Reason Explanation for Encouraging Behaviour Change Intention

Amal Abdulrahman
Computing Department, Macquarie University
Sydney, NSW, Australia
amal.abdulrahman@students.mq.edu.au

Deborah Richards
Computing Department, Macquarie University
Sydney, NSW, Australia
deborah.richards@mq.edu.au

Ayse Aysin Bilgin
Department of Mathematics and Statistics, Macquarie University
Sydney, NSW, Australia
ayse.bilgin@mq.edu.au

ABSTRACT

The demand for intelligent virtual advisors in our rapidly advancing world is rising and, consequently, the need for understanding the reasoning process to answer why a particular piece of advice is provided to the user is directly increasing. Personalized explanation is regarded as a reliable way to improve the user's understanding and trust in the virtual advisor. So far, cognitive explainable agents utilize reason explanation by referring to their own mental state (beliefs and goals) to explain their own behaviour. However, when the explainable agent plays the role of a virtual advisor and recommends a behaviour for the human to perform, it is best to refer to the user's mental state, rather than the agent's mental state, to form a reason explanation. In this paper, we are developing an explainable virtual advisor (XVA) that communicates with the user to elicit the user's beliefs and goals and then tailors its advice and explains it according to the user's mental state. We tested the proposed XVA with university students where the XVA provides tips to reduce the students' study stress. We measured the impact of receiving three different patterns of tailored explanations (belief-based, goal-based, and belief&goal-based explanation) in terms of the students' intentions to change their behaviours. The results showed that the intention to change is not only related to the explanation pattern but also to the user context, the relationship built with the agent, the type of behaviour recommended and the user's current intention to do the behaviour.

KEYWORDS

Explainable agents; Personal virtual advisor; Reason explanation; Behaviour change intention; Working alliance; Trust

ACM Reference Format:
Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2021. Reason Explanation for Encouraging Behaviour Change Intention. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 10 pages.

1 INTRODUCTION

Intelligent agents have become more acceptable in our world in various fields such as health [28], education [5] and marketing [34]. To increase the acceptability and efficiency of intelligent agents, those agents must be transparent by explaining their behaviour. The importance of building explainable agents (XAs) comes from its role in building human-agent trust [18] which plays an important role towards achieving the system goals (e.g. behaviour change [24]). How an XA can communicate the reasoning behind its behaviour is an important open problem for believable and acceptable XA [35].

The vast body of research in the human-agent interaction for behaviour change field is built on theories and findings from the social sciences and it is reasonable to build XAs that mimic the natural human method of explanation. Earlier, Dennett [14] stated that, similar to a human being, an agent can be represented with three stances: physical, design, and intentional stances. While the physical and design stances refer to the hardware and software that construct an artificial agent as an entity, the intentional stance is the rational cognitive representation of the agent which can explain and predict the agent's current and future actions [15]. People explain their intentional behaviours by referring to their mental state (i.e. beliefs and desires/goals), which is called reason explanation [37]. Malle [37], further, emphasised the importance of the use of grammatical markers to refer to beliefs: "*I think/believe...*", goals: "*I want...*", and to signal subjectivity: "*I/He/She think(s)/want(s)...*", especially when explaining the behaviours of others.

Inspired by the Belief-Desire-Intention (BDI) model by Bratman [9], BDI agents have been introduced which include beliefs and goals besides intentions as the main components to drive the agent's actions [2]. The design of BDI agents facilitates the implementation of XAs that use reason explanation to explain their intentional behaviours. This ability of BDI agents is important because human users regard the agent's behaviour as intentional behaviour and they expect to receive a similar explanation from the virtual agent to what they receive from humans when they explain their intentional behaviour [12].

The explanation process is a social process undertaken in a conversational form to close a gap in understanding between the explainer and the explainee [40]. The gap could be the transferred knowledge itself or the inference of the provided knowledge [51]. According to the conversational model by Hilton [26], an explanation must appropriately answer the *why* question and must be relevant to the explainee. Explanation relevancy could be connected to providing the relevant reasons behind the action (beliefs and/or goals) [23, 30] or the relevant information to the explainee's context [49]. However, an agent may derive its behaviour as a result of a series of beliefs and goals, and including all of them in the explanation will generate a long and irrelevant explanation [27]. Selecting the relevant knowledge or the elements to build the reason explanation, using beliefs and goals is challenging.

In this paper, we contribute to the body of research to find relevant explanation patterns that could persuade a user to adopt

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

healthier behaviours. Explanations provided by XAs typically focus on the agent's beliefs and goals and do not consider the user's beliefs and goals. In contrast, we distinguish between the agent's behaviour and the user's behaviour and believe that explanation should include the user's beliefs and goals when the behaviours are required to be performed by the user, not the agent. Hence, in this paper, we investigate the following question: "*How do agent's explanations that refer to the user's beliefs or goals influence the user's intention to change the behaviours recommended by the agent?*".

In Section 2 we review the most related work to this research that leads to form our research hypotheses. In Section 3, we present how we designed our research to test the hypotheses and the methods used. Section 4 presents the experimental results followed by a discussion in Section 5 and conclusion in Section 6.

2 LITERATURE REVIEW

The use of a virtual agent as a virtual advisor/therapist is compelling in the health domain to provide advice and/or support [7, 44]. Over the past decade, several virtual conversational agents have been developed to improve the lifestyle of users with health problems such as obesity and mental health [31].

Virtual agents can build rapport and trust with users via the inclusion of relational and emotional cues [20]. However, such cues have not been proven to be effective in behaviour change [48]. Motivating users towards healthier behaviour change is found to be more effective when the recommendation messages are properly designed and personalized according to users' preferences [4].

