Designing Graphical User Interface To Elicit Personal Values

How does a graphical user interface that uses in comparison questioning influence the accuracy of user value model?

Martynas Krupskis[1]

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

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

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

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

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Behavior support applications require an understanding of the user's values to provide personalized support. User models have been developed to capture this relationship, but they may not always address unforeseen circumstances due to the changing nature of human values. To address this, the ability to update value models in real-time is crucial. This paper presents a study on the design of a graphical user interface (GUI) that utilizes behavior trees to elicit and update personal values. The study focuses on comparing the accuracy and usability of different questioning techniques, specifically comparative questions and isolation questions, and user interfaces. The results show that GUI with isolation questions and text user interface with comparative questions performed better in terms of accuracy compared to GUI with comparative questions. The study also highlights the possible impact of discrepancies in experimental procedures on the results. Furthermore, the limitations of the behavior tree model and the need for model improvements are discussed. Future work is proposed to address these limitations and explore enhancements such as alternative models, supervised processes for baseline tree building, frictionless elicitation techniques, consideration of random value shifts and big life events, and the potential of large language models. Overall, the findings contribute to the understanding of eliciting and updating user value models in personalized systems.

1 Introduction

Behavior support applications are becoming increasingly prevalent in our daily lives, providing personalized support to help individuals achieve their goals. According to [1] the growing symbiotic relationship between individuals and these systems has created a demand for personalized systems that comprehend the unique needs of their users. For example, a healthy lifestyle support agent shouldn't insist on a run if it's currently raining and the user values comfort over staying fit. To achieve this, user models have been developed to capture the relationship between the user's desired behavior and their values. [2] proposed a way to enable agents to consider personal values in decision-making and [3] proposed a framework that enables the behavior support agent to reason about a user's normative behavior and provide the support that is aligned with their values. Although user models can help adapt to users' needs, they may not always be sufficient to address unforeseen circumstances due to the changing nature of human values as outlined in [4].

Given the aforementioned, there is a need for an effective and accurate way to elicit and update user value models. However, in the current research, there is a knowledge gap in how to build a system, which could update the user's value model in run-time, instead of relying on information provided during the setup of a behavior support system.

Under this context, the study aims to bridge the knowledge gap by developing a graphical user interface that uses comparative questions to elicit and update user values. A comparative question is "a question intended to compare two or more entities and it has to mention these entities explicitly in the question". [5, p. 1] Comparing alternatives is a very simple and common framework for evaluating possible decisions. Humans are able through comparing alternatives to make a decision that maximizes positive outcomes.

In this research study, a team of five researchers is conducting experiments to explore various user interface types and questioning methods. The main emphasis of this paper is on investigating Condition *GC*. Among the five conditions considered, Conditions *GI* and *TC* are also relevant to this paper since they share similarities with the condition under study:

- **Condition GC**: Graphical user interface and comparative questions (this study)
- **Condition GI**: Graphical user interface and isolation questions
- **Condition TC**: Textual user interface and comparative questions
- **Condition TI**: Textual user interface and isolation questions
- **Condition AI**: Audio user interface and isolation questions

Using prototype built for Condition *GC*, the following research hypothesis and question are addressed:

- **Research Question** - How does a graphical user interface that uses a comparative questioning technique influence the accuracy of the user value model?
- **RH1** - Graphical User Interface (GUI) using the comparative questioning technique leads to the most accurate behavior tree model when compared to other conditions.
- **RH2** - Graphical User Interface (GUI) using the comparative questioning technique has the highest usability when compared to other conditions.

The impact of this Research Project will allow future behavior support agents to decide on the adoption of in comparison questioning technique and using a graphical user interface type. The study's results will provide insights into the perceived accuracy on the user's value model and will allow to conclude the effectiveness of comparative questioning and graphical user interface.

The rest of the paper will outline how the research question was answered. Section 2 will present the additional background information, explain the functioning of behavior tree value model, and describe different evaluation methods of the prototype and the generated tree. Then, 3 will outline a design of a graphical user interface used in the experiment and major design choices. Additionally, Section 4 presents how the comparative questions were designed and how they can be used to build a behavior tree model. Later, Section 5 presents the design and procedure of the experiment and how data will be collected. Section 6 analyzes ethical aspects of the research and presents the reproducibility of experimental methods. Section 8 offers a comprehensive analysis of the results, while Section 9 provides concluding remarks and suggests potential future improvements.

2 Proposed Approach

This section provides a methodology to address the research question. It begins with Subsection 2.1, which outlines the creation of Misalignment Scenarios. Subsection 2.2 then discusses how these scenarios can be utilized to define behavior tree models. In order to assess the accuracy of the behavior tree model, a baseline tree needs to be established, as described in Subsection 2.3. Finally, Subsection 2.4 details the use of the baseline tree for evaluating the method's accuracy, along with the employment of the System Usability Scale (SUS) to assess the interface's usability.

