Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues

Chen, Pei Yu; Tielman, Myrthe L.; Heylen, Dirk K.J.; Jonker, Catholijn M.; Van Riemsdijk, M. Birna

DOI
10.3233/FAIA230077

Publication date
2023

Document Version
Final published version

Published in
HHAI 2023

HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230077

Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues

An Exploratory Focus Group Study

Pei-Yu CHEN [a,1], Myrthe L. TIELMAN [a] Dirk K.J. HEYLEN [b]
Catholijn M. JONKER [a,c] and M. Birna VAN RIEMSDIJK [b]
[a] *Delft University of Technology, The Netherlands*
[b] *University of Twente, The Netherlands*
[c] *Leiden University, The Netherlands*

Abstract. For personal assistive technologies to effectively support users, they need a user model that records information about the user, such as their goals, values, and context. Knowledge-based techniques can model the relationships between these concepts, enabling the support agent to act in accordance with the user's values. However, user models require updating over time to accommodate changes and continuously align with what the user deems important. In our work, we propose and investigate the use of human-agent *alignment dialogues* for establishing whether user model updates are needed and acquiring the necessary information for these updates. In this paper, we perform an exploratory qualitative focus group study in which we investigate participants' opinions about written examples of alignment dialogues, as a foundation for their design. Transcripts were analyzed using thematic analysis. A main theme that emerged concerns the potential impact of agent utterances on the user's feelings about themselves and about the agent.

Keywords. Human-agent alignment, User modelling, Values, Behaviour support technology, Conversational agents, Dialogue

1. Introduction

Behaviour support technology is being developed to perform tasks on our behalf or to guide our actions. In the area of health and well-being, for example, there are support agents that remind us to take our medicine [1], to help us eat healthier [2], and to coach us on well-being [3]. As support agents become more and more integrated with our daily lives, it becomes even more important that they provide support that is in line with the goals, norms, values, capabilities and context of users [4].

Existing work has proposed computational representations for capturing such human notions as a user model in the agent [5,6,7,8]. These Semantic User Models [9] employ knowledge-based techniques, comparable to a representation of ontologies in a

[1] Corresponding Author: Pei-Yu Chen, p.y.chen@tudelft.nl

semantic web context, through which the user's motivational attitudes and their relations with user actions are modelled explicitly. For example, a user model can describe the daily activities and sub-activities of the user, the user's capabilities in performing these activities, and which values are promoted or demoted by which activities [5]. If the user's goals and value preferences are also modelled, the agent can select support actions that are in alignment with what the user finds important.

A challenge is how to acquire the information that is to be captured in a Semantic User Model. This is particularly challenging because such a model not only records individual pieces of information about the user but also relations between concepts. Some of this information, e.g., regarding the current habits of a user, may be obtained through analysis of behavioural user data. However, doing so for high-level concepts such as values or goals can be challenging [10]. Moreover, behavioural data reflects people's past behaviour rather than their future desired behaviour. Capturing the latter is particularly important for agents intended to support a user in changing their behaviour. In addition, data-driven approaches can lack transparency because of the complex relationship between input data and a model's output [11]. This makes it difficult for users to understand how the system works and also to adapt the system to their preferences.

In our research, we explore a complementary approach for acquiring user model information, namely via interaction between user and support agent, specifically via a user-agent *dialogue*. The idea is that the support agent will have a conversation with the user where the agent asks the user about the activities they need support with and the underlying values [9]. An initial version of such a user model, however, is unlikely to provide a complete and fully accurate picture of the user's needs and contextual factors throughout the period of use of the agent. There can be situations where the user model needs to be updated, for example, because of changes in user needs or context.

In this paper, we take a first step in designing human-machine *alignment dialogues* (introduced in Section 3) that aim to identify and repair such misalignments between user and agent, or prevent future misalignments. We perform an exploratory qualitative focus group user study, in which we show participants different written variants of what such human-agent alignment dialogues might look like and discuss their opinions (Section 4). With this first study, we aim to identify dimensions that are important for designing good alignment dialogues. We analyze the focus group transcripts using thematic analysis, through which we identify main themes, concepts, and their relations. Based on this we highlight several considerations that need to be taken into account in the next step in developing alignment dialogue models (Section 5). We discuss related work in Section 2 and discuss our findings and conclude the paper in Section 6.

2. Related Work

The concept of human-agent alignment dialogues can be positioned at the intersection of research on conversational agents and human-agent teamwork.

2.1. Conversational agents

Conversational agents have already been extensively investigated in the context of healthcare [12], for example, to support users in self-care, retrieving information, or non-

task related interactions. These dialogue approaches are typically frame-based, where users are asked to fill in slots in a template, or they take place through a series of predetermined steps. Elicitation and use of richer Semantic User Models is a novel approach that facilitates more comprehensive and personalized support. This richness means that more nuanced and context-aware information can be integrated into the model, which requires updates as changes occur. Dialogues to facilitate such updates have, to the best of our knowledge, not yet been investigated.