Wheeler et al. [52] found a link between the intention to change a behaviour and the delivered message, i.e. the recommendation to change. Interestingly, they found that people are more likely to be persuaded to change their behaviour when the delivered message forms a link between their own cognition (e.g. beliefs and context) and the recommended behaviour. Persuading a user to change a behaviour is more effective when the motivation is internalized which has been shown to occur when the persuasion attempt is aligned with the user's cognitive state: beliefs and goals [21]. In human-human interaction, it is common to refer to our cognition, including beliefs, goals and intention, to explain our actions [37]. Similarly, virtual agents can use their cognition to explain their behaviour (e.g. Harbers et al. [23]). Proper explanation improves user-agent understanding and trust and, consequently, increases the user's intention to follow the advice recommended by the virtual advisor [39]. However, so far, the introduced explainable virtual agents rely on their own cognition rather than the user's cognition which undermines the concept of personalizing the delivered recommendation to change a behaviour according to the user's thoughts and reasoning processes. In general, personalization and the use of user models in the field of explainable virtual agents is still very limited: about 8% of the current work [3].

2.1 Explainable Agents

Explainable artificial intelligence (XAI) has gained importance with the advancement in automated and persuasive systems. The majority of work in XAI has been done to provide explanations for data-driven systems such as machine learning [46]. Little work, but rapidly increasing, has been done in the area of explainable agents (XA), particularly, goal-driven agents [3]. People perceive virtual agents as social entities and they respond to them socially as they do with other humans [41]. They expect these virtual agents to have a mental state that derives their behaviour and, thus, they expect agents to be able to explain their behaviours [12]. Hence, it is best to build XAs that can mimic the ways humans explain their behaviour to others which is commonly done by referring to their mental state [37]. Harbers et al. [23] found such explanations have been well received by users in terms of understandability. They, further, found that different explanation patterns, referring to beliefs or goals, should be delivered according to the agent's type of action.

BDI agents are built to mimic the human cognitive reasoning process using beliefs, desires and intentions [43]. With these elements, an agent derives its actions, and consequently, explains them. The beliefs are the context or the knowledge of the agent about its environment; the goals are the objectives the agents can achieve through possible stored plans in its memory, and the intentions are the plans the agent is currently committed to perform. A BDI agent triggers an action based on its beliefs and/or goals which could be represented using a goal hierarchy tree (GHT) such as the GHT in [22]. In any GHT, the agent's main goal is placed at the root of the tree and it can be achieved through one or more sub-goals (the branches of the tree) that could be achieved in a sequential or hierarchical order. The leaves of the tree represent the agent's actions. For the agent to perform an action, some conditions must be attained. These conditions are the agent's beliefs and all the beliefs above the action have to be true and, consequently, the goals/sub-goals above are achievable.

Harbers et al. [23] tested the usefulness of four different patterns of explanation according to GHT using: the goal one-level above the action, the goal two-levels above the action, the belief(s) above the action, and the next/previous goal and action following the current action depending on the place of the current action on the GHT. Twenty non-expert new firefighting trainers evaluated the four types of explanations within a training scenario. Trainers preferred belief-based explanation to explain the agent's behaviour when only one action or conditional action(s)/goal(s) were adopted. Goal-based explanations were more preferred but not significantly over the belief-based explanations in procedural actions: a sequence of actions/sub-tasks will be performed by the agent. However, expert firefighters in a similar scenario preferred goal-based explanations. [50].

Kaptein et al. [30] reported a difference in adults' explanation preferences compared to children. They designed a robot to educate children with Type 1 diabetes when they are in a good mood; with the ability to cheer them up, first, if they are not in a good mood. Using GHT, they utilized the beliefs and goals that are directly above the current actions to design the explanations. With 19 children and 19 parents, they found that both children and adults preferred goal-based explanation over belief-based explanation but the preference of goals over beliefs was significantly greater in the adults' group.

The findings of the above-mentioned studies provide evidence of the importance of the user's profile in designing the explanation pattern using goals and beliefs. However, the introduced XAs with GHT utilized the beliefs and goals of the agent in the reasoning and explanation as well. They did not take into account the human

user's beliefs and goals. This use of the agent's beliefs and goals in designing explanation patterns could be acceptable when the agent's actions are related to the agent's environment. However, such explanations may not be perceived as relevant by the user when the agent is a personal assistant or virtual advisor and the actions should be performed by the user.

Thus, using the user's beliefs and goals to create the explanation patterns, we formed our first hypothesis as follows:

Hypothesis 1. *(**H1**) There is a difference in terms of intention to change a behaviour between users who receive belief-based explanation and goal-based explanation.*

While explanation facilitates the transference of knowledge, it is critical to select the proper knowledge to transfer. Kulesza et al. [33] found that a complete explanation that describes the entire decision process is more important than a simple explanation. However, when an agent explains all of its underlying process, the explanation would include many beliefs and goals which could be irrelevant. Walton [51] asserted that the explanation should close a small gap of understanding and not be too lengthy. Long explanations could highly increase the cognitive load and lose its importance [40]. Thus, we are investigating the impact of extending the explanation pattern by providing explanations based on both goal and belief:

Hypothesis 2. *(**H2**) Users who receive longer explanations including both belief and goal will show less intention to change a behaviour compared to those who receive belief only or goal only based explanation.*

As discussed above, the influence of a type of explanation could be linked to the user profile, hence:

Hypothesis 3. *(**H3**) The differences, if any, in change in intention to do a behaviour between users who receive belief-based explanation, goal-based explanation or both are associated with the users' profiles*[1].

In the health domain, a main predictor of adherence is the therapist-patient relationship. This relationship is commonly called therapeutic or working alliance (WA) and it is achieved when the patient and the therapist engage in a positive relationship discussing and reaching an agreement on the therapy outcome (the goal) and how to achieve it (the task) [8]. Therapist-patient WA was associated with significant reduction in stress [11], adherence, satisfaction and quality of life [6]. Therefore, it is of interest to investigate if different explanation patterns can build different levels of user-agent relationship:

Hypothesis 4. *(**H4**) The three types of explanation will develop different user-agent relationships measured by WA.*

3 METHODOLOGY

As above-mentioned in the literature review, the use of BDI agents facilitates explaining the agent's actions using its beliefs and goals. This is because BDI agents use their cognitive mental state in the reasoning process. Hence, to answer the research question, we extend and evaluate the BDI-based cognitive agent architecture FAtiMA (Fearnot AffecTIve Mind Architecture) [16] as described next.