2.1 Misalignment scenarios

Misalignment scenarios occur when the advice provided by a behavior support agent fails to consider unforeseen circumstances or the user's individual values and context. For example, a person with a long-term goal of increasing their water intake may find themselves in a social event where they highly value social acceptance. In this context, they may choose to deviate from their goal and drink fizzy beverages instead. However, a behavior support agent that blindly optimizes for the long-term goal would continue to advise them to drink water when thirsty. Personalized and context-aware support is crucial in such situations to address misalignment and provide relevant guidance. Formalizing the scenario outlined above:

- **Long term goal:** Drinking more water
- **Behavioral Challenge:** Drink Choice
- **Possible behavior choices:** Drink water or drink sugary/sweet drinks
- **Affected value:** Social Acceptance
- **Misalignment reason:** Attending a social event

We extended the misalignment scenarios used in the experiment defined in Section 5, as real-life behavioral challenges might affect not just one, but multiple values. Below is the extended misalignment scenario mentioned earlier. Refer to Appendix A for the complete list of scenario definitions.

- **Long term goal:** Drinking more water
- **Behavioral Challenge:** Drink Choice
- **Possible behavior choices:** Drink water or drink sugary/sweet drinks
- **Affected values:** Health, Enjoyment, Social Acceptance
- **Misalignment reason:** Attending a social event (Party)

2.2 Behavior Tree Model

The behavior tree model is a formal way to store information on the misalignment reasons, values, behavioral choices, and the relationship between them. The effect of behavior choices on values can vary, ranging from positive to neutral or negative, and can be represented using weighted factors. For the context of modeling values into behavior trees the following scale was chosen:

1. -10: Very Negative
2. -5: Negative Impact
3. 0: No Impact
4. 5: Positive Impact
5. 10: Very Positive Impact

In the example scenario described in Section 2.1, Figure 1 depicts a behavior tree. The ”Drink choice” node represents the behavioral challenge of choosing a drink, with ”Drink water” and ”Drink Sugary drink” as the two available choices. An empty circle relationship indicates the options for the challenge. Additionally, the choice of drinking water directly influences the personal values of ”Health,” ”Enjoyment,” and ”Social Acceptance,” as denoted by the arrow relationships. The weights on the arrows represent the impact of behavior choices on values.

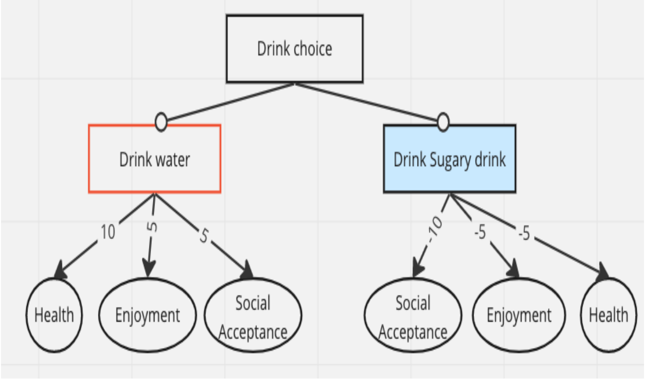


Figure 1: Behavioral challenge of drinking choice with two choice of drinking water and drinking sugary drinks. Empty circle relationship defined ”is instance of” relationship, while arrow is a ”affects” relationship. Given this, drinking water has a very positive effect on health, positive effect on enjoyment and social acceptance. While drinking sugary drinks have very negative effect on social acceptance, negative effect on enjoyment and health.

Misalignment scenarios enable the observation of value importance in specific contextual misalignments. By incorporating the misalignment reason into a behavior tree, we can introduce an additional context node, depicted as the rounded rectangle ”Party” node in Figure 2. The arrow relationships from the misalignment reason node to the values indicate their respective effects. To determine the complete impact of a choice within the context, the weights of the incoming arrow relationships are summed.

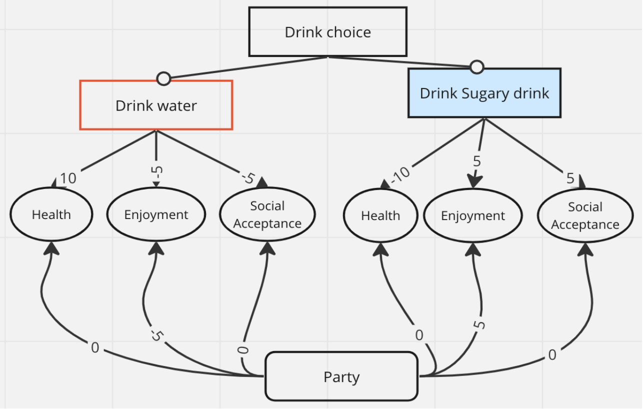


Figure 2: The drinking choice behavioral challenge involves two options: drinking water and drinking sugary drinks. The presence of the misalignment reason (party) influences the effects of these choices on the value nodes. The weights from the Party node to the value nodes indicate how much the misalignment reason modifies the baseline impact of a choice on a value. For instance, while drinking water generally has a negative effect on enjoyment, in the context of a party, the combined incoming weights (-10) reflect a significantly more negative impact on the enjoyment value. In simple terms, although a person may not typically enjoy drinking water, the context of a party intensifies the negative enjoyment associated with it.