Moreover, in the area of conversational agents, a concept related to alignment dialogues is studied, namely *dialogue alignment* [13]. This concerns alignment processes *in* dialogues, as opposed to the use of dialogues *for* human-machine alignment as we introduce in this paper. For example, interlocutors in a conversation tend to develop the same set of referring expressions to refer to specific objects [14]. Dialogue alignment processes may be part of an alignment dialogue in order to achieve successful human-machine alignment and provide the proper support to the user.

2.2. Human-agent teamwork

Furthermore, in our approach, we take inspiration from research in human-agent teamwork, since the user and support agent can be viewed as a team working together to ensure the user is supported appropriately. From shared mental model theory we know that mental model sharedness – defined as "overlapping mental representations by team members that reflect how the group members as a collectivity think or characterize phenomena" [15] – improves team performance (see, e.g. [16,17,18]). Shared mental model theory has been translated to the context of (human-)agent teams, arguing that sharedness is also important when artificial agents are involved [19]. Sharedness of the mental model that the agent has of the user's goals, and the user's own mental model of their goals, can improve the agent's support and alignment with the user's needs. Since the agent cannot directly inspect the content of the user's mental model, alignment dialogues can be a way to elicit the relevant information and update its user model.

Moreover, the coactive design [20] approach to human-agent teamwork argues for the importance of designing for human-machine interdependence for realizing resilient human-machine systems. Alignment dialogues can be viewed as a way of designing for interdependence in the context of support agents, ensuring that human-machine misalignments can be identified and repaired.

3. Human-Agent Alignment Dialogues

In this section, we outline the concept of alignment dialogues. We will start with an illustrating example of how misalignment between users and support agents could happen and what a corresponding alignment dialogue could look like.

Alignment dialogue example:

Scenario: Upon initialization, John has told the agent about his ideal exercise schedule. However, since then, his opinion about this has changed. When the agent asks him to stick to the original schedule, a misalignment situation occurs.

John: I don't like the exercise schedule.

Agent: Are you no longer motivated to exercise?
John: I am, but I just want more variety.
Agent: Okay. Anything in particular that you would like to include?
John: Could you add swimming to my schedule and make suggestions more randomly?
Agent: Yes. Anything else I can do for you?
John: No thank you.

We identify three types of alignment (Figure 1), referring to the desired support as the *purpose* of alignment, the user model as the *means* for aligned support, and behavioral non-compliance as the *trigger* indicating a possible need for starting an alignment dialogue: 1) alignment between *agent's support actions* and the support *users need/want* which is reflected in their actual behaviour, 2) alignment between agent's *user model* and user's *self-model*, and 3) alignment between user's *actual behaviour* and their *desired behaviour*.

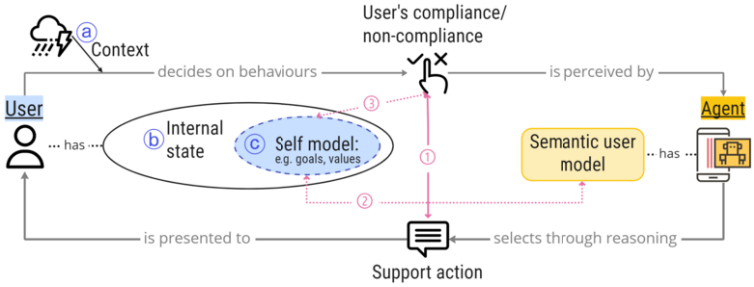

Figure 1. Alignment dynamics in user-agent interaction: types of alignment (1,2,3) and changes that could cause misalignment (a,b,c); see text for further explanation. The primary goal of alignment dialogues is to address the first type of alignment, as indicated by the solid pink arrow.

Alignment types The first type of alignment is the broadest and encompasses the second and third types. As the agent can never fully grasp the user's true self-model, it cannot be certain about the second and third types of alignments. Therefore, the focus is on the first type of alignment, which pertains to the match between the agent's support actions and the user's needs/wants. We define the user's needs as 'the means for them to achieve a specific goal that the agent facilitates and promotes.' A corresponding notion of misalignment can be described conversely as 'a situation where the provided support does *not* match what the user needs or wants.'

It is important to note that what the user wants at a specific moment may not be what they need with regard to their goals. Similarly, there might be a conflict between short-term and long-term goals [21]. This then gives rise to what may be seen as a moral challenge of whether the agent should align with what the user wants now, or what they need long term. In this paper, we focus on *how* the user wants to talk about what they need or want in an alignment dialogue, and how to resolve this via an alignment dialogue. The outcome of this dialogue can then be a way for the agent to determine the most appropriate support.