[1]We use the word profile to include the user's personal details (e.g. age, gender and personality) and his/her context (e.g. study aims, upcoming exams, study stress level).

3.1 Agent Architecture

FAtiMA is an agent architecture that allows the agent to logically reason about its actions according to its emotional and cognitive state. Because we are interested in evaluating the impact of explanation only, we have disabled the emotional appraisal component to control the experiment environment. The agent's emotions have a great influence on the agent-user relationship [20, 36]. Thus, we are using only the cognitive part of FAtiMA and extending it for more tailored reasoning and explaining process.

As a conversational agent, the agent communicates with the user through a designed dialogue. The agent perceives the user through the multi-choice answers available for the user to choose from. The agent interacts with the user by uttering the sentences produced from the agent's reasoning process. Originally, besides the emotional appraisal component, FAtiMA includes the agent's memory where the agent's beliefs and general knowledge are stored. The action selection component takes mainly the agent's beliefs and logically processes them to adopt new goals and trigger the proper actions. In order to tailor the agent's actions towards the user's beliefs and/or goals and to refer to them in the explanation process, we added a user model in the agent's memory where the user's information (belief's and goals) are stored.

In the action selection component, an explanation engine is added that combines the user's beliefs and/or goals that triggered the current action and relevant knowledge. More details about the extended FAtiMA and explanation engine are presented in a previous work [1].

3.2 Agent Dialogue Design

We designed a virtual advisor (VA), Sarah, to encourage university students to follow healthy behaviours shown to correlate with study stress [45]. Those recommended behaviours were designed carefully by specialists in the university Well-being Service Centre and are usually delivered as a pdf or text on the university website.

The agent starts the conversation by introducing herself, welcoming the user and introducing her goal of the interaction: providing some tips to help manage study stress. The dialogue ends with a farewell conversation. Similar to the concept of GHT, the agent's actions (recommendations) are designed to be triggered when a particular belief(s) and/or goal(s) of the user are attained. The agent's main goal is to reduce study stress which includes a series of sub-goals (i.e. recommended activities). In this study, the agent recommends three activities: participating in a study group, doing regular physical activities and meeting new people. Those three behaviours have been found to be more difficult to change in university students [1]. When the agent adopts a sub-goal (to recommend an activity), it enters into a dynamic conversation with the user to elicit his/her beliefs and/or goals to find the relevant recommendation to this user according to the available information. After that, the user asks the agent the *why* question before he/she receives an explanation. Figure 1 presents an example of the conversation between Sarah and a student who believes that studying in a group is wasting time. The student selected this belief from a list of various possible beliefs about studying in a group.

Explanation provides the user with two aspects: the relevancy of the recommendation to the user by stating the user's belief(s) and/or

Sarah: Let's start talking about the first tip which is studying in an online group. What do you think about participating in a study group in general?
Student: Study in a group is wasting time.
Sarah: That can happen if you don't find the right study group.
…
Sarah: May I recommend you to start looking for an online group if you are not in a group yet? You may ask your friends to form one or post a message on iLearn looking for a group.
Student: Can you explain why you are suggesting this?
Sarah: You find studying in a group could waste your time, but, if you allow me, I still would like to recommend you to join a study group. Depending on the size of the study group, it can be incredibly difficult coordinating a time and place that works for everyone. However, there are many other ways to overcome this problem. As for example, you may form an online study group…

Figure 1: Belief-based XA-student conversation snippet

goal(s) and extra information relevant to this context explaining how the user could follow the recommendation and how it could help achieving his/her goal [26, 49]. Following Malle [37], the agent uses the phrase "*you think/find...*" to refer to the user's beliefs and the phrase "*you want to.,.*" to refer to the user's goals.

3.3 Study Design

We designed one VA (Sarah) with three types of settings: belief-based explanation, goal-based explanation, and belief&goal-based explanation. In the three settings, the agent chats with the user to elicit the user's beliefs and goals, recommends the same recommendations in a similar order, but it uses different explanation patterns according to the enabled setting. The XVA was designed using the Unity3D game engine and integrated with FAtiMA. The participants have been recruited through the university channel and participation was completely optional. The students were granted course credit upon completing the study. The study was announced as an online study where the students were able to finish it anytime and anywhere.

Before the interaction, the students received a consent form, and series of questionnaires covering demographics (age, gender, culture), study (achievement aim, if having exam in the following two weeks, course and year of study), personality, propensity to trust and behaviour intention. In the behaviour intention questionnaire, the participants have been asked to rate their intention to do the three activities on 5-point Likert scales (from never to always). Before interacting with the XVA, the students were asked to indicate their emotional feeling towards their studies (stress level) on a scale 0: extremely relaxed to 10: extremely stressed. The scale is designed following the subjective units of discomfort scales (SUD) [47]. Participants are then asked to interact with the XVA. After the interaction, they were asked again to score their study stress level and complete the behaviour intention questionnaire. Moreover, they completed the trust and WA questionnaires.

Although there are several theories to describe personality, the big 5 factors model is a widely used and well regarded personality model [19]. The model is comprised of five factors: extraversion, agreeableness, conscientiousness, openness to new experiences and emotional stability. We used the brief questionnaire developed by Gosling et al. [19], called ten-item personality inventory (TIPI), comprising 10 items to measure the five traits using 7-point Likert scales from strongly disagree to strongly agree.

To measure the agent-user relationship, we utilized two questionnaires: trust and WA questionnaires. The WA inventory [25] is a common measurement of the therapist agent-user relationship; however, the built alliance could be a result of a user tendency to trust others in general (trait-like alliance) or of the therapy process (state-like alliance) [54]. Therefore, we included the trust questionnaire which is adapted from Mayer and Davis [38] to measure the propensity to trust others in general besides trust and trustworthiness sources: ability, benevolence and integrity. The WA questionnaire is the short form of working alliance inventory [25] that measures three elements of WA: task, goal and bond. We also asked the participant to rate their liking for the XVA using one question: "I like the agent". The trust, liking the XVA, and WA questionnaires' items are measured using 5-point Likert scales: from strongly disagree to strongly agree for items of the trust questionnaire and liking the XVA, and from seldom to always for items of the WA questionnaire. Further, we provided the participants with the option "Not applicable" next to the scales to choose when they think the question is not applicable to the situation because human-human measures are not always perceived as appropriate for measuring human-agent interactions or relationships [42].

