Multimodal Technologies and Interaction

Article

Adapting a Virtual Advisor's Verbal Conversation Based on Predicted User Preferences: A Study of Neutral, Empathic and Tailored Dialogue

Hedieh Ranjbartabar [1,*], Deborah Richards [1], Ayse Aysin Bilgin [2], Cat Kutay [3] and Samuel Mascarenhas [4]

[1] Computing Department, Macquarie University, Balaclava Rd, Macquarie Park NSW 2109, Australia; deborah.richards@mq.edu.au
[2] Mathematics and Statistics Department, Macquarie University, 12 Wally's Walk, Macquarie Park NSW 2109, Australia; ayse.bilgin@mq.edu.au
[3] College of Engineering, IT and Environment, Charles Darwin University, Ellengowan Drive, Darwin NT 0815, Australia; cat.kutay@uts.edu.au
[4] INESC-ID & Institute Superior Técnico, Universisade de Lisboa, R. Alves Redol 9, 1000-029 Lisboa, Portugal; samuel.mascarenhas@gaips.inesc-id.pt
* Correspondence: hedieh.ranjbartabar@hdr.mq.edu.au

Received: 15 July 2020; Accepted: 14 August 2020; Published: 17 August 2020

Abstract: Virtual agents that improve the lives of humans need to be more than user-aware and adaptive to the user's current state and behavior. Additionally, they need to apply expertise gained from experience that drives their adaptive behavior based on deep understanding of the user's features (such as gender, culture, personality, and psychological state). Our work has involved extension of FAtiMA (Fearnot AffecTive Mind Architecture) with the addition of an Adaptive Engine to the FAtiMA cognitive agent architecture. We use machine learning to acquire the agent's expertise by capturing a collection of user profiles into a user model and development of agent expertise based on the user model. In this paper, we describe a study to evaluate the Adaptive Engine, which compares the benefit (i.e., reduced stress, increased rapport) of tailoring dialogue to the specific user (Adaptive group) with dialogues that are either empathic (Empathic group) or neutral (Neutral group). Results showed a significant reduction in stress in the empathic and neutral groups, but not the adaptive group. Analyses of rule accuracy, participants' dialogue preferences, and individual differences reveal that the three groups had different needs for empathic dialogue and highlight the importance and challenges of getting the tailoring right.

Keywords: user model; human computer interaction; virtual humans; intelligent virtual agents; rapport; study stress; virtual advisor; agent's expertise

1. Introduction

There are many potential applications of intelligent virtual agents, virtual humans with the ability to interact intelligently with humans. They could perform social roles, such as a virtual advisor. Intelligent interaction requires the agent to adapt to the user based on real-time understanding of the context including the user. The verbal and non-verbal behaviors of virtual humans have reached a level of sophistication and believability that allow them to use situation awareness and conversational functions such as listening, turn taking and feedback, reasoning over the changes in their environment, and responding accordingly. There are a number of frameworks and agent architectures such as MultiSense/SimSensei [1], Greta [2], ODVIC [3] and FAtiMA [4] and Agent United Open Platform.

Humans will adapt their responses to other humans in different ways based on more than the inputs they receive during an interaction. The reasons can be varied, including their past experiences, current feelings and past interaction [5]. Agents who convey their feeling and react emotionally to events, are more believable [6]. Many agent architectures, such as FAtiMA (Fearnot Affec Tive Mind Architecture) [4], allow the agent to have their own emotions and autobiographical memories gained through past interactions or events that allow them to respond in a humanlike and socially believable way. Our focus, however, is not on agent believability but on responding in ways that are appropriate to the needs of each user. To meet the needs of users, the agent must learn from the events in the world and adapt to them accordingly [7]. Learning requires the agent to develop models about that user and allows them to grow in their expertise about how to handle the needs of the users they service. For that reason, our work involves the agent being friendly and responding empathically (expressed via dialogue). Our work focuses more on the human's features and less on the agent's features.

Many human features need to be modeled and reasoned over to support social, educational or health scenarios involving human–agent interactions. For a human in helping roles such as teacher, coach, mentor, advisor or therapist, they need to draw on deeper knowledge of the individual user, beyond their own memories (such as in medical records, learning management systems) and their own expertise acquired through years, even decades, of dealing with other people in their care.

For the agent to be able to adapt to different users, not just to different user feedback, we need to include understanding of the user (i.e., model of the individual) and knowledge of how to respond to different types of users (i.e., past experience/knowledge of similar users) as part of the agent architecture. To achieve this, we have extended FAtiMA with an Adaptive Engine comprised of a collection or repository of User Models, one for each user, and an Agent Expertise module that represents what the agent has learnt by interacting with a range of users in the past. The User Model can include complex elements such as the user's verbal and non-verbal responses, personality, preferences, emotional state.

In this paper, we present our proposed Adaptive Engine, its implementation and a study to evaluate the adaptive behavior. In this article, we have used the terms adaptive and tailoring interchangeably. We view tailoring as a subset of adaptation, where the agent may adapt its behaviour based on many factors including factors internal to the agent (i.e., its knowledge, personality, emotion, reasoning, etc.) or its environment. We see the user as more than just part of the environment, where each individual user is the focus of the agent and the agent adapts its dialogue based on its knowledge of the user (the individual user model) and its expertise on how to deal with users of this type. In this respect, our agent uses tailoring (expertise of how to behave with certain types of users) and personalization (to respond to specific user).

In this study we use three variations of a scenario (empathic, neutral, tailored/adaptive). All of these scenarios are aimed at Reducing Study Stress. In all three, the dialogue differs according to the 10 empathic dialogue cues identified by Bickmore [8], where the neutral scenario uses no empathic cues; the empathic scenario uses all 10 cues and in the adaptive scenario the dialogue is adapted to include or omit one or more of the empathic cues according to the agent's knowledge of the individual user and the agent's expertise gained from dealing with users over time. The evaluation seeks to answer these research questions:

- RQ1 Do users feel less stressed after interacting with a virtual human when it uses tailored, empathic or neutral dialogue?
- RQ2 Do users establish more sense of rapport with a virtual human when it uses tailored, empathic or neutral dialogue?
- RQ3 Does increased rapport lead to less study Stress?

In the following section, we first provide some background review of empathic and adaptive agents. We then introduce in Section 3 the extensions to FAtiMA to support adaptation to the user followed by our methodology in Section 4 to evaluate the extensions and answer the above research questions. Results are presented in Section 5, followed by discussion, and conclusion and future work, in Sections 6 and 7, respectively.

2. Background & Related Literature

First, we consider the role of empathy and approaches to empathic agents, followed by a review of ways in which agents may adapt to the user.

2.1. Expressing Empathy Using Virtual Humans

Empathic agents are designed to detect the user's emotion and respond in an emotionally intelligent way. With artificial agents being increasingly called upon to improve health care support and advice systems, especially in the domain of mental health, many studies on empathic agents focus on health care domain such as ODVIC [3] and MultiSense/Simsensei [1]. Empathy can involve cognitive or affective attributes. The former involves reasoning to understand and demonstrate this understanding of the user. Affective attributes involve physiological responses and rapid expression change responding to someone else's display of emotions, such as mimicry. This work focuses on the cognitive attributes.

Previous work suggests that empathic ability can improve user satisfaction [9] and user affinity for the agent [10] both of which are helpful for patient engagement in treatment. In fact, any application in artificial social interaction can be enhanced by improved interaction with the user [11]. Research on empathic virtual agents has shown that these agents are perceived as more likeable, trustworthy, and caring [10] and they have successfully developed long-term relationships with human partners [12]. However, use of empathy does not always improve intended outcome or improve rapport with the agent in comparison to use of neutral dialogue [13].