2.3 Baseline tree

To assess the accuracy of the generated behavior trees, a baseline is required, representing the user's actual perceived values. In this study, participants are asked to enhance the generated trees to align with their perceived values. Each participant receives a tree for each scenario, generated based on their responses. They manually review and adjust each weight to create an *improved* behavior tree that better reflects their perceived values.

2.4 Evaluation methods

Accuracy of the generated tree

Given a behavior tree: a generated tree $G(V, E)$ in which V is a set of nodes and E is a set of edges connecting some two nodes $i, j \in V$. An improved tree $T(V, E)$ has the same structure as the generated tree G, meaning the same set of nodes and connections between nodes, except for each edge $e_{i,j} \in E$ the weight value $w_{i,j}^{base}$ denoting weight value between $i, j \in V$ in G might be different from $w_{i,j}^{improved}$ denoting weight value between $i, j \in V$ in T. We can thus compare G and T using the following distance measures

$$Hamming\ Distance = \#\ of\ weight\ changes$$
$$between\ G\ and\ B\ trees$$

$$AWC = \sum_{e_{i,j} \in E} |w_{i,j}^{baseline} - w_{i,j}^{generated}|$$

Usability

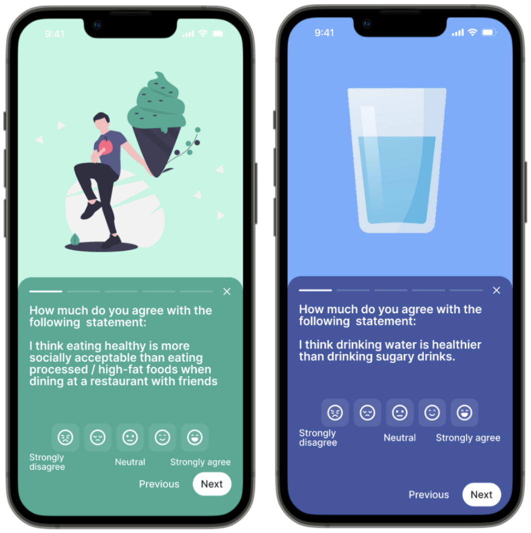
Each participant performs a System Usability Scale (SUS), as outlined in [6], to evaluate the usability of the interface.

3 Interface Design

A mid-fidelity prototype was created using Figma[1], a collaborative wireframing tool. Using Figma we can create interactive prototypes that will allow for experiment subjects to interact with an interactive med-fidelity prototype while allowing us to deliver the prototype reasonably fast. Figure 3 shows snippets of the prototype. The full prototype can be found in this Figma link.

3.1 Rationale for design choices

Below a discussion about how certain design principles are ensured can be found.


How much do you agree with the following statement:

I think eating healthy is more socially acceptable than eating processed / high-fat foods when dining at a restaurant with friends

How much do you agree with the following statement:

I think drinking water is healthier than drinking sugary drinks.

Figure 3: A snippet of the prototype that was created for the study. The left screen is from the third misalignment scenario described in Appendix A, while the right screen is from the first.

Visibility of system status

1. The status bar at the top of the bottom sheet displays the current question and progress, informing users about their task completion.
2. Clicking on a choice highlights the corresponding emoticon, and the answers are preserved when navigating forward or backward.

Recoverability

1. Users can correct mistakes by selecting a different answer.
2. Accidental clicks on the "Next" button can be reverted using the "Previous" button.

Multi-modality

1. The interface utilizes textual and visual cues to immerse users in the context.

Minimalistic design

1. Each screen has at most three interactive elements, all crucial to answer questions and to navigate through them.

[1] https://www.figma.com

4 Questions for eliciting values

4.1 Questions

The questions were designed in a way that could be answered on a 5-point Likert Scale. There are two types of statements: one asks about user's values in general, and the other sets the question in a specific context, a misalignment reason. For example:

- **General question:** I think drinking water is more socially acceptable than drinking sugary drinks.
- **Context-specific question**: I think drinking water is more socially acceptable than drinking sugary drinks when I am at a social event.