To see whether the explanations had a lasting effect, three weeks after completing the study, participants were sent an email invitation to complete a short follow-up survey containing the same behaviour intention questionnaire for the three behaviours. The surveys have been designed using Qualtrics and the data has been analysed using SPSS. Due to the use of ordinal data (Likert scales) in the questionnaires and the number of the participants in every group, we opted to use the non-parametric tests in analysing the data [13].

4 RESULTS

4.1 Participants

In total, 91 university students participated voluntarily in the study and were assigned randomly to one of the experiment groups: belief group (age: $mean = 26.00, median = 19.00, STD = 12.222$), goal group (age: $mean = 25.87, median = 19.00, STD = 11.471$), or belief&goal group (age: $mean = 27.67, median = 28.5, STD = 8.938$). About 70.3% of the participants were under 30 years of age and from different cultural backgrounds, mainly: 29.7% Oceania, 16.5% Northern-Western Europe, and 16.5% South-East Asian. About 50% of the participants completed the follow-up survey as shown in Table 1. The table also reports the personality test results for the participants in each group. There were no significant between-group differences in terms of participants' age or personality.

4.2 Study Stress

Participants showed statistically significant reduction in their study-related stress after interacting with the XVAs in the three groups.

Table 2 reports the means, standard deviations and the analysis results using Wilcoxon signed ranks (SR) test. The Kruskal-Wallis test reported no statistically significant difference in stress between the three groups before interaction, indicating a fair distribution of

Table 1: Participants distributions among the three groups and their personality stats

Setting	#Interacted with VA	Extraversion		Agreeableness		Conscientiousness		Openness to experiences		Emotional stability		#Completed followup
		M	*std*	*M*	*std*	*M*	*std*	*M*	*std*	*M*	*std*	
Belief	33	4.05	1.655	4.68	0.999	4.79	1.186	4.92	1.016	3.74	1.485	19
Goal	34	3.57	1.431	4.91	0.830	4.99	1.464	4.90	1.153	3.84	1.397	13
Belief&Goal	24	3.71	1.301	4.79	1.179	4.98	1.202	5.00	0.821	4.02	1.339	14

Table 2: Study stress stats

Group	Before interaction		After interaction		Wilcoxon SR test	
	Mean	*std*	*Mean*	*std*	*Z*	*p*
Belief	6.52	2.108	5.21	2.132	-3.543	**<.001**
Goal	5.74	2.416	4.21	2.293	-3.534	**<.001**
Belief&goal	5.58	2.020	4.50	2.187	-2.913	**<.01**

the participants among groups, and no significant between-group differences in terms of stress reduction.

4.3 User-agent Relationship

4.3.1 Trust. The participants in the three groups reported average propensity to trust others in general: $mean = 3.08$ with $std = .378$ for the belief group, $mean = 3.12$ with $std = .445$ for the goal group, and $mean = 3.18$ with $std = .344$ for the belief&goal group with no between-group significant difference, which confirms the fair distribution of the participants among the groups. The reliability of the trust questionnaire was high with Cronbach's $\alpha = .914$.

Table 3 reports the results of analysing the responses of the trust questionnaire. Kruskal-Wallis test reported no significant differences between the three groups in terms of ability, benevolence, integrity or trust. The table, further, presents the total number of times (in percentages) the participants responded to any item of the specified construct as not applicable (NA). Thus, the reported means of the constructs are calculated as the average of only the valid responses on the Likert scales.

4.3.2 Working Alliance. The statistics from analysing the WA questionnaire are presented in Table 3. The questionnaire reliability, Cronbach's α, was .960. Similar to the trust's statistics, the table presents the number of participants who selected the "not applicable" option, in percentages, for each construct. Kruskal-Wallis test reported no significant difference between the three groups for task, goal, or bond scales.

4.4 Behaviour Change Intentions

Table 4 presents the statistics of the participants' intentions to do the three behaviours before interacting with the assigned XVA, immediately after the interaction and three weeks later. The analysis reveals significant greater intentions to do the three behaviours after interacting with belief-based XVA and goal-based XVA compared to their intentions before the interaction. Participants who interacted with belief&goal-based XVA showed significant change in their intentions to do physical activity and to meet new people but not to join a study group. There were no between-group significant differences in the intention change of the three behaviours.

Analysing the responses of the participants to do the three behaviours after 3 weeks of the interaction revealed no significant change in the participants' intentions to do the behaviours compared to their intentions after interacting with the XVA immediately. The intention changes at both points (immediately after interaction and after 3 weeks) were significantly correlated for joining a study group (Spearman's $\rho = .437$ *at* $p = .002$), doing physical activity (Spearman's $\rho = .325$ *at* $p = .028$) and meeting new people (Spearman's $\rho = .443$ *at* $p = .002$).

To find if the changes in the intentions are related to the users' profiles, several tests have been conducted depending on the type of factors: age, gender, personality, and study stress level. About 55.8% of the participants were 20 years old and younger, hence, the participants were assigned into two groups: older than 20 years, and 20 years or younger. For participants aged older than 20 years, there was no significant difference between the three experiment groups in their intention changes with the three behaviours. For participants 20 years and younger, Mann-Whitney test reported significant differences in the change in intention to join a study group between belief and belief&goal groups ($Z = -3.021$ *at* $p = .003$), and between goal and belief&goal groups ($Z = -2.627$ *at* $p = .005$) where participants in both the belief group and goal group showed higher intentions to change than those in the belief&goal group. For the study stress factor, the stress level was moderately correlated with the change in intention to join a study group in the belief group only (Spearman's $\rho = .441$ *at* $p = .010$). No further association was found between the changes in the intentions and other factors: gender, personality, achievement aim (high: distinction and high distinction vs. low: credit and pass) and having exam (yes/no).