Reasons for misalignment In the example above, misalignment arises because the user changes their mind with regard to their desired exercise schedule. We identify three main reasons why misalignment can arise in general. These are derived from the way the agent's support actions would be chosen, i.e., using reasoning based on information in the user model. A misalignment between the agent's support action and the user's support needs can thus arise because, first, the agent's reasoning process itself can be wrong. Second, the agent's user model can be wrong initially. Third, something can change that requires an adapted interpretation of the situation compared to the information captured in the existing user model. Regarding the last one, we further identify three aspects that could change and cause misalignment (illustrated in Figure 1):

(a) **Context**: This includes the external factors or environments of the user, such as the weather, special occasions, or events. When the context changes, the original support may no longer match the user's needs.
(b) **User's internal state**: A user's internal states encompass the user's emotions, stress level, physical or mental conditions, etc.
(c) **User's desired behaviour**: As time passes, the user may want to adjust their goals or other motivational attitudes. In this case, the user model may have been correct at the beginning; however, it is the user themselves who changes over time, requiring the agent to adapt the user model to ensure alignment.

Alignment dialogues We define alignment dialogues as 'dialogues with which the agent and user try to achieve or maintain alignment.' This can include first establishing if there is a misalignment, as it might not always be obvious to the agent if an observation of the user's behaviour points to a misalignment. If there is a misalignment, the conversation could shift to talking about how to solve the situation, where the user and the agent take on a question-answering approach. The cycle will continue until the misalignment no longer exists, and the agent will have obtained a better understanding of the current situation.

4. User Study

In the previous section, we have outlined what we mean by (mis)alignment and how we see the role of alignment dialogues. To better understand how we could shape alignment dialogues, we performed a qualitative user study to explore people's opinions and ideas about alignment dialogues.

4.1. Focus group with scenarios

The user study was performed in the form of focus groups using a scenario-based approach. We chose to conduct focus groups because the interaction between the participants could spark more discussions regarding what they like or dislike about certain aspects in alignment dialogues [22,23]. The moderator encouraged participants to express different opinions and ensured participants got a chance to share their views. In the focus groups, the participants were presented with six scenarios with accompanying variants of human-agent dialogues described in textual form, similar to the example in Section 3. We followed the definitions and procedures by [24,25] to create six scenarios. We identified

factors that could lead to misalignment (Section 3) and what the corresponding dialogues could be. From a large number of possibilities, we chose six scenarios with differences and diversity to cover various challenges. For each scenario, we had one or two variants of alignment dialogues to address it. Details are further discussed in Section 4.3.

4.2. Participants

Eligible participants were those who were fluent in English and current or potential users of behaviour support technology. A total of 13 adults participated in two focus groups, seven in the first group and six in the second group (eight males and five females, age = 26.08 years, SD = 2.72 years) from various countries of origin. The participants were recruited through our networks or through advertisements on social media. We obtained approval to conduct the study from the Human Research Ethics Committee of Delft University of Technology (ID nr 1673). We e-mailed the informed consent forms, including the request for consent to record videos, to the participants before the study and asked them to sign them.

4.3. Materials

The study was divided into two parts. The first part consisted of general questions with regard to behaviour support agents. The purpose of the first part was to familiarize the participants with this type of agent and its role. For the second part, we focused on participants' opinions on alignment dialogues. Six misalignment scenarios and their corresponding alignment dialogues were shared with the participants in textual form over several pages.

Scenarios In all six misalignment scenarios, the behaviour support agent had conversations with a fictional persona named John. John's age, profession, social relationships, hobbies, and the behaviour change he needs were detailed in written form, alongside a picture of a white man. These misalignment scenarios occur because the user (John) deviates from their goals due to the unpleasant weather (Scenario 1), the user's mood (Scenario 2), an occasional birthday party (Scenario 3), and changes in desired activities (Scenario 4). Additionally, we included Scenario 5 in which the provided support is in fact in line with user needs but it is so 'accidentally' (the agent suggests the user go to work by bike for health reasons but the user does so for an environmental reason). In Scenario 6 the agent is pushy by asking the user to exercise repeatedly even if rejected multiple times, which in itself might be a misalignment as it may deviate from the support the user needs.

Dialogues For each scenario, one or two versions of alignment dialogues were created, with variations in several aspects. One of the key differences between the versions is the **depth of reasoning**: In one dialogue variant, the agent asks surface-level questions, while in another it aims to elicit user values. Values are considered to be a driving factor in human behaviour [26,27]. [28] defined values as "what a person or group of people consider important in life". For example, in an alignment dialogue, the user may say "because it's raining" when asked why they do not want to go running. If the agent continues asking, the user's values (e.g. *comfort*) may be revealed.