The full set of questions for all four misalignment scenarios can be found in Appendix B.

4.2 Building Behavior Trees from user's answers

Using the general and context-specific questions a behavior tree can be generated. To simplify the manual process of drawing trees four scripts, one for each scenario, have been written that generate the trees. The scripts have been open-sourced and can be accessed in this GitHub repository.

Manually generating a behavior tree

General questions with no specific context are used to elicit general values from the user. The answer of "Strongly agree" to "I think drinking water is more socially acceptable than drinking sugary drinks." will result in two edges:

1. $Drink\ Water \rightarrow Health : 10$
2. $Drink\ Sugary\ Drink \rightarrow Health : -10$

So the first subject of the questions gets assigned the weight of the answer, while the second subject of the question gets assigned an additive inverse (a number that together adds up to a zero) of the weight of the first edge.

Context-specific questions are used to discover how user's values changes given a misalignment reason. The answer of "Strongly agree" to "I think drinking water is more socially acceptable than drinking sugary drinks when I am at a social event." shows no change in user's value when compared to general case. The answer will result in two edges of 0 weight:

1. $Party \rightarrow Health\ (Water) : 0$
2. $Party \rightarrow Health\ (Sugary\ Drink) : 0$

5 Experimental Setup

The experiment will involve a sample group of 15 people ranging from 18 to 65 years of age, all of whom are technologically literate and capable of successfully using a smartphone. Ideally, the participants will be chosen from diverse disciplines to ensure a varied representation within the sample group. Within the study, 2 experimental setups are chosen: single-condition and multiple-condition. Single-condition is a between-subject design with subjects testing one of the 5 conditions. Multiple-condition setup is a within-subject experiment where each participant tests every condition.

5.1 Single-Condition Setup

Procedure Each participant:

1. Sends a consent form to be signed and participant information to get more familiar with the study.
2. In a call together with a participant the procedure is explained and a base and a context tree as defined in Subsection 4.2 is built.
3. Using a custom Python script generate trees and import them to a Miro board
4. During a call, the Miro board will be shared, and each weight will be reviewed together with the participant. The participant's task is to adjust any weight that does not align with their perceived values. This step aims to create a baseline tree that accurately measures the generated tree's correctness.
5. Lastly, each participant perfrosm System Usability Scale (SUS), as outlined in [6], to evaluate the usability of the interface.

5.2 Multiple-Condition Setup

In the second part of the experiment, five of the participants carried out the experiments of all conditions, exactly as defined in a single-condition setup. The same accuracy measures will be collected and used as in a single-condition setup. Additionally, the conditions will be ranked by each participant based on how accurately they represent our values. The ranking together with accuracy measures will allow us to see how well-defined the accuracy metrics are.

Five participants in the Multiple-Condition setup are peers in the research group. It's important to acknowledge the inherent biases in this setup. However, due to time and resource limitations, this approach was chosen. Despite the absence of a scientifically sound sample, the lack of any learning effect makes the results valuable for the study.

6 Responsible Research

As the study involves an experiment with human participants, ethical considerations and implications in this section will be discussed. Additionally, it is important to consider reproducibility of experimental methods and results. In this section, steps taken to ensure the responsible conduct of the study and experiment will be discussed.

6.1 Ethical Considerations

Human Research Ethics Commite (HREC) at TU Delft has granted approval to conduct this research and experiment with human participants. Regardless it is important to discuss and analyze any other ethical considerations.

1. **Informed consent:** Prior to participating in the experiment, all participants will be provided with a detailed explanation of the study, the background information, and the study's goals. They are also informed of their right to withdraw from experiment at any moment and about the data privacy policy and how data will be anonymized and stored.

2. **Accurate value models pose security risk:** It is fair to argue accurate value models of oneself are personal data. If a method to uncover the accurate models is discovered, it can be used by bad actors to learn about people's behavior and values and used against a person.
3. **Behavior support agent manipulation:** Behavior support agents given access to accurate value models might be used to manipulate and influence people's behaviors. This can be used by bad actors to influence people's choices and over a long time change their value profiles.

6.2 Reproducibility

Reproducibility of the results is important to ensure the trustworthiness of the results and guarantees any findings are replicable.

1. **Releasing prototype and code:** In Section 3 and Subsection 4.2 a link to the full prototype and Python code to build Behavior Trees from users' answers are provided. This ensures that while reproducing results one can use the exact same prototype and code that was used by part of this study.
2. **Detailed methodology:** In Section 2 and the methodology is clearly outlined, including the full description of the Behavior Tree Model, how to build behavior trees from users input and a methodology to measure accuracy of generated trees. In Section 5 the procedure of the experiment, the target audience are clearly defined to be reproduced.
3. **Published results**: In Section 7 there is an overview of study's participants and the results from the study. This ensures that using methods defined in Sections 2 and 5 one can compare the produced results.