To study how the user-agent relationship impacts intention change, binary logistic regression was run to explore the factors that can explain the variance in the intention to change the three recommended behaviours. The models' outcomes (the intentions changes) were coded as 0 for negative and no change in the intentions to do the behaviours, and as 1 for positive changes. Two levels of binary logistic regression models were performed to explore if the source of the variation in intention change were related to the users' profiles only or also to the relationship built with the XVA. In the first level, the factors in the users' profiles were used as predictors including: age, gender, personality, stress level, study achievement aim, having exam or not, and the intention to do the behaviour before the interaction with the XVA. The models were statistically significant in predicting the intentions to do the three

Table 3: Trust and WA stats measured on 5-point Likert scales. NA stands for "not applicable" and indicates how many NA was reported, and VR stands for "valid response" on the Likert scales

Construct	Belief-based explanation				Goal-based explanation				Belief&goal-based explanation			
	NA	VR	*Mean*	*std*	NA	VR	*Mean*	*std*	NA	VR	*Mean*	*std*
Trust and trustworthiness												
Ability	0%	100%	3.45	0.667	0%	100%	3.59	0.846	2.5%	97.5%	3.46	0.671
Benevolence	0%	100%	3.12	0.919	0%	100%	3.36	1.009	4.2%	95.8%	3.30	0.765
Integrity	0%	100%	4.00	0.718	0%	100%	3.99	0.702	0.0%	100.0%	4.02	0.634
Trust	0%	100%	2.95	0.863	0%	100%	3.06	0.857	1.0%	98.9%	2.94	0.618
WA												
Task	2.3%	97.7%	2.67	0.936	11.0%	89.0%	2.93	1.239	5.2%	94.8%	2.80	0.918
Goal	5.3%	94.7%	2.45	0.96	12.5%	87.5%	2.90	1.195	8.3%	91.7%	2.59	1.096
Bond	19.7%	80.3%	2.62	1.176	16.9%	83.1%	2.92	1.213	24.0%	76.0%	2.69	1.162
Liking the XVA	15.2%	84.8%	2.68	1.124	5.9%	94.1%	3.03	1.307	16.7%	83.3%	3.00	1.124

Table 4: Behaviour change intentions stats

Activity	Before interaction		After interaction		Follow-up		Wilcoxon signed ranks test (before vs after interaction)	
	Mean	*std*	*Mean*	*std*	*Mean*	*std*	*Z*	*p*
Belief-based explanation group								
Study in a group	2.18	0.917	2.70	0.951	2.42	.902	-3.532	**<.001**
Do physical activity	3.03	1.262	3.48	1.121	3.79	.976	-2.879	**.004**
Meet new people	2.55	1.063	2.97	0.984	2.84	1.015	-3.300	**.001**
Goal-based explanation group								
Study in a group	2.24	0.89	2.62	0.779	2.54	1.050	-2.427	**.015**
Do physical activity	3.03	1.087	3.38	0.954	3.00	1.155	-3.207	**.001**
Meet new people	2.56	0.786	3.00	0.921	2.62	.768	-3.095	**.002**
Belief&goal-based explanation group								
Study in a group	2.29	0.859	2.46	0.977	2.29	1.069	-1.633	.102
Do physical activity	3.25	1.073	3.71	1.122	3.00	.961	-2.598	**.009**
Meet new people	2.54	0.658	3.13	0.68	2.93	.730	-2.841	**.005**

behaviours. For joining a study group ($\chi^2(5) = 16.820$, *at* $p = .005$, Nagelkerke $R^2 = .230$ (i.e. the explained variation in the dependent variable based on the model= 23%)), the significant source of variations in the intention changes were: the intention to do the behaviour (odd ratio (OR)=.507, 95% confidence interval (CI)=(.280–.919)), and stress level (OR=1.379, 95% CI=(1.069–1.779)). And for doing a daily physical activity ($\chi^2(3) = 14.319$, *at* $p = .003$, Nagelkerke $R^2 = .230$), the significant source of variation were the intention to do the behaviour (OR=.457, 95% CI=(.262–.795)), and agreeableness (OR=1.758, 95% CI=(1.009–3.065)). Finally, for meeting new people ($\chi^2(3) = 25.653$, *at* $p =< .001$, Nagelkerke $R^2 = .339$), the significant source of variations in the intention changes were the intention to do the behaviour (OR=.250, 95% CI=(.119–.523)), openness to experience (OR=2.099, 95% CI=(1.218–3.619)), and having upcoming exam (OR=3.083, 95% CI=(1.081–8.796)). In the Second level, the user-agent relationship factors (liking the XVA, trustworthiness, trust, and WA scales) were included as predictors in the regression models. The models were also statistically significant in predicting the intentions to: join a study group ($\chi^2(5) = 26.660$, *at* $p < .001$, Nagelkerke $R^2 = .401$ with 76.9% classification accuracy), do a daily physical activity ($\chi^2(5) = 28.094$, *at* $p < .001$, Nagelkerke $R^2 = .380$ with 72.7% classification accuracy) and meeting new people ($\chi^2(4) = 32.598$, *at* $p =< .001$, Nagelkerke $R^2 = .415$ with 70.5% classification accuracy). Table 5 presents the details of the second level regression model.

5 DISCUSSION

The main research goal of this study is to build a virtual advisor that tailors its advice and explains *why* this particular advice is given by citing the user's beliefs and goals instead of the agent's beliefs and goals. The influence of the use of belief-based explanation vs goal-based explanation or a combination of both were measured in terms of the system outcome: behaviour change intentions.