The second variant is the agent's **reactions to the user's non-compliance** with regard to their goals: in one dialogue, the agent acknowledges the importance of the values

behind the action, while in the other dialogue, the agent suggests an alternative and asks if this is a one-time exception to know if the user model should be updated. The third variation relates to the **dialogue initiation**, specifically the timing and the initiator. The complete material can be found in the Supplementary.

4.4. Procedure

Due to the measures regarding Covid-19, both sessions were conducted online via Microsoft Teams. All sessions were video recorded for the purpose of making transcriptions. The recordings were deleted once they were transcribed. The participants were given vouchers worth 15 euros as a thank-you for their contributions. We used only reading material in the focus group, with no other physical prompts. The session lasted around 1.5 hours. Therefore, we believe the online setting was appropriate.

At the beginning of the sessions, the overall objective of the study was explained to the whole group: to explore how end-users prefer to discuss misalignment with support agents. The material was shared with the participants while going through each page as per the moderator's instructions. In the first part, each participant was asked about their personal attitudes toward support agents. In the second part, the participants read misalignment scenarios and alignment dialogues, and were then asked to compare and discuss them by imagining themselves as the persona in the scenarios. This continued until all six scenarios were discussed. To guide the participants, discussion questions were prepared.

- Which version of the dialogue do you prefer, or which part of which dialogue do you prefer? Why?
- Is there a certain part of the dialogues that you particularly like/not like? Why? How would you want to do it instead?

Furthermore, we are interested in how users felt about their relationship with the agent after engaging in an alignment dialogue. Although the participants did not interact with a system, we asked them to answer the questions as if they were the user in the presented dialogues. Users' perception of the agent can ultimately affect the agent's effectiveness and resultant user behaviors. Our questions were inspired by the autonomous agent teammate-likeness (AAT) model, which aims to understand humans' perceptions of their intelligent partners.

- Which dialogue is more 'intelligent', as in has more capability in providing support? (related AAT construct: perceived agentic capability)
- Do you feel one dialogue is more supportive than the other? (related AAT construct: perceived benevolent intent)
- After which dialogue do you think the agent would be more 'on the same page' as you? (related AAT construct: synchronized mental model)

The validity and reliability of the AAT are not applicable due to the modifications to open questions. However, the primary focus of this study is to gather participants' attitudes and the underlying reasons for their responses, rather than to obtain ratings of the statements. Thus we used AAT only as an inspiration for preparing the discussion and did not rely on its predefined constructs in our data analysis.

4.5. Data analysis method

We transcribed the focus group sessions and analyzed the transcriptions using qualitative data analysis methods. Qualitative data analysis is sometimes criticized for being subjective and lacking reproducibility and generalizability [29,30]. However, when it comes to understanding people's beliefs, attitudes, and values, a qualitative approach may be more appropriate than quantitative methods [29]. As our study aims to uncover the reasons behind individuals' opinions on engagement in alignment dialogues, we believe that a qualitative approach is suitable.

In our study, we seek to understand how users want to engage in alignment dialogues and what are the reasons behind their opinions *grounded* in the data. To achieve this, we chose inductive thematic analysis as our analysis method where the themes identified are strongly linked to the data themselves [31] without trying to fit it into a pre-existing coding frame. This form of thematic analysis is similar to the 'lite' version of grounded theory [32]. We incorporated the coding stages from [33] and [34] as they provide a clear series of steps and descriptions of how each step takes place [35].

1. Familiarization: at this step, we familiarized ourselves with the data by reading the full transcriptions several times in an effort to immerse in the details and get a sense of the interview as a whole before breaking it into parts [35].
2. Open coding: the data are chunked into small units and coded with a number of words that represent key points in the data.
3. Indexing and charting: the quotes are lifted from their original context and rearranged to prepare for the next step.
4. Axial coding: similar codes are grouped together to create categories from the open codes.
5. Selective coding: central categories that connect all the codes are identified.
6. Interpretation: emergent themes are linked and visualized. The focus is the relationship between the quotes, and the links between the data as a whole [35].

The coding results and the model were evaluated by an independent researcher who coded the passages with the coding schema and answered questions regarding the terminology used, consistency, completeness, and the grouping of the codes. In instances where two coders disagreed on the coding of a passage, we engaged in further discussion to arrive at a consensus.

We followed the guidance from [36] to ensure the credibility, transferability, dependability, and confirmability of our study. We used data source triangulation (literature, potential users) to validate the credibility of our findings [37,38]. Transferability and dependability assessments are supported with detailed descriptions of the research methods. Confirmability was ensured through a pilot study and a second coder [39].