7 Results

This section presents the study results. Subsection 7.1 provides an overview of the recruited participants. Subsection 7.2 presents the results of the Single-Condition experiment, while 7.3 discusses the results of the multi-condition experiment. Subsection 7.5 presents qualitative data on user-reported limitations.

7.1 Participants

In the single-condition experiment, 15 subjects with diverse backgrounds, including industrial design, mechanical engineering, computer science, economics, business, finance, chemistry, and marketing & communication, participated. Their ages ranged from 19 to 24 years old.

In contrast, the multi-condition experiment involved 5 computer science bachelor students aged between 21 and 24.

All participants in both experiments were recruited from personal networks and did not receive any compensation for their participation.

7.2 Single-Condition experiment

In this section, the results from conditions A, B (Graphical Interface and isolation questions), and C (text interface and comparative questions) are reported, as B and C are the conditions that overlap with the condition of this study.

1. **GUI**

Comparative			Isolation		
Median	Mean	SD	Median	Mean	SD
40	36.87	21.92	0	8	13.07

Table 1: Median, mean and standard deviation of absolute weight change for graphical user interfaces.

The mean of AWC of GUI using comparative questions is 28.87 higher than mean of distance of GUI using isolation questions. The null hypothesis of no mean difference is rejected with a p-value of asd.

Comparative			Isolation		
Median	Mean	SD	Median	Mean	SD
6	5.33	3.53	0	1.33	2.19

Table 2: Median, mean and standard deviation of Hamming distance for graphical user interfaces.

The mean of the Hamming Distance of GUI using comparative questions is 4 higher than the mean of the Hamming distance distance of GUI using isolation questions. The null hypothesis of no mean difference is rejected with a p-value of m

2. **Text Interface and Comparative questions**: absolute weight change (AWC) measure is reported to have a median of 0, a mean of 9.67, and a standard deviation of 14.20. On the other hand, the hamming distance measures have a median of 0, a mean of 0.8, and a standard deviation of 0.98.

The mean of AWC of GUI using comparative questions is 27.2 higher than the mean of the distance of text user interface using comparative questions. The null hypothesis of no mean difference is rejected with a p-value of 0.00214.

The mean of the Hamming Distance of GUI using comparative questions is 4.53 higher than the mean of the Hamming distance of text user interface using comparative questions. The null hypothesis of no mean difference is rejected with a p-value of 0.0003.

Average change per changed edge

GI exhibited lower average Hamming Distance, Distance and change per changed edge than GC. While TC exhibited lower average Hamming Distance and Distance, a change per changed value was almost twice as high.

	Hamming Distance (avg)	AWC (avg)	Normalized AWC
GC	5.33	36.87	6.92
GI	1.33	8.00	6.15
TC	0.80	9.67	12.09
TI	5.07	30.87	6.09
AI	3.60	13.50	3.75

Table 3: Results for single-condition experiment. Normalized AWC is AWC divided by Hamming distance.

Hamming distance and absolute weight change are more readable when plotted on a 2D Cartesian coordinate system as seen in Figure 4. Points closer to the origin represent more accurate conditions.

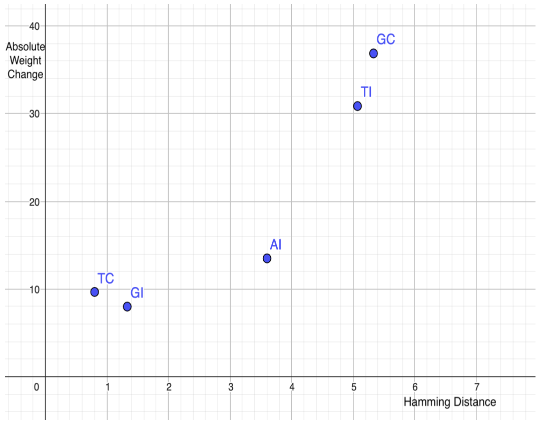


Figure 4: Each condition's Average Absolute Weight Change and Hamming distance plotted on 2D Cartesian coordinate system. Points closer to the origin are more accurate.

7.3 Multiple-Condition experiment

	Hamming (avg)	AWC (avg)	Normalized AWC
GC	8	50.5	6.31
GI	3.5	17.5	5
TC	6.75	55	8.15
TI	6.5	40	6.15
AI	0.5	2.5	5

Table 4: Results of a multiple-condition experiment.

7.4 System Usability Scale (SUS)

The System Usability Scale (SUS) is a metric used to measure the usability of a system. In this case, the mean SUS score is 78.17, indicating an average level of usability. The standard deviation is 7.04.