A study by Gratch, et al. [14] shows that people tend to disclose more personal information when they interact with a virtual human rather than a real human. Ochs, et al. [15] compared participants' responses to empathic expression of emotion with no expression of emotion. They found that if the agent is empathetic, they are perceived as expressive and cheerful regardless of whether the emotion being expressed is positive or negative, as long as the empathic cues are congruent. In contrast, if the empathic cues are not congruent, the agent is perceived as more irritating, cold and strange when emotion is expressed compared to no expression of emotion. [15]. In health applications, expressing empathy is important for information extraction, which is key to supporting the patient [16].

Empathy for the user can be expressed by verbal or nonverbal cues (e.g., head nods and facial expressions). In the context of health, verbal empathy refers to the ability of the doctor to use verbal cues to help their patient disclose their symptoms ([17] p. 223). When used for giving advice, agents expressing empathy through social dialogue, reciprocal self-disclosure and other techniques have been shown to establish more rapport [8]. Our work draws on these techniques to express empathy to the user. To generate empathy in agents, there are two methodologies: analytical and empirical [18]. In the analytical method, a model of empathy is developed by analysing the findings of the empathy literature. The empirical approach relies on data collected in studies that examine peoples' social and emotional interactions. Empirical data collected prior to engaging with the agent can be used to train computational models using algorithms that generalise patterns of empathic behaviour relating to the user's preconditions [18]. In this paper, we discuss a model that used data collected on relevant aspects of the user profile prior to engaging with the agent.

2.2. Adaptation

Literature shows that in designing agents' features, the most attention has been focused on speech and facial expression of the agents in the health domain [19]. In an agent with the capability of communicating through speech tailored to different users, designing a user model is vital. Early adaptive agents were distinguished by the number of modalities they supported. For instance, in the mimicking agents framework [20], agents copy or mirror the non-verbal behaviors of the user such as head nods or facial expression, without needing to understand the user's verbal or non-verbal behaviours. The process of adapting the agent's affective state to match the changing affective state of the target person to display empathic behaviour, based on the observed cues from the target, can be implemented through ongoing feedback or analysis of target [21]. However, a listening agent [22] responds according to the verbal and non-verbal inputs from the user with non-verbal behaviors to convey understanding and empathy. With advances in agent architectures, adaptation also focuses on agents reasoning over the environment. In ORIENT [7], an educational role play game, the aim is to create non-player characters more believable in multi-agent worlds who respond to the player both cognitively and emotionally. As in this work, ORIENT is based on FAtiMA [4] agent architecture taking into account the PSI [23] motivational and learning system to increase the agent's adaptive ability. In this architecture, drives lead the goals. The five basic drives from PSI are energy, integrity, affiliation, certainty, and competence [7].

SAIBA framework [24] allows multimodality generation (i.e., verbal and nonverbal behaviours) by conveying different ways of interaction through different inputs and outputs. In particular in agents designed to deal with psychological distress, such as SimSensei Kiosk [25], the agent utilizes Multi-Sense framework, which is a multimodal perception system and is sensitive to the user's behaviours in real time to identify indicators of psychological distress. The core functionalities are dialogue processing, multimodal perception and nonverbal behaviour generation [25]. Simsensei Kiosk uses dialogue processing (i.e., speech recognition, language understanding and dialogue management) to process users' verbal behaviours. The system seeks to make the user feel comfortable by hearing from an agent that listens to the user continuously and responds empathically.

In scenarios where agents need to help the user in multi-objective decision making, understanding the user's preferences has a key role in how the agent responds in the most efficient way. For instance, to choose the best holiday trip or the best university for the user, the agent needs to know the user's preferences and needs. Zintgraf, et al. [26] have proposed elicitation strategies using multi-objective decision scenarios where each decision has several properties that have an influence on what user's preferences are stored.

To find an appropriate response during a user's varying state, active learning can be used. "Active learning is where we alternate between updating the model of user's utility and querying for the user for feedback by using reactive feedback queries." ([26] p. 1478). To avoid asking too many questions from users and still create an optimal user preference, two methods were used: absolute feedback and relative feedback. Absolute feedback is scoring an item, while relative feedback is comparing items. Zintgraf, et al. [26] used more focus on relative feedback since it has been said to be more consistent given human variation [27].

Collecting data from the user during the game or interaction can be distracting or imprecise. While online pre and post surveys are used for data collection from users for modelling and evaluation, less intrusive approaches are being investigated. Social media provides opportunity for future pre-analysis of user models. Verhagen, et al. [28] developed a Facebook application to analyse the user's personality and Chen, et al. [29] used log data from a health app to understand user engagement. For situations where the information being sought is more short-term reactions such as stress and there is no access to the user's background, our user model integrates detailed information from a pre-survey to generate a relational user model to select empathetic/neutral responses in a therapy agent system.

3. Extensions to FAtiMA

FAtiMA [4] is a modular agent architecture for creating autonomous, engaging and believable characters. Being modular helps researchers to utilize a subset of the architecture components or extend the architecture to create their own simpler or more complicated design based on their own scenarios and goals, such as adding the agent's motivation [7] or culture [30]. The inspiration for FAtiMA came from earlier work involving interactive pedagogical drama, such as Carmen's Bright Ideas [31], and has been used in multiple projects for role-playing and social simulations including the FearNot! System for teaching school children how to cope with bullying and cultural sensitivity towards refugees [32–34]. In these applications, users observed virtual humans acting out scenarios. Thus, FAtiMA was initially designed for multi-agent interactions, where the human user could select options regarding appropriate responses to the scenario and potentially take on a role of one of the characters. There was no user model in Fatima but rather agent character models.

3.1. Conceptual Design of Proposed Extended Architecture

Our work has been to extend FAtiMA to support capture of a collection of user models and development of agent expertise (i.e., knowledge/rule base) that is either seeded using machine learning from a suitable data source, or which develops over time as the agent learns how best to meet the needs and/or respond to types of users (e.g., groups of users with similar characteristics such as preferences or biographical features).

In FAtiMA, the verbal and non-verbal behaviours of the agent are influenced by the emotional state and personality of the agent, which are influenced by the perceived events. The character has its own personality, emotion and belief but is not capable of reasoning on the user's features and adapting or reacting accordingly. Thus, we propose an extended architecture for "Study Stress" scenario by incorporating a User Model repository and expertise module to be used by the adaptive agent.

Figure 1 shows our proposed extended FAtiMA architecture for the adaptive agent. The modules from FAtiMA are illustrated in grey with a double-line border and the extended modules are in blue with a single-line border.

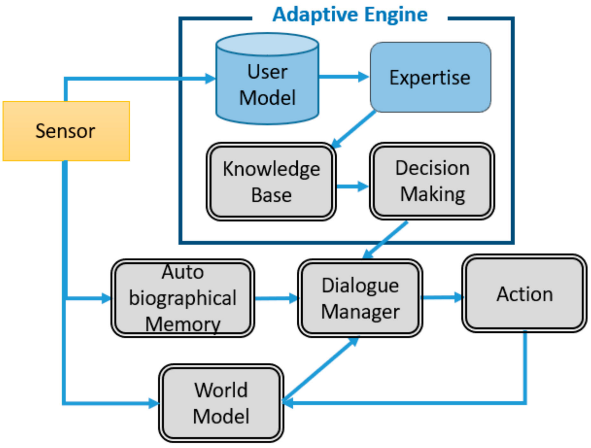


Figure 1. Extended FAtiMA with Adaptive Engine.

In FAtiMA, the Knowledge Base module stores the character's beliefs about the world, which could include properties of an object or relationships. In the Autobiographical Memory, past events are stored. In Decision Making, a set of action rules is authored. Each action rule must specify a list of activation conditions that are verified against the beliefs of the agent through a logic unification algorithm. If the unifier finds a valid set of substitutions that verify the action's conditions, the agent forms an intention to execute the action.

In the extended architecture, once the agent perceives an event in the world (e.g., visiting a new user), the user profile is loaded into the User Model. The user profile could be collected via a number of alternative methods including incremental acquisition through a survey or conversation with the character or from an existing data source such as a medical or academic record. For the study presented in the next section, we ask each user to complete an online survey using Qualtrics, the online survey tool, before the interaction, which the agent is then able to access via their input sensor. In parallel, Autobiographical Memory and World Model will store this event.