5. Results

During the study, we asked participants to imagine themselves as if they were the users having an alignment dialogue with a support agent. In the following sections, we use the term 'participants' when referring to the opinions of those participating in the study, and 'user' when referring to their envisioned role in the human-agent dialogue.

5.1. Dialogue variants and how they are perceived differently

Prior to conducting the qualitative analysis, we examined participants' responses regarding the different variants of alignment dialogues (discussed in Section 4.3) by contrasting what has been said about different variants of dialogues. With regard to the first variation - **the depth of reasoning** - the majority of participants did not prefer the dialogue in which the agent probed further. They found it annoying or missing the point, even if they understood its purpose. In terms of the **reactions to the user's non-compliance**, participants preferred when the agent offered suggestions rather than simply accepting non-compliance. However, participants disliked user model-related conversations, finding them passive-aggressive or sarcastic when values were acknowledged, and weird when asked about the incident being a one-time thing. With respect to the **initiation of the dialogue**, participants preferred the agent to initiate the dialogue but with an option for the user to give input. The timing of initiation varied depending on the situation and severity of the outcome.

It is important to note the aforementioned summary is not intended to provide a conclusive "solution" as to which variant is better. As a result of the qualitative nature of the study, we did not conduct a quantitative analysis. Nevertheless, the observations were intended as an initial exploratory first step to identify which types of considerations and dimensions we need to take into account when designing alignment dialogues.

5.2. Tree of codes

We used QSR NVivo [40] to perform the qualitative analysis. First, we derived a preliminary coding schema from a thorough reading of the material (*Step 1* in Section 4.5). In the second round of analysis, we annotated each piece of text with appropriate codes (*Step 2*), and grouped relevant codes together, resulting in a tree of codes.(*Step 3 & Step 4*).

Figure 2. Final tree of codes. Due to space constraints, the codes at the lowest level are displayed side by side using semicolons.

The tree of codes is shown in Figure 2. At the highest level, there are five categories. These categories are groupings of codes that together represent the main elements emerging in alignment dialogues:

- **Agent utterances** include the codes with regard to the agent's support actions or utterances.
- **Use of language** focuses on how a sentence or a piece of information is expressed.
- **User's opinions & feelings** cover a rather broad theme that consists of the user's feelings or opinions arising in the alignment dialogues, such as the agent being annoying or the user feeling guilty.

- **User characteristics** represent attributes of the user that could play a role in human-agent alignments, such as their personality or personal preference.
- **Additional functionality** refers to functions desired by participants, such as recommendations for activities or integration of menstrual cycles for exercise advice.

5.3. Connections between the themes

In the last stage of analysis, we explored the relationships between categories(*Step 5* and *Step 6*). We queried the data with all the combinations of codes. The main intersections were found between *user's feelings* and *agent utterances*, *user's feelings* and *the use of language*, and *user's feelings* and *user characteristics*. We further looked at the quotes that contained these combinations of themes. Some example quotes are presented in the Supplementary. By examining the quotes containing these intersections, we gained insights into how the categories are related and the potential impact of alignment dialogues on the user, as shown in Figure 3. The conceptual model explains how the different aspects relate to each other, focusing on *why* people have certain views and opinions towards the dialogue content, and *how* it has an impact on the user.

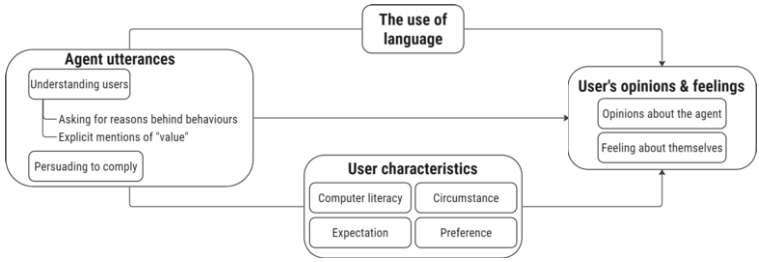

Figure 3. Overview of how alignment dialogues may affect user feelings.

5.3.1. Relationship between agent utterances and user's opinions and feelings

The content of alignment dialogues includes interactions aimed at understanding the user and persuading them to comply with their goals, as explained in Section 3. Participant responses indicate that these agent utterances could have an impact on the **user's opinions and feelings**, which can be further classified into *user's feelings about themselves* and *how they think about the agent*. It is important to distinguish between these two types of feelings as the objects to which the feelings are directed are essentially different, and separating them helps to gain deeper insights into their underlying causes. For instance, the dialogue may lead the user to feel guilty, or the user may perceive the agent as unsupportive. This differentiation is critical in comprehending the factors driving these feelings and in elucidating their interplay with the alignment dialogue.