The XVA delivered three tips to the participating students to reduce their study-related stress. The results showed that students statistically significantly felt less stressed after interacting with the three versions of the XVA which indicates that students found the XVAs' advice relevant and helpful to them in study planning and reducing their stress. Students' comments at the end of the study confirm this conclusion such as *"It was very interesting to see the*

Table 5: The binary logistic regression models with user's profile and user-agent relationship scales as predictors

Predictor	B	Stand. error	Wald	df	p	$Exp(B)$	(95% CI)
Behaviour: join a study group							
Intention to do the behaviour	-.882	.354	6.210	1	**.013**	.414	(.207–.828)
Age	.045	.026	2.920	1	.087	1.046	(.993–1.102)
Task	.907	.277	10.708	1	**.001**	2.477	(1.439–4.264)
Constant	-3.378	1.593	4.495	1	**.034**	.034	-
Behaviour: do a physical activity							
Intention to do the behaviour	-1.003	.304	10.848	1	**.001**	.367	(.202–.666)
Agreeableness	.557	.294	3.582	1	.058	1.746	(.980–3.108))
Openness to experiences	.501	.301	2.770	1	.096	1.650	(.915–2.975)
Integrity	1.390	.659	4.448	1	**.035**	4.015	(1.103–14.612)
Trust	.923	.424	4.749	1	**.029**	2.517	(1.097–5.774)
Constant	-5.443	2.597	4.393	1	**.036**	.004	-
Behaviour: meet new people							
Intention to do the behaviour	-1.545	.402	14.778	1	**<.001**	.213	(.097–.469)
Openness to experiences	.670	.297	5.103	1	**.024**	1.954	(1.093–3.495)
Having exam (yes)	1.411	.590	5.719	1	**.017**	4.101	(1.290–13.040)
Trust	.952	.381	6.243	1	**.012**	2.591	(1.228–5.467)
Constant	-3.500	1.742	4.036	1	**.045**	.030	-

advice and responses to my own personal encounters with studying.", "This experiment was useful. It is reflecting on where I was in terms of my study schedule and plans. I will be using the tips provided within my daily routines.".

Participants in the belief group and goal group reported significant increase in their intentions to do the three behaviours recommended by the assigned XVA. The results failed to capture any between-group difference in terms of intention to change to support **H1**. The belief&goal group reported no significant change in their intentions to join a study group. Although this low intention to change was only found in one behaviour, it supports **H2** that longer explanation can inhibit a user's intention to change which could be a result of increase in the cognitive load [40]. The belief&goal-based XVA's dialogue is longer than the belief-based and goal-based dialogues with about 35% and 29% more words, respectively. Younger participants (20 years old and younger) showed significantly less intention change to join a study group after receiving belief&goal explanation compared to the other two types of explanations. This significant difference persists up to age 30 years. Further, highly stressed students can be motivated to join a study group by receiving belief-based explanation where stress level and intention change in the belief group were positively correlated. Hence, elements of the user profile can be determinants of which type of explanation a user should receive to motivate their intention change, which supports **H3**. Results from binary logistic regression explained how different factors from the user profile can predict the intentions to change. Context-aware XVAs should tailor its advice and explanation further according to the user profile and user's current context which could help motivating the user to change. For example, for students having an upcoming exam, the XVA can recommend the student to meet new people to cope with the study stress.

In the university setting, emotional stability has been found to be a significant predictor of stress vulnerability [10]. As a functional stress coping behaviour, students seek study support from university peers, particularly for first year students as they experience higher stress levels compared to others in the following years [17]. In this study, stress level was significantly correlated with emotional stability (Spearman's $\rho = -.400$ *at* $p < .001$) and, considering the user profile factors only, stress level was the significant predictor of the change in the intentions to join a study group. For every one-unit increase in the stress level, we expect a 1.379 increase in the log-odds of intention to join a study group, holding all other independent variables constant. This can been seen as a compelling reason to use an XVA, particularly belief-based XVA, that can be available any time to motivate students to join a study group and help them deal with their study stress.

Agreeableness was found as a significant predictor of the change in the intention to do physical activity using the user profile factors only. Prior meta-analysis reported that agreeableness and openness to experiences are weakly to not correlated with doing physical activity; however, this general conclusion, as reported in the meta-analysis, did not take age into consideration where age was found to mediate the association between the agreeableness and physical activity. The effect of agreeableness was clearer with people under 35 years old. About 76.9% of the participants in this study were under 35 years of age, thus, it is more likely to have agreeableness as a predictor for the intention to do physical activity.

Moreover, it is noteworthy to mention that the study has been conducted during the COVID-19 pandemic (February-June 2020). The advice to do the activities has been adapted to suit the government's guidelines in place during this special time. The XVA, for example, recommended walking alone or with a friend considering the social distancing and within the restricted local area only. No regular physical activity such as going to a gym or group activities

were recommended. Similarly, the XVA recommended different strategies to meet/interact with new people virtually over the internet. These COVID-19 specific modifications could explain the variation in the results of this study compared to previous findings. Recent studies on personality and complying with the COVID-19 restrictions reported that higher agreeableness is a main predictor of restriction compliance and open to experiences are highly correlated with behaviour awareness during this pandemic [32, 53]. As a result, agreeableness and openness to experiences are the only personality traits found as predictors for the intention change to do physical activity and to meet new people. Moreover, as an impact of COVID-19, the participants in the follow-up survey reported the difficulty to do regular exercise so their intentions decreased in the two groups (goal and beliefgoal groups). However, the intention to do physical activity was statistically significantly higher in the belief group at $p < .05$ and borderline in the goal group, $p = 0.05$ compared to the baseline (before interaction).

The two level binary regression models indicate that some factors from the user profile only can significantly explain around 23% of the variance in the intention change to join a study group and to do physical activity, and 34% of the variation in the intention change to meet new people as presented in Section 4.4. These variances could be explained further, up to 40%, 38%, and 41% for the three behaviours (Table 5), respectively, by including factors related to the user-agent relationship. Further, the differences in the models for the 3 behaviours revealed that the process of motivating a user to follow a particular recommendation or change a specific behaviour depends on different factors. Previously, for instance, Hagger et al. [21] found that personal attitudes impact the intention to change the dieting behaviour but not exercise while perceived control increases the intention to exercise more than dieting.