The adaptive rules are in a separate module called Expertise derived by applying machine learning and data mining techniques on data collected in previous studies involving our virtual advisor. The User Model will trigger appropriate rules for the user in the Expertise module. The Expertise module has the potential to update the rules by learning from feedback on the fly. Then, the Knowledge Base is updated for each user based on the Expertise (i.e., if the current user likes a particular sentence to be uttered empathically or neutrally). Next, the decision-making module selects the appropriate dialogue for communication from the dialogues defined in the Dialogue Editor, which holds all the dialogues that our adaptive character selects from them according to the triggered action rules in the Knowledge Base. Thus, the agent is able to look at the user model, know about the user and react adaptively. More details are described in Methodology in Section 4.

3.2. *Design of Proposed Extended Architecture*

To develop the Adaptive Engine, we considered how to answer three questions: (1) what to adapt to? (2) when to adapt? and (3) how to adapt? We wanted Sarah, our Virtual Human, to know about the user's features, create the individual user model and then adapt accordingly.

There are many aspects of a virtual human that can be varied including, but not limited to, their appearance and other physical features; role or purpose, which will affect the domain knowledge they possess; verbal and non-verbal language; reasoning processes; and affective behaviours. To narrow what would be adapted, consistent with our previous investigations, we focused on adapting the agent's dialogue to contain empathic or neutral dialogue following the 10 empathic cues outline by Bickmore [8,35]. To decide when to adapt, we needed to capture the profiles of individuals and their preferences concerning those empathic cues and learn rules that could be applied to predict the preferences of others according to their profile using the same features. This would provide us with a set of user models and also a set of rules to reason over those models. In answering how to adapt, we needed to modify the agent so that it would be able to create or update a model of each user it interacts with, use its knowledge about what types of cues people prefer and then modify the FAtiMA architecture to adapt its dialogue in real time. Finally, we needed to conduct a study that investigated whether the tailored dialogue delivers better results than dialogue that is not tailored and evaluate the Adaptive Engine.

4. Methodology

Section 4.1 presents our experimental design to perform this design and evaluation. Section 4.2 describes the materials and scenario used for the experiment. Section 4.3 presents the experimental procedure and data collected. Section 4.4 describes how the rules for adaptation were created.

4.1. *Adaptive Dialogue Rules*

Creating the character's adaptive dialogue, which is tailored to different users, involved the use of data captured in three previous studies conducted from 2017 to 2018 that also used the Reducing Study Stress scenario with Sarah. The study design is shown in Figure 2.

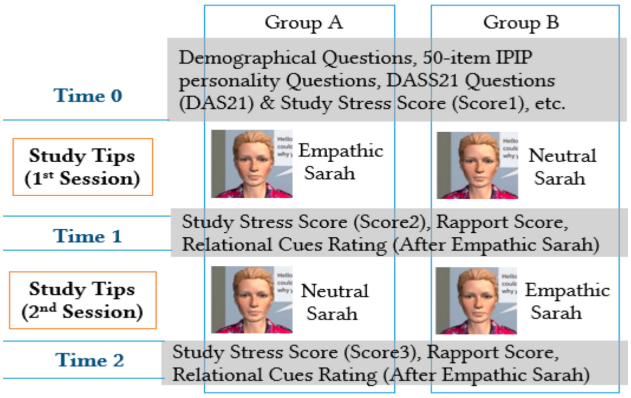

Group A
Group B
Time 0
Demographical Questions, 50-item IPIP personality Questions, DASS21 Questions (DAS21) & Study Stress Score (Score1), etc.
Study Tips (1st Session)
Empathic Sarah
Neutral Sarah
Time 1
Study Stress Score (Score2), Rapport Score, Relational Cues Rating (After Empathic Sarah)
Study Tips (2nd Session)
Neutral Sarah
Empathic Sarah
Time 2
Study Stress Score (Score3), Rapport Score, Relational Cues Rating (After Empathic Sarah)

Figure 2. Design of previous studies that captured data used for Machine Learning.

As illustrated in Figure 2, before interacting with the character, we collected demographic data including gender, age, cultural background, degree being studied, computer game activity, and attitude/aim towards study. Then, followed by IPIP (International Personality Item Pool), DASS21 (Depression Anxiety Stress Scores) questionnaires we also collected for other parameters such as their study stress score. The experiment consists of two groups to which participants were randomly assigned. Group A interacted with Empathic Sarah first, followed by Neutral Sarah. Group B did the reverse.

We collected 376 participants' data. The empathic dialogues were created using 10 relational cues (RC) identified from the literature [8]. For each relational cue Table 1, there are multiple sentences uttered by the character in different situations. To evaluate the empathic dialogues, at the end of each empathic interaction (i.e., after the 1st session in group A, and 2nd session in group B), participants were asked if the sentences using the relational cues seemed helpful, empathic and/or stupid. We combined the RC values for positive responses, Helpful and/or Empathic (HE), by using 'Exclusive Or' function to finally have two fields for each cue (i.e., RC_HE and RC_Stupid). More details of the study design can be found in our previous publications [36].

Table 1. Empathic Cues used in the Dialogues.

Relational Cues	Example	RCs
Social Dialogue	"How are you going?"	RC1
Meta-Relational Dialogue	"As a last thing together, … "	RC2
Empathic Feedback	"I am here for you."	RC3
Humor	"If I actually have a mouth, I think I'd eat healthy food."	RC4
Continuity behaviors	"I am waiting here for you."	RC5
Self-Disclosure	"I got the tip from my friend."	RC6
Reference to mutual/sharing knowledge	"We think alike."	RC7
Solidarity and rapport- mirroring	"So, you are a day person like me."	RC8
Politeness	"Please make yourself comfortable."	RC9
Inclusive pronoun	"Sometimes it's nice to have our own time."	RC10

To figure out when and how to adapt to the user, we generated the adaptive rules for our user model by using C5.0 classification modelling methods in IBM SPSS modeler version 18.0. Classification models use the value of input field/s to predict the value of output or target field/s. The modelling techniques in C5.0 includes machine learning, statistical methods, and rule induction. As well as generating a decision tree, C 5.0 also ranks the predictors based on their importance for the creation of the decision tree.

We created the C5.0 model for each 10 RC_HE, using three methods. In method 1, for each 10 RC, group A data was the training and group B data was the testing partitions to create the 10 models. Then, in method 2 we split the full dataset, including group A and group B, into 70% randomly assigned to the training and 30% to the testing data. We replicated method 2 to form method 3 by assigning 60% of the full dataset to the training and 40% to testing data. When the best model was identified for each RC, we used that model for adaptation of the speech. To measure the success of the three predictive models, we calculated the Lift value and accuracy for each outcome category and chose the best model for each RC.

For instance, for RC1, the model with 70% randomly assigned data for training dataset and 30% for testing dataset had the highest accuracy. The severity of pruning the C5.0 model was 85% for RC1 and no misclassification cost was required. For RC3, we used the model where group A was the training and group B was the testing partitions. Without misclassification cost, we were only able to identify one important predictor (stress), however, after including misclassification cost to 2, the number of predictors increased to four (i.e., stress, score1, extraversion, examsoon). Misclassification cost was used in RC3. For RC2, we used the model with 60% and 40% partitions for training and testing respectively and no misclassification cost was utilised.

In all models, each RC_HE is the target (outcome) and the 18 parameters from the user profile are the input (predictors). The predictors of the models are user profile information including age, gender, cultural group, degree being studied, computer game activity, grade being aimed to achieve, grade being expected to achieve, study attitude, any exam coming soon, DASS21 and IPIP results, and study stress score.