Regarding the *user's feelings about themselves* component, it is worth noting that particular events can trigger various emotions. For instance, when the agent uses comparative language to describe the user's decision (see [Q1] in the Supplementary), it is likely to generate a feeling of being judged. Similarly, negative emotions arise when the agent highlights the user's non-compliance behaviors [Q2]. When the agent asks the user about their values based on its observations of their behavior, participants indicated they would feel confused, disoriented, or annoyed [Q3].

At first glance, one may question why the alignment dialogues were designed to elicit negative emotions in participants. However, they were not created with such intention in mind, except for Scenario 6. Reflecting on the results, we identify two underlying facets of the nature of behaviour support agents that could make them prone to evoking some negative feelings in the user: 1) the agent's role, at least at times, is to address non-compliance behaviours of the user; 2) the agent needs to support the user towards their goals even if it conflicts with their short-term desires.

These findings about alignment dialogues potentially triggering negative emotions in users align with existing research in psychology and behavior change/support. The presence of alternatives that simultaneously cue both long-term goals and short-term desires can lead to conflicts [41]. In such conflict scenarios, either option (i.e. compliance or non-compliance) will inevitably elicit both positive and negative emotions [42]. Additionally, [43,44] suggest that friction in interactive technologies is necessary to make people stop, think, and ultimately change.

5.3.2. Factors that moderate human-agent relationship

We observed that there are additional factors that could play a role in the relationship between the dialogue content and the user's opinions or feelings: the use of language and user characteristics.

Regarding **the use of language**, the participants' responses indicate that the way a sentence or a piece of information is expressed could have a significant impact on both the user's feelings about themselves and about the agent [Q1][Q4]. This is in line with [45], which suggests that the use of language in persuasive technologies sets the stage for outcomes. [46] also demonstrated the effect of source credibility and message framing on promoting physical exercise. It is noteworthy that our findings revealed that users perceived mentions of personal values by the agent as negative [Q5], regardless of whether the intention was to acknowledge the values or verify their accuracy. Users perceive such references as passive-aggressive or sarcastic and express their feelings of being judged.

User characteristics can explain how various aspects of users influence their feelings towards the agent. Although we did not ask the participants for their characteristics, we deduced four user characteristics from their quotes. The first one is the user's *computer literacy*. Participants who lacked computer literacy and did not understand the relevance of certain questions experienced confusion [Q6], while those with more knowledge did not report confusion but did express concerns about the user-friendliness of the agent [Q7].

The second characteristic is the *user's expectation* [Q8][Q9]. [47] has identified the need for "high performing, smart, seamless and personal" agents. However, the reality does not live up to these expectations. This coincides with [48]'s findings on the dissonance between user expectations and their assessment of the intelligence of the conversational agents. Norman's Gulf of Execution [49] illustrated this mismatch between the user's intentions and the allowable actions. The smaller the gulf, the more satisfying the user experience.

The third characteristic is *personal preferences.* Throughout different focus groups discussing different scenarios, it was repeatedly emphasized that personal preference plays a role in human-agent interaction. As one participant expressed, "What if we can choose how we are spoken to," highlighting the importance of personalization. This res-

onates with a large body of research on personalization. For behaviour support, personalization plays an important role as effective strategies are likely to depend on user characteristics [50,51,52].

The last user characteristic is the *user circumstance* [Q10]. For instance, the agent gives a reminder when the user is not available, and then the user forgets, which in turn makes them think the agent is not useful. There is vast research on modelling and reasoning about context and situation, e.g. [53,54]. Behaviour support agents need to understand a user's situation to provide comprehensive support [55]. By improving the agent's context awareness, the richness and usefulness of the agent increase as well [56].

6. Discussion & Conclusion

Limitations The limited number and similar age of the participants means there are some limitations regarding the generalization of the results. Moreover, written dialogues were used which means that participants did not interact with a dialogue agent personally. Therefore, the results cannot be interpreted as yielding general a theory of how alignment dialogues affect users. Rather, this study is intended as a first step in the design process of alignment dialogues, and our results provide directions for further investigation in our next steps.

Discussion We have observed that the participants' attitudes can be negative regarding parts of the dialogue where the agent tries to ask about abstract, broader reasons behind the user's actions such as their values and what is important to users in general. However, it is not yet clear what the reasons for this are. It could be that participants do not like the parts of the dialogue in which the agents ask about values due to their abstract nature, that the unfamiliarity with this type of conversation causes misunderstandings, that the conversation becomes too deep or personal too quickly, or that the timing is wrong. Moreover, it could be that asking the user for an *explanation* of the reasons behind certain choices of action is perceived by participants as the agent asking for a *justification*. This might be reinforced by the dialogue-based setup, which could invoke the perception of the agent as a social other' with opinions about the behavior of the user.