The user-agent relationship was measured using the trust and WA questionnaires. The results are reported in Table 3. No between-group differences were captured for all the scales and, hence, **H4** cannot be accepted. The trustworthiness scales (propensity, ability, benevolence, and integrity) are utilized in the study to capture the source of trust in the XVA. The propensity to trust strangers was not correlated with trust in the XVA which indicates that the participants perceived the XVA as trustworthy (state-like trust) and not because they tend to trust strangers (trait-like trust) [54]. Trust was moderately to strongly correlated with ability (Spearman's $\rho = .656$ *at* $p < .001$), benevolence (Spearman's $\rho = .461$ *at* $p < .001$), and integrity (Spearman's $\rho = .547$ *at* $p < .001$). Although the participants had the chance to mark the trustworthiness and trust questions as "not applicable", they did not take that option after interacting with belief-based or goal-based XVA, whereas 0%-4.2% did after interaction with the belief&goal-based XVA. The user-agent trust was associated with the changes in intentions to do physical activity and to meet new people. Participants who trusted the XVA were about 2.5 times more likely to change their intentions to do physical activity and to meet new people than those who did not build trust in the XVA. The XVA integrity was significantly associated with the intention change for doing physical activity with high odd ratio equal to 4.015. This positive and trust relationship built with the agent can explain the persistence of the intention to do the behaviours after 3 weeks of the interaction.

The WA results presented in Table 3 show that the participants marked the bond questions the most as not applicable, compared to other constructs, to describe their alliance with the XVAs. The participants in the belief&goal group scored the highest percentage of perceiving the bond questions as not applicable (24%) which could be due to the long discussion about their beliefs and goals, followed by participants who discussed beliefs only (19.7%) and then the goal only group (16.9%), who engaged in the briefest discussion. A similar pattern can be noticed with the participants' responses to how much they liked the XVAs. A prior study by Jeong et al. [29] found that when the agent interferes with the user's privacy, it could impose a feeling of discomfort. Thus, future work should incorporate measuring how the user perceives the agent asking to disclose his/her beliefs and goals to the agent and if this act influences the treatment outcome.

6 CONCLUSION

Prior studies concluded that the use of beliefs or goals in the explanation should be delivered according to the agent's action and the user's profile. However, little is known how those types of explanations are perceived by the user when the action should be performed by the user such as those actions recommended by virtual advisors. In this study, we investigated the influence of referring to the user's beliefs and goals to explain *why* an action should be taken by the user. Three patterns of explanations have been investigated including user's beliefs, goals, or both beliefs and goals. Although there was no difference between receiving belief-based explanation and goal-based explanation on the user's intention to change a behaviour, receiving a longer explanation that includes beliefs and goals tends to hinder the motivation to change the intention to do some behaviours. The preference to receive belief-based or goal-based explanation over belief&goal-based explanation was linked to the user profile with some behaviours (e.g. younger students were less motivated to join a study group after receiving belief&goal-based explanation).

We can conclude that the explanation pattern does not influence the user-agent relationship which is a great predictor of behaviour change. The participants rated the XVAs almost similarly as trustworthy and they built similar level of WA with the different XVAs. Besides the user profile, this relationship can predict the intention to do the recommended behaviours. However, different behaviours can be predicted by different factors. Defining these factors is a great advantage towards building a tailored XVA that can understand the user's characteristics and context and, then, tailor its advice and explanation accordingly to build a better relationship with the user and motivate behaviour change. While the experiment does not show clear overall difference in behaviour change between type of tailored explanation, except for some specific contexts, it is encouraging that, in general, making the explanation user-specific is adequate to motivate intention to change. Hence, to support this finding and as a future direction, we are interested in comparing the impact of providing explanations based on the agent's mental state to explanations based on the user's mental state. Further, it is of interest to measure how users perceive XVAs that discuss their beliefs and goals and see if that impacts on the XVAs efficacy for delivering behaviour change.

REFERENCES

[1] Amal Abdulrahman and Deborah Richards. 2019. Modelling working alliance using user-aware explainable embodied conversational agents for behavior change: framework and empirical evaluation. In *40th International Conference on Information Systems, ICIS 2019*. Association for Information Systems.

[2] Carole Adam and Benoit Gaudou. 2016. BDI agents in social simulations: a survey. *The Knowledge Engineering Review* 31, 3 (2016), 207–238.

[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[4] Tania Bailoni, Mauro Dragoni, Claudio Eccher, Marco Guerini, and Rosa Maimone. 2016. PerKApp: A context aware motivational system for healthier lifestyles. In *2016 IEEE International Smart Cities Conference (ISC2)*. IEEE, 1–4.

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[6] Jennifer K Bennett, Jairo N Fuertes, Merle Keitel, and Robert Phillips. 2011. The role of patient attachment and working alliance on patient adherence, satisfaction, and health-related quality of life in lupus treatment. *Patient education and counseling* 85, 1 (2011), 53–59.

[7] Timothy W. Bickmore, Laura M. Pfeifer, and Michael K. Paasche-Orlow. 2009. Using computer agents to explain medical documents to patients with low health literacy. *Patient education and counseling* 75, 3 (2009), 315–320.

[8] Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice* 16, 3 (1979), 252.

[9] Michael Bratman. 1987. *Intention, plans, and practical reason*. Vol. 10.

[10] Adomas Bunevicius, Arune Katkute, and Robertas Bunevicius. 2008. Symptoms of anxiety and depression in medical students and in humanities students: relationship with big-five personality dimensions and vulnerability to stress. *International Journal of Social Psychiatry* 54, 6 (2008), 494–501.

[11] Alice E Coyne, Michael J Constantino, Holly B Laws, Henny A Westra, and Martin M Antony. 2018. Patient–therapist convergence in alliance ratings as a predictor of outcome in psychotherapy for generalized anxiety disorder. *Psychotherapy Research* 28, 6 (2018), 969–984.

[12] Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

[13] Joost CF De Winter and Dimitra Dodou. 2010. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. In *Practical Assessment, Research & Evaluation*. Citeseer.

[14] Daniel Clement Dennett. 1978. Three kinds of intentional psychology. *Perspectives in the philosophy of language: A concise anthology* (1978), 163–186.