Figure 3 shows the training and testing C5.0 models for RC7_HE. Method 2 is used here when the full dataset is partitioned randomly into 70% for training and 30% for testing models. The accuracy of the training model is 78.5%. The lift values for the two outcomes are greater than one, so we can conclude that the accuracy of the model is better than randomly assigning the sentences.

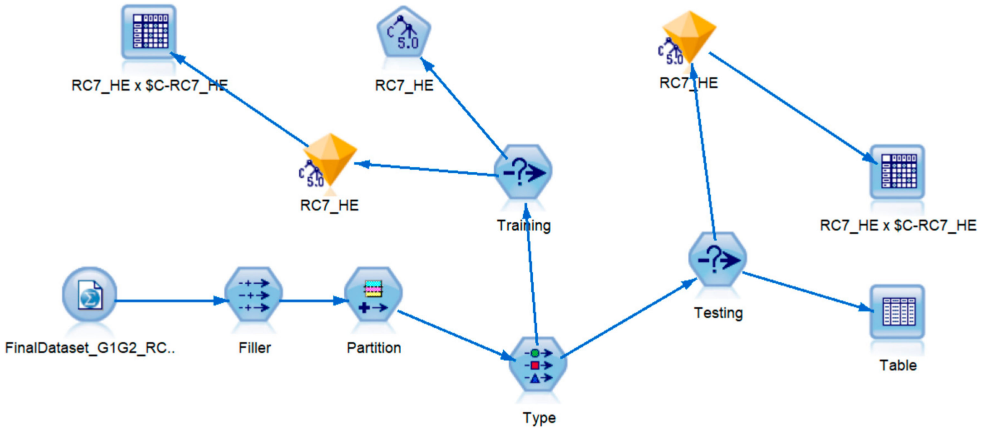
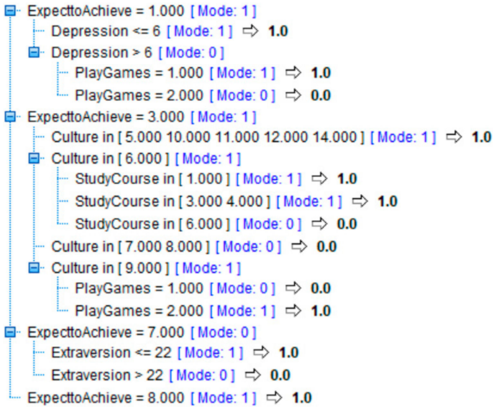


Figure 3. Training and Testing Models for RC7_HE.

Figure 4 shows the rules for RC7_HE. To interpret the decision tree, after expanding all the branches, we can see that there are 13 leaves/rules and 7 predictors. If the target is 1 (rules predicted that RC7 is helpful/empathic to the user), then the user will receive all the sentences using RC7. If the outcome of the decision tree is 0 (rules did not predict that the cue was helpful/emphatic to the user), then the user will not receive the empathic sentences, instead they will receive the neutral versions of the sentences.



```
ExpecttoAchieve = 1.000 [ Mode: 1 ]
    Depression <= 6 [ Mode: 1 ] ⇨ 1.0
    Depression > 6 [ Mode: 0 ]
        PlayGames = 1.000 [ Mode: 1 ] ⇨ 1.0
        PlayGames = 2.000 [ Mode: 0 ] ⇨ 0.0
ExpecttoAchieve = 3.000 [ Mode: 1 ]
    Culture in [ 5.000 10.000 11.000 12.000 14.000 ] [ Mode: 1 ] ⇨ 1.0
    Culture in [ 6.000 ] [ Mode: 1 ]
        StudyCourse in [ 1.000 ] [ Mode: 1 ] ⇨ 1.0
        StudyCourse in [ 3.000 4.000 ] [ Mode: 1 ] ⇨ 1.0
        StudyCourse in [ 6.000 ] [ Mode: 0 ] ⇨ 0.0
    Culture in [ 7.000 8.000 ] [ Mode: 0 ] ⇨ 0.0
    Culture in [ 9.000 ] [ Mode: 1 ]
        PlayGames = 1.000 [ Mode: 0 ] ⇨ 0.0
        PlayGames = 2.000 [ Mode: 1 ] ⇨ 1.0
ExpecttoAchieve = 7.000 [ Mode: 0 ]
    Extraversion <= 22 [ Mode: 1 ] ⇨ 1.0
    Extraversion > 22 [ Mode: 0 ] ⇨ 0.0
ExpecttoAchieve = 8.000 [ Mode: 1 ] ⇨ 1.0
```

Figure 4. Decision Tree for RC7_HE.

Based on the rules we generated, we provide three examples below based on three different types of adaptive dialogues for three different user groups. In the first one, RC2, RC3 and RC6 are triggered. In the second one, only RC10 is triggered and in the last one there is no RC triggered.

Adaptive Agent: Let's talk about socializing, which is good for our mental health (RC2 & RC10). That's why I'm here (RC2 & RC3). It helps reduce the symptoms of depression and anxiety. I'm supported by my friends and family (RC6). Do you feel supported in your life?

Adaptive Agent: Let's talk about socializing, which is good for our mental health (RC2 & RC10) It helps reduce the symptoms of depression and anxiety. Do you feel supported in your life?

Adaptive Agent: Socializing is good for your mental health. It helps reduce the symptoms of depression and anxiety. Do you feel supported in your life?

4.2. Experimental Design for Evaluation

The aim of this study is to compare the results of user interactions with three different types of virtual advisors: neutral, empathic and adaptive. The adaptive virtual advisor used a scenario designed by the authors called "Reducing Study Stress" where Sarah talks to the user and provides study tips using neutral, empathic or tailored dialogue. The tailored dialogue was generated by adaptive dialogue rules described in Section 4.1.

As shown in Figure 5, participants in Empathic, Neutral and Adaptive groups interacted with empathic, neutral and adaptive Sarah, respectively. Data for Empathic and Neutral groups were collected from our previous study as depicted in Figure 2. Since in the previous study each participant in group A and group B interacted with both Empathic and Neutral characters in the different orders in two consequent sessions, we only extracted the first session of each group, to form Empathic and Neutral groups for the current study as shown in Figure 5. However, Adaptive group includes both adaptive sessions. For Adaptive group, we use the stress score and rapport obtained after the second session.

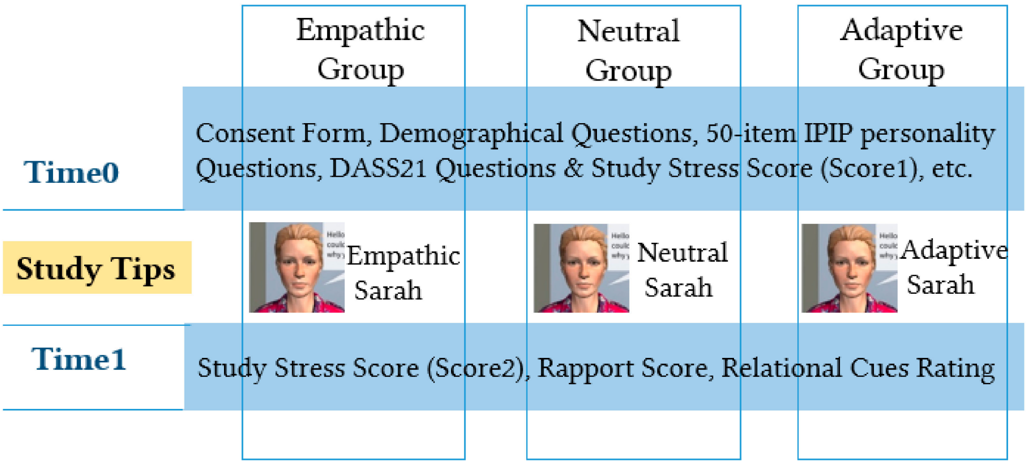

Empathic Group
Neutral Group
Adaptive Group
Time0
Consent Form, Demographical Questions, 50-item IPIP personality Questions, DASS21 Questions & Study Stress Score (Score1), etc.
Study Tips
Empathic Sarah
Neutral Sarah
Adaptive Sarah
Time1
Study Stress Score (Score2), Rapport Score, Relational Cues Rating

Figure 5. Study Design.

