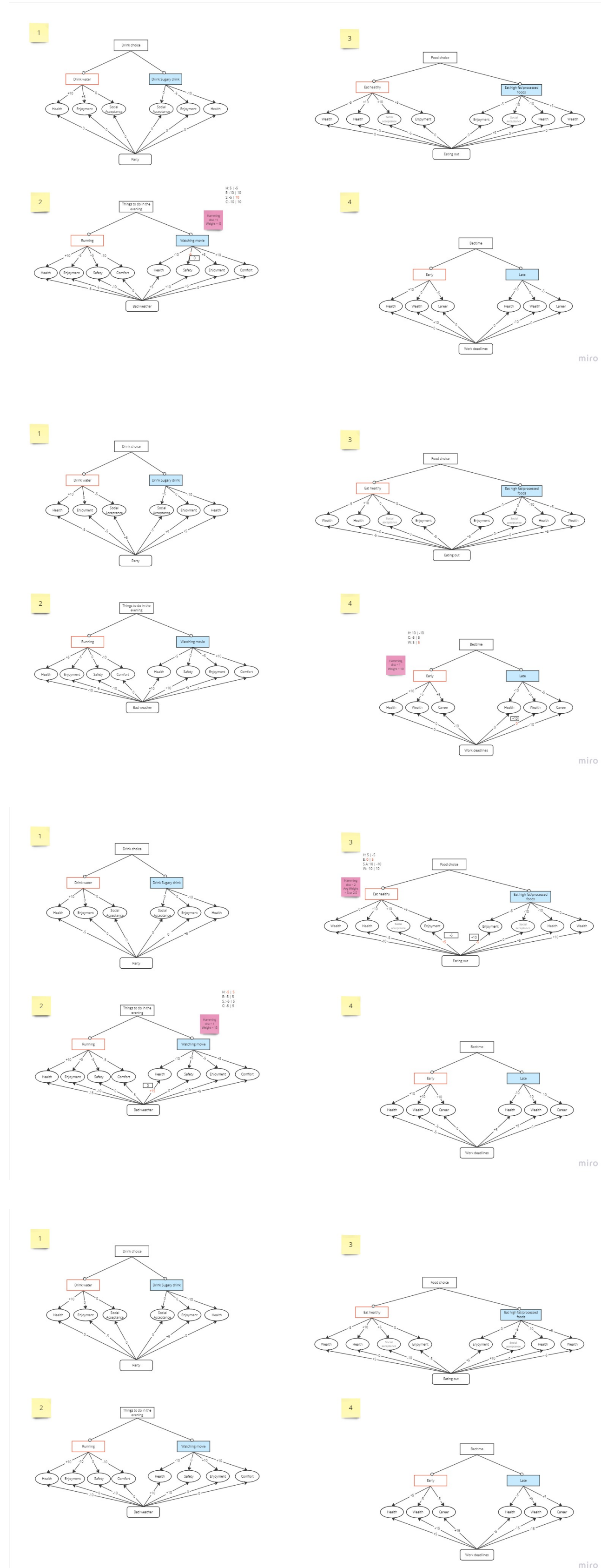

Figure 11: Calculated Behavior Trees (Part 4)

D.2 SUS

Q1	Q3	Q5	Q7	Q9	Positional scale	Scale A	Q2	Q4	Q6	Q8	Q10	Positional scale 2	Scale B	Sum = Scale A + Scale B	Sum × 2.5
2	3	2	3	3	13	12	1	1	2	2	1	7	2	14	35
3	3	3	3	2	14	13	2	0	1	1	0	4	1	14	35
0	3	2	3	3	11	10	2	1	2	2	1	8	-3	7	17.5
0	4	1	4	4	13	12	2	0	2	2	0	6	-1	11	27.5
2	3	2	3	3	13	12	1	0	1	1	0	3	2	14	35
3	4	4	4	2	17	16	0	0	0	0	0	0	5	21	52.5
4	4	4	4	4	20	19	0	0	0	0	0	0	5	24	60
2	4	2	4	2	14	13	2	0	1	1	0	4	1	14	35
4	4	4	4	2	18	17	0	0	0	0	0	0	5	22	55
4	4	3	4	2	17	16	0	0	0	0	0	0	5	21	52.5
3	4	2	4	4	17	16	0	0	1	1	0	2	3	19	47.5
0	3	1	3	3	10	9	3	0	3	3	0	9	-4	5	12.5
2	4	3	4	4	17	16	1	0	1	1	0	3	2	18	45
1	4	2	4	4	15	14	3	0	0	0	0	3	2	16	40
2	3	2	3	3	13	12	1	1	1	1	1	5	0	12	30

Table 8: Usability Scores