[15] Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

[16] Joao Dias, Samuel Mascarenhas, and Ana Paiva. 2014. *Fatima modular: Towards an agent architecture with a generic appraisal framework*. 44–56.

[17] Rebecca Erschens, Teresa Loda, Anne Herrmann-Werner, Katharina Eva Keifenheim, Felicitas Stuber, Christoph Nikendei, Stephan Zipfel, and Florian Junne. 2018. Behaviour-based functional and dysfunctional strategies of medical students to cope with burnout. *Medical education online* 23, 1 (2018), 1535738.

[18] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 227–236.

[19] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[20] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *International workshop on intelligent virtual agents*. Springer, 125–138.

[21] Martin S Hagger, Nikos LD Chatzisarantis, and Jemma Harris. 2006. From psychological need satisfaction to intentional behavior: Testing a motivational sequence in two behavioral contexts. *Personality and social psychology bulletin* 32, 2 (2006), 131–148.

[22] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, and John-Jules Meyer. 2010. Guidelines for developing explainable cognitive models. In *Proceedings of ICCM*. Citeseer, 85–90.

[23] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and evaluation of explainable BDI agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. IEEE, 125–132.

[24] Katherine Harman, Marsha MacRae, Michael Vallis, and Raewyn Bassett. 2014. Working with people to make changes: a behavioural change approach used in chronic low back pain rehabilitation. *Physiotherapy Canada* 66, 1 (2014), 82–90.

[25] Robert L Hatcher and J Arthur Gillaspy. 2006. Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy research* 16, 1 (2006), 12–25.

[26] Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[27] Denis J. Hilton, John McClure, and Robbie M. Sutton. 2010. Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology* 40, 3 (2010), 383–400.

[28] David Isern and Antonio Moreno. 2016. A systematic literature review of agents applied in healthcare. *Journal of medical systems* 40, 2 (2016), 43.

[29] Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2020. A Robotic Positive Psychology Coach to Improve College Students' Wellbeing. *arXiv preprint arXiv:2009.03829* (2020).

[30] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 676–682.

[31] Lean L Kramer, Silke Ter Stal, Bob C Mulder, Emely de Vet, and Lex van Velsen. 2020. Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review. *Journal of Medical Internet Research* 22, 2 (2020), e14058.

[32] Emily B Kroska, Anne I Roche, Jenna L Adamowicz, and Manny S Stegall. 2020. Psychological flexibility in the context of COVID-19 adversity: Associations with distress. *Journal of Contextual Behavioral Science* 18 (2020), 28–33.

[33] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

[34] V. Kumar, Ashutosh Dixit, Rajshekar Raj G. Javalgi, and Mayukh Dass. 2016. Research framework, strategies, and applications of intelligent agent technologies (IATs) in marketing. *Journal of the Academy of Marketing Science* 44, 1 (2016), 24–45.

[35] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems.. In *AAAI*, Vol. 17. 4762–4763.

[36] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 19.

[37] Bertram F. Malle. 1999. How people explain behavior: A new theoretical framework. *Personality and social psychology review* 3, 1 (1999), 23–48.

[38] Roger C Mayer and James H Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology* 84, 1 (1999), 123.

[39] Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[40] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[41] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are People Polite to Computers? Responses to Computer-Based Interviewing Systems 1. *Journal of Applied Social Psychology* 29, 5 (1999), 1093–1109.

[42] Hedieh Ranjbartabar and Deborah Richards. 2018. Should we use human-human factors for validating human-agent relationships? A look at rapport. In *workshop on Methodology and the Evaluation of Intelligent Virtual Agents (ME-IVA) at the Intelligent Virtual Agent conference (IVA2018)*. 1–4. https://iva2018methodologyworkshop.wordpress.com/proceedings/

[43] Anand S Rao, Michael P Georgeff, et al. 1995. BDI agents: From theory to practice.. In *ICMAS*, Vol. 95. 312–319.

[44] Deborah Richards and Patrina Caldwell. 2017. Improving Health Outcomes Sooner Rather Than Later via an Interactive Website and Virtual Specialist. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1699–1706.

[45] Lauren A Rutter, Robin P Weatherill, Sarah C Krill, Robert Orazem, and Casey T Taft. 2013. Posttraumatic stress disorder symptoms, depressive symptoms, exercise, and health in college students. *Psychological Trauma: Theory, Research, Practice, and Policy* 5, 1 (2013), 56.

[46] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Mller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[47] Francine Shapiro. 2017. *Eye movement desensitization and reprocessing (EMDR) therapy: Basic principles, protocols, and procedures*. Guilford Publications.

[48] Silke ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design Features of Embodied Conversational Agents in eHealth: a Literature Review. *International Journal of Human-Computer Studies* 138 (2020), 102409. https://doi.org/10.1016/j.ijhcs.2020.102409

[49] Philip E. Tetlock, Jennifer S. Lerner, and Richard Boettger. 1996. The dilution effect: Judgmental bias, conversational convention, or a bit of both? *European Journal of Social Psychology* 26, 6 (1996), 915–934.

[50] Karel van den Bosch, Maaike Harbers, Annerieke Heuvelink, and Willem van Doesburg. 2009. Intelligent agents for training on-board fire fighting. In *International Conference on Digital Human Modeling*. Springer, 463–472.

[51] Douglas Walton. 2004. A new dialectical theory of explanation. *Philosophical Explorations* 7, 1 (2004), 71–89.
[52] S Christian Wheeler, Richard E Petty, and George Y Bizer. 2005. Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. *Journal of Consumer Research* 31, 4 (2005), 787–797.
[53] Marcin Zajenkowski, Peter K Jonason, Maria Leniarska, and Zuzanna Kozakiewicz. 2020. Who complies with the restrictions to reduce the spread of COVID-19?: personality and perceptions of the COVID-19 situation. *Personality and individual differences* 166 (2020), 110199.
[54] Sigal Zilcha-Mano. 2017. Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *American Psychologist* 72, 4 (2017), 311.