From these previous studies, we found that students significantly reduced their study stress levels through discussion with the virtual adviser but they did not necessarily establish more rapport with the empathic character and/or find them more helpful in terms of stress reduction, as compared to the neutral advisor. This led us to explore whether an adaptive agent would make a difference. Thus, we collected further data after implementing the Adaptive Engine and creating an adaptive Sarah, as described in the next section. These studies were approved by the Macquarie University's Human Research Ethics Committee (approval reference number 5201700595).

4.3. Materials and Method

Sarah provides study tips to reduce study stress. The contents of the dialogues were the same as the previous study and derived from the campus wellbeing of our university and included work, study and life balance, exercise and healthy eating, overcome exam stress and socialising tips. For empathic Sarah's dialogue, we modified neutral Sarah's dialogue to include Bickmore's 10 empathic (relational) cues to create Empathic Sarah's dialogue. The cues with related examples are shown in Table 1.

Adaptive Sarah's default dialogue is neutral. Based on her knowledge about the user on the fly, her dialogue will become adaptive. FatiMA uses a Hybrid approach to create the dialogue. In the hybrid state machine, each state is identified by a meaning, style, utterance and next state. The values in meaning and style help us to link the dialogue with other parts of the architecture. For each neutral sentence in the dialogue, there is an empathic version. The empathic and neutral sentences are distinguished by the value in style. For instance, if the value in style is RC2, it means that the sentence includes the second empathic cue. The process of identifying which cues were suitable for which user is presented in the next section.

The Unity3D game engine and FAtiMA Toolkit were used to implement Sarah (Figure 6), the virtual adaptive advisor. Sarah uses lip-synching. She did not have any non-verbal behaviours except a smile. FAtiMA toolkit allowed us to import a dialogue file and use TextToSpeech feature to automatically generate Sarah's synthesized voice, which is also selected based on users' selection of answer options.

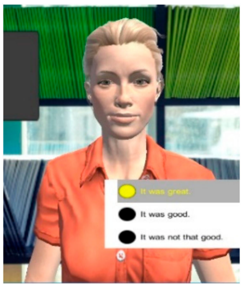


Figure 6. Sarah, the Virtual Advisor.

To implement the Adaptive Engine in FAtiMA, we specified Sarah as the role-playing character and used the simulator to test the dialogues. The Adaptive Engine was able to access the user model of each individual user from a repository of stored user profiles. Individual user profiles were populated by asking the user to complete a Qualtrics survey before the interaction. The user profile consists of 18 factors that cover the user's demographics, personality, emotional state, and study goals and attitude. The adaptive rules are implemented in the knowledge base to provide the best adaptive responses to the user based on these 18 parameters.

4.4. Procedure and Data Collection

To create the individual user model for each participant, we collected demographic data including gender, age, cultural background, degree being studied, computer game activity, attitude and goal towards study and study stress score. The reasoning of Sarah is also influenced by the personality and psychological emotional state of the user. Thus, to model the personality of the user, we used International Personality Item Pool (IPIP), which is called the Big Five personality factor [37], and to measure psychological emotional state of the user we used DASS21 questionnaire [38]. The IPIP questionnaire consists of 50 questions, each with 10 items measuring one of the five personality traits. The big five factors include Openness to experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N), hence also known as the OCEAN model. Each trait is a dimension with two poles: Openness vs Closed; Conscientiousness vs Disorganized; Extraversion vs Introversion; Agreeableness vs Disagreeableness and Neuroticism vs Emotional Stability. A higher score on the 5-point likert scale indicates greater openness, conscientiousness, extraversion, agreeableness, and emotional stability. Thus, for increased comprehension and easier comparison between the five personality dimensions, we label and refer to the Neuroticism dimension as Emotional Stability. The DASS21 questionnaire measures depression, anxiety and stress level of the user. All these data form the 18 factors in the user model.

To measure the sense of rapport built between the user and the character, after each interaction participants were asked to answer 20 rapport questions. The questions used a 5-point Likert-type scale from "strongly disagree" (1) to "strongly agree" (5), derived from five studies [25,39–42]. For each response to a rapport question, the negatively worded questions were reverse coded. Participants could also choose the option "not applicable".

To determine if the adaptive behaviour and study tips were useful, Sarah asked the user twice (at the beginning and the end of the interaction) to 'think about your emotional feeling towards your study on a scale of 0 to 10. Zero means "extremely good and relaxed" and 10 means "extremely bad and stressed"'. To evaluate the dialogue provided by adaptive Sarah, we presented 20 empathic sentences (i.e., two for each relational cue) and asked the participants to specify if they found it empathic, helpful and/or stupid. They could select more than one option (i.e., relational cue rating questionnaire, depicted in Figure 5). Participants were able to download Sarah through the Qualtrics survey and any interaction with Sarah through the keyboard and mouse was captured into a separate MySQL database.

The normality assumptions of the variables for the statistical analysis were checked by using Kolmogorov–Smirnov test of normality. For variables that were significantly different than normal distribution, we used non-parametric tests such as Kruskal–Wallis and Wilcoxon Signed-Ranks tests. To test the relationship between the categorical variables, chi-square test was conducted. We chose 0.05 for significance level.

5. Result

In this section, after introducing the participants' demographics, we compare the results of the last experiment (i.e., Adaptive group) with the other two experimental groups, Empathic and Neutral groups. In the final subsection we include analysis of the results from applying the rules generated from machine learning. This analysis will allow us to determine whether Sarah was adapting correctly or not.

The result of normality tests for the full dataset and for each group showed that for some variables, the departure from normality was large ($p < 0.005$), and therefore, non-parametric statistical analysis were used. Due to underlying algorithms of Kolmogorov and Shapiro, different results were produced for Score2, Openness, Rapport Average, and Rapport Sum. Graphical checks by using histograms of the variables showed that the distributions of them did not seem to be significantly different than normal distribution, therefore, where possible, we used parametric statistical tests.

5.1. Participants

Our study involved a collection of 59 participants' data for Adaptive group (mean age = 19.88, SD = 2.485). To compare the results of interaction by adaptive Sarah with neutral and empathic dialogue, we chose to use data from our most recent previous study in 2018. There are 52 participants in empathic (mean age = 20.54, SD = 5.816) and 43 participants in neutral (mean age = 20.23, SD = 3.146) groups. The final number of 154 students aged between 18 and 57 (mean age = 20.20, SD = 4.051) participated in all parts of the study. Only three participants were over 30 years old. Table 2 shows the gender distribution across each group for the 154 participants. Overall, there were more female participants than male participants. For Neutral and Adaptive groups, the number of male and female participants are almost equal. The Empathic group had 10% more females than males.

Table 2. Gender Distribution Across Experimental Groups.

Groups	Female		Male		Other		Total	
	%	N	%	N	%	N	%	N
Empathic	38	32	28	19	100	1	34	52
Neutral	26	22	30	21	0	0	28	43
Adaptive	36	30	42	29	0	0	38	59
Total	100	84	100	69	100	1	100	154

Table 3 shows the largest cultural group is Oceania (28%) and the next largest groups are South-East Asian and Northern-Western European (19% and 16% respectively).

Table 3. Cultural Groups Distribution by Group.