According to the research [7] a score of 78.17 falls within the 80th to 84th percentile range. This means that the usability of the system is relatively good compared to other systems. Furthermore [7] assigns a B+ grade to this score, indicating a positive evaluation.

In terms of adjective descriptions, a SUS score of 78.17 for the prototype falls between the categories of ”Good” and ”Excellent” according to [7]. This suggests that the usability of the prototype is quite favorable and surpasses the average level.

7.5 User-reported limitations

1. Model-related comments
   - ”I really don't enjoy, don't think it's safe and have no comfort when running in general. It is even more so when it is raining, but I have no way to express it as the values are capped at -10.”
   - ”Capping does not allow to further decrease values. In this case I think safety and enjoyment should decrease further.”
2. Comparative questioning technique related comments
   - ”Drinking water enjoyment is neutral,”
3. Lack of context
   - ”I would put -10 on the health of watching movies if I had to judge long-term effects on my health. Now I will leave it at 0, as I think about these trees as one-off decisions”
   - ”When answering the questions it's hard to identify values for each scenario so it's easy to generalize too much.”

8 Discussion

In this section the results presented in the Section 7 will be discussed.

8.1 Accuracy of generated trees

Accuracy significance

On average using GUI and comparative questioning techniques there were 5.33 weights changed, meaning that on average there are 1.33 mistakes per scenario per person. Additionally, a weight was changed on average by 6.92, which is bigger than one step in the Likert-like scale proposed in Section 2. According to [8] trust plays a vital role in all scenarios that involve collaborative decision-making, as the level of trustworthiness attributed to an agent determines the user's ability and willingness to interact with it. This means that any deviations from user's actual value models are detrimental to the agent-human interaction, therefore an average change of 6.92 is undesirable.

Contradictory results

Notice in Figure 4 how GC is less accurate than GI, therefore assuming all things equal between conditions we can conclude that isolated questions lead to more accurate models. Two reasons for inaccuracies in the generated GI model are the limitations of the behavior tree model and the nature of comparative questions. Comparative questions efficiently provide scoped information about user preferences, requiring fewer questions to build a behavior tree. However, the behavior tree model stores isolated weights between values, which misaligns with how values are elicited (comparatively) and observed (in isolation) by the user. For example, when comparing exercising outside to watching a movie in rainy and windy weather, a person's answer of ”Much less healthy” implies that watching a movie becomes ”Much more healthy” in that context. However, the model stores these as separate data points, resulting in potential incorrect weights.

However, Figure 4 shows how TI is less accurate than TC, therefore assuming all things equal between conditions we could conclude that comparative questions lead to more accurate models. This is a contradictory conclusion to the one in the previous paragraph. No clear reason for TI being less accurate than TC have been found. A hypothesis could be done that due to discrepancies in the experimental procedures, the results are unreliable. Similar contradictory results were found in the relationship between user interface modality and accuracy.

Comparative versus isolation questions
The mean of behavior tree distance of GUI using comparative questions is 36.87, compared to 8.00 of a GUI using isolation questions. Since behavior trees store isolated relationships between values and choices, the in-isolation questioning technique is much more aligned with the behavior tree model.

8.2 Fatigue from questions

Comparative questions are concise and the prototype requires twice as few questions to build the same behavior tree, which leads to less fatigue and less likely to observe survey speeding. This can be clearly visible with the following three questions:

1. How much do you enjoy water?
2. How much do you enjoy soda?
3. How much do you enjoy water over water?

A single comparative question (c) provides enough information to learn about users' preferences for water in comparison to soda, whereas to achieve the same effect two isolated questions (a) and (b) are required.

The participants doing conditions that use comparative questioning techniques had to answer 28 questions, while isolated technique survey contained 56 questions. Coupled with the fact that a personal network was used to recruit participants and participation were not compensated, it is likely that survey fatigue and survey speeding might have taken place.

8.3 Limitations

1. **A vs B $\neq$ B vs A** [9] states that in comparative questions that compare X to Y, X is the subject and Y is the referent of the question. Furthermore, [9] argues that individuals tend to primarily focus on the characteristics of the subject being compared and overlook the characteristics of the referent. This inherent flaw in comparative questions should be taken into account when developing techniques to elicit personal values, as responses may be biased toward the subject.

2. **Lack of unification of experimentation procedure.** During the experimentation procedure, participants who were left alone during the baseline tree building phase were observed to be significantly quicker in expressing satisfaction with the results. The number of weights involved in tweaking the generated tree differed between conditions, with comparative question trees having 56 weights and isolated question trees having 112 weights. To ensure careful task completion, the experimentation procedure included supervision during baseline tree building.

   In terms of conducting the experiment, there were differences among conditions. For instance, in condition TC, participants were left alone to tweak the weights. Although there is no strong evidence suggesting that this led to higher accuracies, observations from test runs indicated that participants expressed satisfaction with the baseline tree more quickly compared to those who carefully considered each weight.