Contributions and future work The research on alignment dialogues is still in its early stage. This study introduces and sharpens the notion of 'alignment dialogue' and sheds light on what is needed for future development and research on an interactive approach to human-machine alignment in support agents, in particular regarding the potential effects of the dialogues on users' feelings. To further understand these effects, it is essential to conduct qualitative and quantitative user studies where the participants are experiencing the dialogues as they unfold, as opposed to reading pre-written dialogues.

Acknowledgement This work is part of the Hybrid Intelligence Gravitation Programme, with project number 024.004.022, which is financed by the Netherlands Organisation for Scientific Research (NWO).

Supplementary The scenario and dialogue materials used in the focus group, and example quotes are available at 4TU.ResearchData. https://doi.org/10.4121/b7a321df-640a-483d-8c32-a18fe21e7204

References

[1] Milić E, Janković D, Milenković A. Health care domain mobile reminder for taking prescribed medications. In: International Conference on ICT Innovations. Springer; 2016. p. 173-81.

[2] Schoffman DE, Turner-McGrievy G, Jones SJ, Wilcox S. Mobile apps for pediatric obesity prevention and treatment, healthy eating, and physical activity promotion: just fun and games? Translational behavioral medicine. 2013;3(3):320-5.

[3] Wabeke TR. Recommending tips that support well-being at work to knowledge workers. 2014.

[4] Van Riemsdijk MB, Jonker CM, Lesser V. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In: Proceedings of the 2015 international conference on autonomous agents and multiagent systems; 2015. p. 1201-6.

[5] Tielman ML, Jonker CM, van Riemsdijk MB. What should i do? deriving norms from actions, values and context. In: MRC@ IJCAI; 2018. .

[6] Kließ MS, Stoelinga M, Riemsdijk M. From Good Intentions to Behaviour Change. In: International Conference on Principles and Practice of Multi-Agent Systems. Springer; 2019. p. 354-69.

[7] Cranefield S, Winikoff M, Dignum V, Dignum F. No Pizza for You: Value-based Plan Selection in BDI Agents. In: IJCAI; 2017. p. 178-84.

[8] Ajmeri N, Murukannaiah PK, Guo H, Singh MP. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems; 2017. p. 230-8.

[9] Berka J, Balata J, Jonker CM, Mikovec Z, van Riemsdijk MB, Tielman ML. 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. 2022:1-17.

[10] Armstrong S, Mindermann S. Occam's razor is insufficient to infer the preferences of irrational agents. Advances in Neural Information Processing Systems. 2018;31.

[11] Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer; 2019.

[12] Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association. 2018;25:1248-58.

[13] Pickering MJ, Garrod S. Toward a mechanistic psychology of dialogue. Behavioral and brain sciences. 2004;27(2):169-90.

[14] Brennan SE, Clark HH. Conceptual pacts and lexical choice in conversation. Journal of experimental psychology: Learning, memory, and cognition. 1996;22(6):1482.

[15] Klimoski R, Mohammed S. Team mental model: Construct or metaphor? Journal of management. 1994;20(2):403-37.

[16] Mathieu JE, Heffner TS, Goodwin GF, Salas E, Cannon-Bowers JA. The influence of shared mental models on team process and performance. Journal of applied psychology. 2000;85(2):273.

[17] Van den Bossche P, Gijselaers W, Segers M, Woltjer G, Kirschner P. Team learning: building shared mental models. Instructional Science. 2011;39(3):283-301.

[18] Westli HK, Johnsen BH, Eid J, Rasten I, Brattebø G. Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design. Scandinavian journal of trauma, resuscitation and emergency medicine. 2010;18(1):1-8.

[19] Jonker CM, van Riemsdijk MB, Vermeulen B. Shared Mental Models. In: De Vos M, Fornara N, Pitt JV, Vouros G, editors. Coordination, Organizations, Institutions, and Norms in Agent Systems VI. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 132-51.

[20] Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M. Coactive Design: Designing Support for Interdependence in Joint Activity. Journal of Human-Robot Interaction. 2014;3(1).

[21] Gabriel I. Artificial intelligence, values, and alignment. Minds and machines. 2020;30(3):411-37.

[22] Kitzinger J. Qualitative research: introducing focus groups. Bmj. 1995;311(7000):299-302.

[23] Krueger RA. Focus groups: A practical guide for applied research. Sage publications; 2014.

[24] Schnaars SP. How to develop and use scenarios. Long range planning. 1987;20(1):105-14.

[25] Spaniol MJ, Rowland NJ. Defining scenario. Futures & Foresight Science. 2019;1(1):e3.

[26] Schwartz SH. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in experimental social psychology. vol. 25. Elsevier; 1992. p. 1-65.