Cultural Groups	Empathic		Neutral		Adaptive		Total	
	N	%	N	%	N	%	N	%
Oceania	12	23%	12	28%	19	32%	43	28%
Northern-Western European	7	13%	8	19%	9	15%	24	16%
Southern-Eastern European	2	4%	0	0%	3	5%	5	3%
North African and Middle Eastern	4	8%	2	5%	3	5%	9	6%
South-East Asian	10	19%	7	16%	12	20%	29	19%
North-East Asian	1	2%	3	7%	2	3%	6	4%
Southern and Central Asian	2	4%	3	7%	4	7%	9	6%
People of the Americas	2	4%	2	5%	0	0%	4	3%
Sub-Saharan African	0	0%	0	0%	1	2%	1	1%
I don't identify with any cultural group	12	23%	6	14%	6	10%	24	16%
Total	52	100%	43	100%	59	100%	154	100%

As shown in Table 4, there was a mix of participants in each personality dimension. While most participants (44% on average) are in the medium range of each factor, some differences between the groups are notable. In the Neutral group, 51.2% of participants had medium level of emotional stability and 37.2% had low level of agreeableness; in the Empathic group, 53.8% had medium level of agreeableness; 52.5% of participants in the Adaptive group had medium level of openness; 15.3% of participants in the Adaptive group had low level of openness; in the Empathic group, 21.1% of participants had low level of conscientious and 44.2% had medium level conscientious. Participants in the Adaptive group were a little more open and agreeable, while participants in the Empathic group were slightly more conscientious, extraverted and emotionally stable.

Table 4. Personality Dimension Distribution by Group.

Personality Factors	Group	Openness		Conscientious		Extravert		Agreeable		Emotional Stability	
		N	%	N	%	N	%	N	%	N	%
Low	Empathic	16	30.8	11	21.2	16	30.8	11	21.2	13	25.0
	Neutral	14	32.6	20	46.5	15	34.9	16	37.2	10	23.3
	Adaptive	9	15.3	24	40.7	22	37.3	15	25.4	22	37.3
Medium	Empathic	18	34.6	23	44.2	20	38.5	28	53.8	21	40.4
	Neutral	18	41.9	11	25.6	14	32.6	16	37.2	22	51.2
	Adaptive	31	52.5	21	35.6	24	40.7	24	40.7	17	28.8
High	Empathic	18	34.6	18	34.6	16	30.8	13	25.0	18	34.6
	Neutral	11	25.6	12	27.9	14	32.6	11	25.6	11	25.6
	Adaptive	19	32.2	14	23.7	13	22.0	20	33.9	20	33.9

The results for DASS21 (Table 5) show that in total, normal levels of depression (D), anxiety (A) and stress (S) are 39%, 34% and 68%, respectively. Nineteen percent of participants had extreme level of anxiety (A) and this is shown slightly more in male than female. The level of normal stress was 68% where 67% of females and 71% of males reported having normal stress.

Table 5. DASS 21 Results by Gender (D: Depression; A: Anxiety; S: Stress).

	DASS 21 Categories	**Female**	**Male**	**Other**	**Total**	**Female**	**Male**	**Other**	**Total**
		N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
D	Normal	31	37	29	42	0	0	60	39
	Mild	16	19	20	29	0	0	36	23
	Moderate	25	30	14	20	0	0	39	25
	Severe	6	7	3	4	1	100	10	6
	Extr severe	6	7	3	4	0	0	9	6
A	Normal	29	35	24	35	0	0	53	34
	Mild	14	17	13	19	0	0	27	18
	Moderate	17	20	11	16	0	0	28	18
	Severe	9	11	7	10	0	0	16	10
	Extr severe	15	18	14	20	1	100	30	19
S	Normal	56	67	49	71	0	0	105	68
	Mild	9	11	8	12	0	0	17	11
	Moderate	9	11	6	9	1	100	16	10
	Severe	8	10	3	4	0	0	11	7
	Extr Severe	2	2	3	4	0	0	5	3

Table 6 shows that most participants in all groups had normal, mild or moderate level of depression (D) (i.e., 92% for empathic, 87% for neutral and 84%). The level of stress was normal for 79% in empathic, 70% in neutral and 58% in the Adaptive groups.

Table 6. DASS 21 Results by Group (D: Depression; A: Anxiety; S: Stress**).**

		Empathic	**Neutral**	**Adaptive**	**Empathic**	**Neutral**	**Adaptive**
		N	**%**	**N**	**N**	**%**	**N**
D	Normal	23	44%	17	40%	20	34%
	Mild	12	23%	9	21%	15	25%
	Moderate	13	25%	11	26%	15	25%
	Severe	1	2%	4	9%	5	8%
	Extr severe	3	6%	2	5%	4	7%
A	Normal	20	38%	13	30%	20	34%
	Mild	9	17%	6	14%	12	20%
	Moderate	11	21%	9	21%	8	14%
	Severe	5	10%	3	7%	8	14%
	Extr severe	7	13%	12	28%	11	19%
S	Normal	41	79%	30	70%	34	58%
	Mild	5	10%	9	21%	3	5%
	Moderate	3	6%	3	7%	10	17%
	Severe	1	2%	1	2%	9	15%
	Extr Severe	2	4%	0	0%	3	5%

5.2. Study Stress Score Result

Table 7 shows that the mean value of stress score for Adaptive group is higher than the other two groups (before and after the interaction with Sarah). Neutral group has the lowest average stress score after interaction with Sarah.

Table 7. Descriptive Statistics for Stress Score before and after interaction with Sarah.

	Empathic Group		**Neutral Group**		**Adaptive Group**	
	Mean	StDV	Mean	StDV	Mean	StDV
Score 1	4.68	2.46	4.83	2.41	5.31	2.00
Score 2	3.74	2.05	3.54	2.32	5.19	2.30
Difference	0.94	0.41	1.30	0.09	0.12	-0.70

A Wilcoxon Signed-Ranks test was conducted to compare the study stress score before (score 1) and after interaction (score 2) with Sarah, in each group. There is a significant difference in the score for before interaction and score after interaction in the Empathic group (mean rank = 15.92, z = −3.599, $p < 0.01$) and Neutral group (mean rank = 12, z = −4.590, $p < 0.01$). In contrast, in the Adaptive group there are no significant decreases in study stress scores (i.e., mean rank = 26.84, z = −0.581, $p = 0.568$).

A Kruskal–Wallis H test showed that there was not a statistically significant difference in study stress score 1 between the different groups, $\chi^2(2) = 1.601$, $p = 0.449$, with a mean rank of 72.41 for stress score for Empathic group, 76.22 for Neutral group and 82.93 for Adaptive group. However, there was a significant difference in study stress score 2 between the groups, $\chi^2(2) = 16.342$, $p < 0.001$, with a mean rank of 67.26 for stress score for Empathic group, 64.76 for Neutral group and 95.81 for Adaptive group.

The pairwise comparison test indicates that score 2 is not significantly different between empathic and neutral groups. In contrast, score 2 is significantly different between Neutral and Adaptive groups and Empathic and Adaptive groups.

A Kruskal–Wallis H test has been conducted and showed that there is a statistically significant difference in the amount of reduced study stress between the different groups, $\chi^2(2) = 9.655$, $p = 0.008$, with a mean rank of 78.26 for stress score for Empathic group, 92.98 for Neutral group and 65.55 for Adaptive group.

Pairwise comparison test shows that the differences between score1 and score2 are significantly different between Neutral and Adaptive groups. There are no statistically significant results in the other two groups' comparisons (Table 8).

Table 8. Score Differences Pairwise Comparison of Groups.

Pairwise Comparisons of Group				
Sample 1-Sample 2	Test Statistic	Standard Error	Significance	Adjusted Significance [a]
Adaptive-Empathic	12.709	8.383	0.130	0.389
Adaptive-Neutral	27.426	8.837	0.002	0.006
Empathic-Neutral	−14.717	9.085	0.105	0.316

[a] Significance values have been adjusted by the Bonferroni correction for multiple tests.

In order to identify significant associations between the study stress score and individual differences such as personality and DASS21, we categorized score1 and score 2 into three groups (i.e., low, moderate and high). We aimed to create the categories in a way that the frequencies of scores in all categories were similar. The low category covers the score range between 0 and 3. The score1 and score 2 ranges between 4 to 6 and 7 to 10 were classified as moderate and high, respectively. Table 9 shows the frequencies for each category within each group.