3. **Validity of the baseline**. There is no clear evidence if the baseline tree represents the ground truth of a person's values, therefore no evidence that the used accuracy measures are sound. To create the baseline, participants were instructed to ”tweak the weights, if you feel like they are incorrect”. The task instruction of correcting mistakes might lead to people being more suspicious about the weights, thus making the baseline tree on average more distant from the generated tree than it actually is.

9 Conclusions and Future Work

The study discussed in [1] highlights the need for personalized systems that can understand the unique needs of individual users, as the symbiotic relationship between individuals and support agents continues to grow. While there are proposed solutions for incorporating user values into decision-making, there is still a gap in knowledge regarding how to elicit and update user value models over the lifespan of the system. Updating user values is crucial due to the dynamic nature of values, which can change over time. One approach to capturing user values is through the use of misalignment scenarios, where unforeseen circumstances alter human behavior. In this study, we focus on the design of a graphical user interface that utilizes behavior trees to elicit and update personal values. The graphical user interface employs Likert Scale comparative questions specifically designed to elicit values, which are then stored within the behavior trees. These were the results of the study

- Mean distance between generated and baseline trees: 6.92, which given the 5-point Likert Scale means that on average a mistake was a bit larger than one step.
- Mean Hamming Distance (number of changes) of 1.33 per tree, which means that on average more than one mistake per tree was exhibited.
- A System Usability Scale (SUS) score of 78.17, putting the usability of the prototype inbetween the 80th and 84th percentile, a grade of B+ and adjective description of between ”Good” and ”Excellent”. The results suggest that the usability of the prototype surpasses the average level and is quite favorable.
- Graphical user interface with isolation questions and text user interface with comparative questions had much lower Hamming and behavior tree distances.
- The Hamming Distance of GUI (Graphical User Interface) using isolation questions shows a comparable change per changed value when compared to GUI with comparative questions, while the Hamming Distance of text user interface using isolation questions demonstrates an approximately twofold higher change per changed value. This observation suggests that, on average, there are fewer mistakes in the text user interface, but they are of a larger magnitude.

- Relationship between questioning techniques lead to a contradiction. TC is more accurate than TI, but GC is less accurate than GI. Although there is a reasonable explanation of isolated questions leading to more accurate results, none was found to support comparative questions leading to more accuracy. A hypothesis that experimental procedure discrepancies lead to unreliable and hardly comparable results could be made. Similar contradictory results were found in the relationship between user interface modality and accuracy.

All in all, the performance in terms of accuracy of Graphical User Interface with comparative questions was worse than both Graphical User Interface with isolated questions and the text user interface with comparison questions. The difference between graphical with isolated questions can be explained by the fact that comparative questions are not aligned with behavior tree model, while the reason for the difference between textual interface remains unclear, although it is worth to mention that discrepancies between experimental procedures might have been the cause, but would require further research to strongly conclude it.

In terms of usability, there was no statistically significant difference between the averages of usability scores between textual and graphical interfaces using comparative questioning. While GUI with isolated questions performed significantly better than GUI with comparative questions.

9.1 Future work

The experiment and the use of comparative questioning with behavior tree models could be improved on multiple different axes.

1. **Model Improvements** Comparative questions efficiently provide scoped information about user preferences, requiring fewer questions to build a behavior tree. However, the behavior tree model stores isolated weights between values, which is misaligned with how values are elicited (comparatively) and observed (in isolation) by the user. Therefore, an alternative model is able to store value preferences comparatively. The model should additionally have an infinite range of values to allow for further decreases or increases to weights provided the misalignment scenario.
2. **Supervised process for cognitive tasks** The construction of the baseline tree demands significant cognitive focus and a thorough comprehension of the behavior tree model. While there is no substantial evidence, direct observations from trial tests indicated that unsupervised baseline tree building yields significantly higher accuracy measures. In this experiment, there were discrepancies within the procedure of the experiment, where the text user interface and comparative questions condition used unsupervised tweaking. To ensure consistency, future research should implement a supervised process that is fully aligned across different conditions and ideally conducted by the same person or according to a strict protocol.
3. **Frictionless elicitation technique.** The value elicitation method lacks integration into a human's life, as the 28 questions employed resemble a survey-like procedure. Additionally, the System Usability Scale fails to measure user fatigue, which is an important aspect to consider. Tracking fatigue could provide valuable data regarding the usability and feasibility of eliciting personal values through lengthy surveys.
4. **Random value shifts and big life events.** This study explores updates arising from misalignment scenarios. However, changes in human values can be triggered by various factors, including random shifts in desires or interests, significant life events such as attending university or experiencing a global pandemic [10; 11]. These factors may lead to more substantial shifts in values compared to misalignment scenarios. Future research should investigate re-trigger mechanisms or regular check-ins on a monthly or quarterly basis. Given the system's widespread adoption, employing a machine learning technique to learn and predict potential goal updates based on similar user profiles could be beneficial.
5. **Potential of large language models.** Given the progress made in large language models (LLMs), it is worth considering the potential enhancements in inference when applied to textual representations of value models. Furthermore, by providing LLMs with ample context and training data, we can enable more intelligent weight updates, surpassing the existing proposed update method of the behavior tree model.