[27] Rokeach M. The nature of human values. Free press; 1973.

[28] Friedman B, Kahn PH, Borning A, Huldtgren A. Value sensitive design and information systems. Early engagement and new technologies: Opening up the laboratory. 2013:55-95.
[29] Draper AK. The principles and application of qualitative research. Proceedings of the nutrition society. 2004;63(4):641-6.
[30] Mays N, Pope C. Qualitative research: rigour and qualitative research. Bmj. 1995;311(6997):109-12.
[31] Patton MQ. Qualitative evaluation and research methods. SAGE Publications, inc; 1990.
[32] Braun V, Clarke V. Using thematic analysis in psychology. Qualitative research in psychology. 2006;3(2):77-101.
[33] Strauss A, Corbin J. Basics of qualitative research techniques. Citeseer; 1998.
[34] Krueger RA. Focus groups: A practical guide for applied research. Sage publications; 1994.
[35] Rabiee F. Focus-group interview and data analysis. Proceedings of the nutrition society. 2004;63(4):655-60.
[36] Lincoln YS, Guba EG, Pilotta J. Naturalistic inquiry. Beverly Hills. CA: Sage Publications Lee, WS (2001) Parents divorce and their duty to support the expense of bringing up their child Asian Women. 1985;13(1):85-105.
[37] Carter N. The use of triangulation in qualitative research. Number 5/September 2014. 2014;41(5):545-7.
[38] Patton MQ. Enhancing the quality and credibility of qualitative analysis. Health services research. 1999;34(5 Pt 2):1189.
[39] McDonald N, Schoenebeck S, Forte A. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proceedings of the ACM on human-computer interaction. 2019;3(CSCW):1-23.
[40] (2020) QIPL. NVivo (released in March 2020); 2020. Available from: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.
[41] Fishbach A, Zhang Y. Together or apart: when goals and temptations complement versus compete. Journal of personality and social psychology. 2008;94(4):547.
[42] Ozkaramanli D, Özcan E, Desmet P. Long-term goals or immediate desires? Introducing a toolset for designing with self-control dilemmas. The Design Journal. 2017;20(2):219-38.
[43] Laschke M, Diefenbach S, Hassenzahl M. "Annoying, but in a nice way": an inquiry into the experience of frictional feedback. International Journal of Design. 2015;9(2).
[44] Laschke M, Diefenbach S, Schneider T, Hassenzahl M. Keymoment: Initiating behavior change through friendly friction. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational; 2014. p. 853-8.
[45] Fogg BJ. Persuasive technology: using computers to change what we think and do. Ubiquity. 2002;2002(December):2.
[46] Jones LW, Sinclair RC, Courneya KS. The effects of source credibility and message framing on exercise intentions, behaviors, and attitudes: An integration of the elaboration likelihood model and prospect theory 1. Journal of applied social psychology. 2003;33(1):179-96.
[47] Zamora J. Rise of the chatbots: Finding a place for artificial intelligence in India and US. In: Proceedings of the 22nd international conference on intelligent user interfaces companion; 2017. p. 109-12.
[48] Luger E, Sellen A. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In: Proceedings of the 2016 CHI conference on human factors in computing systems; 2016. p. 5286-97.
[49] Norman D. The design of everyday things: Revised and expanded edition. Basic books; 2013.
[50] Masthoff J, Vassileva J. Tutorial on personalization for behaviour change. In: Proceedings of the 20th international conference on intelligent user interfaces; 2015. p. 439-42.
[51] Bonneux C, Dendale P, Coninx K. Investigating Motivations and Patient Profiles for Personalization of Health Applications for Behaviour Change. In: Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments; 2022. p. 146-54.
[52] Celis-Morales C, Livingstone KM, Marsaux CF, Macready AL, Fallaize R, O'Donovan CB, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4me European randomized controlled trial. International journal of epidemiology. 2017;46(2):578-88.
[53] Bettini C, Brdiczka O, Henricksen K, Indulska J, Nicklas D, Ranganathan A, et al. A survey of context modelling and reasoning techniques. Pervasive and mobile computing. 2010;6(2):161-80.
[54] Ye J, Dobson S, McKeever S. Situation identification techniques in pervasive computing: A review. Pervasive and mobile computing. 2012;8(1):36-66.
[55] Kola I, Jonker CM, van Riemsdijk MB. Who's that?-Social situation awareness for behaviour support

agents: A feasibility study. In: Engineering Multi-Agent Systems: 7th International Workshop, EMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers. Springer; 2020. p. 127-51.
[56] Abowd GD, Dey AK, Brown PJ, Davies N, Smith M, Steggles P. Towards a better understanding of context and context-awareness. In: Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1. Springer; 1999. p. 304-7.