Table 9. Frequencies of Stress Score Categories in Each Group.

Score Category	**Empathic**				**Neutral**				**Adaptive**				**Total %**	
	Score 1		**Score 2**		**Score 1**		**Score 2**		**Score 1**		**Score 2**		**Score 1**	**Score 2**
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Low	18	12%	22	14%	14	9%	20	13%	10	6%	14	9%	27.3%	36.4%
Moderate	20	13%	26	17%	15	10%	18	12%	29	19%	27	18%	41.6%	46.1%
High	14	9%	4	3%	14	9%	5	3%	20	13%	18	12%	31.2%	17.5%

The chi-squared test and Fisher's exact test can assess for independence between two variables when the comparing groups are independent and not correlated [43]. We used Fisher's exact tests to answer the research questions related to whether the proportions for one variable are different among values of the other variable(s). When the sample size is bigger and expected values are at least 5 or more, we could have used the Chi-Square tests but for our sample we could not use Chi-Square test(s) since the assumption that expected values should be at least 5 was not satisfied, and the sample size(s) were small.

A Fisher's exact test revealed significant relationships in Empathic and Adaptive groups between a few variables as shown in Table 10. We compared the observed and expected counts of the relevant variables through the cross tabulations. A big difference between observed value and expected value implies that the assumption of independence between the variables tests is not true, in other words, it implies that the variables are associated with each other. The corresponding cross tab showed that in the Empathic group, the expected counts for participants with normal anxiety and low score 1 are higher than the observed counts. Similarly, the expected counts for normal anxiety and moderate score 1 are higher than the observed counts. For the relationship between score 2 and conscientiousness in the Empathic group, we observed a greater number of low score 2 for high conscientiousness than the expected count. Similarly, the observed count of high score 2 for low conscientiousness is more than the expected counts. Moreover, in the Empathic group, the observation of low score 2 for low depression and moderate score 2 for moderate depression are more than the expected counts. We also observed a smaller number of low score 2 for moderate depression and moderate score 2 for low depression than the expected counts. For anxiety and stress in the Empathic group, we observed a greater number of low score 2 for low anxiety and high number of low score 2 for low stress than the expected counts. Finally, in the Adaptive group, the observed count of low score 1 for high emotional stability and high score 1 for low emotional stability are more than the expected count. The expected count of low score 1 for low emotional stability is higher than the observed count in the Adaptive group.

Table 10. Association Results of Fisher's exact test.

Group	Variable 1	Variable 2	*p* Value
Empathic	Score 1	anxiety	0.037
	Score 2	conscientiousness	0.077
	Score 2	depression	0.046
	Score 2	anxiety	0.014
	Score 2	stress	0.011
Adaptive	Score 1	emotional stability	0.006

5.3. Rapport Result

Table 11 shows the mean and standard deviation for the rapport score sum and rapport average. Sum of the rapport is calculated by adding up the values assigned to each likert-scale. The observed total values varied from minimum 48 to maximum 66. The observed minimum rapport sum was 48 and the highest was 66. Table 11 also shows the result of the Kruskal–Wallis H test for rapport. The average and sum of the rapport scores are not significantly different between the different groups, $\chi^2(2) = 2$, $p = 0.308$ for rapport sum and $\chi^2(2) = 1$, $p = 0.601$ for rapport average. The mean ranks for each group are presented in Table 11.

Table 11. Mean, Mean Rank and Standard Deviation (SD) for Rapport Sum and Average (Kruskal–Wallis).

Groups	N	Rapport Sum			Rapport AVG		
		Mean	**SD**	**Mean Rank**	**Mean**	**SD**	**Mean Rank**
Empathic	52	56.65	2.821	71.26	3.21	0.49	78.70
Neutral	43	57.51	3.254	85.28	3.23	0.32	81.95
Adaptive	59	57.24	3.436	77.33	3.14	0.46	73.19
Total	154	57.12	3.187		3.19	0.42	

To identify any possible significant relationships between rapport and demographic categories, we categorised the rapport score total. To categorise rapport total into three similarly sized buckets capturing low, moderate and high rapport, as shown in Table 12, we have grouped the weighted numbers (i.e., 56, 57 and 58 total rapports) as moderate rapport. Less than 56 s were considered as poor and greater than 58 s were categorized as good rapport. No participants reported high rapport, which we considered would need a score of at least above 70 (mostly chosen “agree”). The total number of participants who established poor, moderate and good rapport, were 45, 60 and 49 respectively.

Table 12. The Number of Participants per Range of Rapport Score

Total_Rapport	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	65	66
no of participants	1	1	2	5	3	4	13	16	18	17	25	16	15	7	5	3	1	2
Rapport Categories	Low								Moderate			High						

Table 13 and corresponding Figure 7 show the level of rapport category for each group. The first column in Table 13 shows the number of participants and the second column represents the percentage within the group. In the Neutral group, 40% felt a good sense of rapport while this amount is 32% and 25% for Adaptive and Empathic groups, respectively. Neutral group has the highest level in moderate, and good rapport categories compare to the other two groups. In the full data, the level of moderate rapport is the highest. No one selected ’not applicable’ for any question. Furthermore, Chi-square tests reported no significant relationship between the rapport score and personality or psychological emotional state of the users. Finally, we did not find any significant relationship between the rapport score and participants’ study stress score after or before the interaction.

Table 13. Categorised Rapport Scores distribution across the Groups.

Rapport	**Empathic**		**Neutral**		**Adaptive**		**Total**	
	N	%	N	%	N	%	N	%
Poor	18	35	9	21	18	31	45	29
Moderate	21	40	17	40	22	37	60	39
Good	13	25	17	40	19	32	49	32
Total	52	100	43	100	59	100	154	100

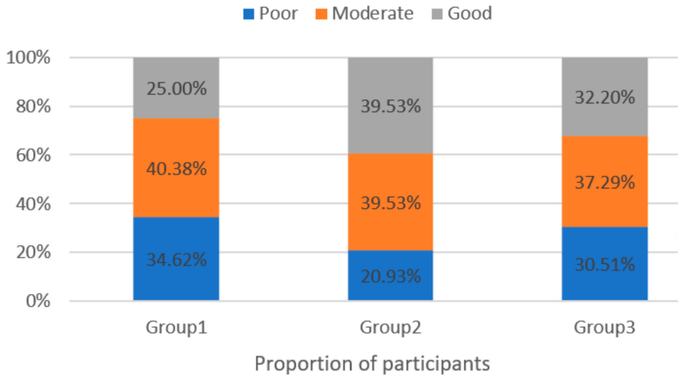


Figure 7. Categorised Rapport Scores by Groups

5.4. Accuracy of Empathic Cues

It was important to validate that the rules generated using data from previous studies accurately predicted the preferences of the participants in the Adaptive group. If accuracy was low, then Sarah would not be adapting according to their actual preferences. This would affect the interpretation of our comparisons between Groups Empathic, Neutral and Adaptive.

As a first analysis, we sought to determine how often RCs were triggered. If there was little difference between adaptive dialogue and the empathic or neutral dialogue, then we could not expect to see much difference between the groups. If RCs were not triggered very often, then the difference between the neutral (default) dialogue and the adaptive dialogue would be minimal, and thus differences between neutral and adaptive would be insignificant. However, if RCs were triggered very often, the adaptive dialogue and empathic dialogue would have been similar. For each participant, we calculated how many RCs had been triggered. With 10 RC, the possible range was 0–10, with 0 meaning no cues were triggered (neutral dialogue) and 10 meaning all cues were triggered (empathic dialogue). Figure 8 shows the total number of participants who received each frequency (from 0 to10) of RC cues. The number of triggered RCs varies from 0 to 8. Only three participants received six or more empathic cues. However, 48 participants received two, three or four empathic cues. This result indicates that the adaptive dialogue was more similar to the neutral dialogue since most of the users received four or less RCs.