A Misalignment Scenarios

1. - **Long term goal:** Drinking more water
   - **Behavioral Challenge:** Drink Choice
   - **Possible behavior choices:** Drink water or drink sugary/sweet drinks
   - **Affected values:** Health, Enjoyment, Social Acceptance
   - **Misalignment reason:** Attending a social event (Party)
2. - **Long term goal:** Exercising
   - **Behavioral Challenge:** Things to do in the evening
   - **Possible behavior choices:** Running or Watching a movie
   - **Affected values:** Health, Enjoyment, Safety, Comfort
   - **Misalignment reason:** Bad weather, rain
3. - **Long term goal:** Eating Healthy
   - **Behavioral Challenge:** Food Choice
   - **Possible behavior choices:** Eat healthy or eat high fat/processed foods
   - **Affected values:** Enjoyment, social acceptance, health, wealth
   - **Misalignment reason:** Eating out with friends majority of whom order junk food.
4. - **Long term goal:** Better Sleep Schedule
   - **Behavioral Challenge:** Bedtime

- **Possible behavior choices:** Early or late
- **Affected values:** Health, Wealth, Career
- **Misalignment reason:** Work deadline

B Scenario Questions

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?

References

[1] C. Stephanidis, G. Salvendy, M. Antona, J. Y. C. Chen, J. Dong, V. G. Duffy, X. Fang, C. Fidopiastis, G. Fragomeni, L. P. Fu, Y. Guo, D. Harris, A. Ioannou, K.-a. K. Jeong, S. Konomi, H. Krömker, M. Kurosu, J. R. Lewis, A. Marcus, G. Meiselwitz, A. Moallem, H. Mori, F. Fui-Hoon Nah, S. Ntoa, P.-L. P. Rau, D. Schmorrow, K. Siau, N. Streitz, W. Wang, S. Yamamoto, P. Zaphiris, and J. Zhou, "Seven hci grand challenges," *International Journal of Human–Computer Interaction*, vol. 35, pp. 1229–1269, 07 2019.

[2] S. Cranefield, M. Winikoff, V. Dignum, T. Delft, M. Dignum@tudelft, F. Nl, Dignum, F. Dignum@uu, and Nl, "No pizza for you: Value-based plan selection in bdi agents," 2017. [Online]. Available: https://www.ijcai.org/proceedings/2017/0026.pdf

[3] M. Tielman, C. Jonker, and M. Birna Van Riemsdijk, "What should i do? deriving norms from actions,values and context." [Online]. Available: https://ceur-ws.org/Vol-2134/paper10.pdf

[4] K. C. Calman, "Evolutionary ethics: can values change," *Journal of Medical Ethics*, vol. 30, pp. 366–370, 08 2004.

[5] S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li, "Comparable entity mining from comparative questions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 1498–1509, 07 2013.

[6] J. Brooke, "Sus -a quick and dirty usability scale," 1986. [Online]. Available: https://digital.ahrq.gov/sites/default/files/docs/survey/systemusabilityscale%2528sus%2529_comp%255B1%255D.pdf

[7] J. Lewis and J. Sauro, "Item benchmarks for the system usability scale," *Journal of Usability Studies*, vol. 13, pp. 158–167, 2018. [Online]. Available: https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Lewis_May2018.pdf

[8] S. Daronnat, L. Azzopardi, M. Halvey, and M. Dubiel, "Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration," *Frontiers in Robotics and AI*, vol. 8, 07 2021.

[9] M. Wanke, N. Schwarz, and E. Noelle-Neumann, "Asking comparative questions: The impact of the direction of comparison," *The Public Opinion Quarterly*, vol. 59, p. 347–372, 1995. [Online]. Available: https://www.jstor.org/stable/2749757?seq=1

[10] V. R. Krishnan, "Impact of mba education on students' values: Two longitudinal studies," *Journal of Business Ethics*, vol. 83, pp. 233–246, 11 2007.

[11] A. Bojanowska, D. Kaczmarek, M. Koscielniak, and B. Urbańska, "Changes in values and well-being amidst the covid-19 pandemic in poland," *PLOS ONE*, vol. 16, p. e0255491, 09 2021.