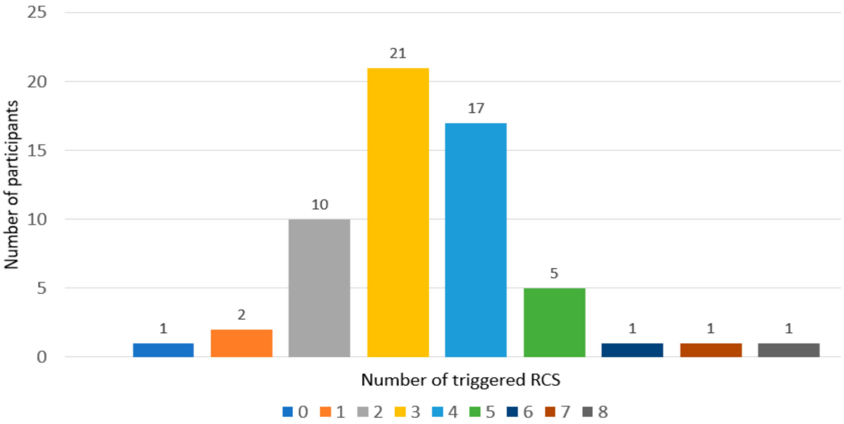


Figure 8. The distribution of Triggered RCs by the number of participants in Adaptive group.

Further, to see how good our rules are and whether they accurately predicted the preferences of users, we compared the triggered cues with their responses to whether they found 20 dialogue cues (2 examples of each of the 10 empathic dialogue cues) to be Helpful/Empathic or Stupid. Since helpful and empathic are both positive responses, scoring either of them by participants was counted as 1. For RC4, which is humor, we count the number of people who selected the stupid option instead of empathic/helpful because the number of participants who have found humor statement stupid in our studies is significantly higher than helpful and empathic. Hence, we decided to choose the stupid option for only RC4 and learnt rules to predict participants finding humor stupid and, if RC4 was NOT triggered, we added humor to the dialogue.

We compared each participant's survey responses to the 20 questions (10 statements × 2 questionnaire) with the triggered cues based on the rules for each participant and calculated four variables in each RC category, which are number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

In Table 14, the total number of the four variables is presented for all participants. TP is counted as 1, when the RCn has been triggered for the participant via the rules and s/he also found the RCn helpful and/or empathic through the survey after the interaction. On the other hand, FP is considered as 1 when the RCn has been triggered but the participant did not find it helpful and/or empathic. TN is the condition where the RCn has not been triggered and has not been desired by the participant. However, FN is when the RCn has not been triggered in the dialogue, although according to the survey, the participant found the RCn helpful and/or empathic. In the ideal scenario, if the rules generate the right RCs for all 59 participants (i.e., when the accuracy is 100%), the total number of TP and TN would be 10 (i.e., number of RCs) times 59 (i.e., number of participants), which is 590. However, the result shows that the prediction was right in 224 cases (i.e., TP + TN), and in 366 (i.e., 590 minus 224) cases the rules prediction was wrong. We can conclude that the overall prediction accuracy of the model is 38%.

Table 14. RC Questionnaires Result (H: Helpful, E: Empathic, S: Stupid).

RCs	RC Name	Total TP	Total FP	Total TN	Total FN	Accuracy	
						True%	False%
RC1	Social Dialogue	7	0	1	51	13.6	86.4
RC2	Meta-Relational Dialogue	17	1	1	40	30.5	69.5
RC3	Empathic Feedback	47	1	1	10	81.4	18.6
RC4	Humor	0	0	23	36	39	61
RC5	Continuity behaviors	16	1	4	38	33.9	66.1
RC6	Self-Disclosure	23	4	1	31	40.7	59.3
RC7	Reference to mutual/sharing knowledge	0	0	1	58	1.7	98.3
RC8	Solidarity and rapport-mirroring	29	2	2	26	52.5	47.5
RC9	Politeness	36	0	0	23	61	39
RC10	Inclusive pronoun	15	0	0	44	25.4	74.6
	Total	190	9	34	357		
	Accuracy/Error %	95%	5%	9%	91%		

We calculated the total number of cues being reported as helpful and/or empathic and the total number of triggered cues for all participants. In our analysis, total number of either helpful or empathic responses was calculated by identifying the total number of participants who found the RCn to be helpful and/or empathic in response to either of the sentences provided. To elaborate, in the first RC survey, 54 participants found *Social Dialogue* (RC1) helpful and/or empathic whereas 47 participants reported the same in the second survey. As depicted in Figure 9, the 54 participants who found RC1 helpful and/or empathic were distributed as follows: 43 participants in both the 1st and 2nd survey, 11 only in the 1st RC survey (survey 1), and 4 only in the 2nd RC survey (survey 2). If participants found either of those helpful and/or empathic, the rules were supposed to trigger the RC1 on. If we had to find both sentences helpful and/or empathic, there would be 43 participants that reported RC1 helpful/empathic (Figure 9).

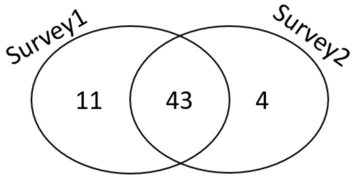


Figure 9. Comparison of the number of either helpful or empathic responses for RC1 examples in two surveys (Survey1 = RC Rating Survey before interaction; Survey2 = RC Rating Survey after interaction).

Although 58 participants reported RC1 as empathic and/or neutral, either in survey 1 or survey 2, RC1 was triggered (correctly) only for seven of them (TP). This means that 51 participants (FN) said RC1 was helpful/empathic but RC1 did not trigger for them. Our results show that there is high accuracy for TP and TN when the model makes a prediction to trigger the empathic cues. For instance, for RC1, the accuracy is 100% for both TP and TN, and for RC2, the accuracy of TP is 94%. Moreover, there are not many FPs in the results. In contrast, the number of FN is very high, which means participants would have wanted the empathic cues, but they were not triggered. The last two columns of Table 14 show the overall accuracy (Formula (1)) and inaccuracy (Formula (2)) of each RC (accuracy defined here as the lost opportunity to trigger the empathic cues). The accuracies of RC3, RC8 and RC9 are over 50%. Formula (3) to Formula (6) show the accuracy or inaccuracy rates for TN, TP, FN, and FP, respectively. The results are shown in Table 14.

$$\text{True} = (\text{TP} + \text{TN})/\text{Total Participants} \tag{1}$$

$$\text{False} = (\text{FP} + \text{FN})/\text{Total Participants} \tag{2}$$

$$\text{Accuracy} = \text{TN}/(\text{TN} + \text{FN}) \tag{3}$$

$$\text{Error} = \text{FN}/(\text{TN} + \text{FN}) \tag{4}$$

$$\text{Accuracy} = \text{TP}/(\text{TP} + \text{FP}) \tag{5}$$

$$\text{Error} = \text{FP}/(\text{TP} + \text{FP}) \tag{6}$$

The overall accuracy of the prediction for each user (True Positives plus True Negatives divided by total number of RCs each participant exposed to) shows that for 16 out of 59 participants, Sarah's dialogue had 50% or more accuracy. However, for 43 of them, the accuracy was lower than 50%. For one participant, the accuracy was 0%.

In Table 15, we compare predicted (based on rules generated from previous studies) with expected (based on Adaptive group survey responses) number of triggered RCs. Row 2 shows the number of participants who received 0–10 RCs. The last row shows the number of RCs that Adaptive group participants considered to be Helpful and/or Empathic and indicates that all participants would have liked to receive at least six cues and more than half of them (52.5%) wanted to receive all of the cues and 33.8% wanted almost all (i.e., 9) of the cues.

Table 15. Comparison of Expected versus Predicted Total Number of Triggered RCs.

No of Triggered RCs	0	1	2	3	4	5	6	7	8	9	10	Total
No of Participants from rules	1	2	10	21	17	5	1	1	1	0	0	59
No of Participants from survey	0	0	0	0	0	0	1	4	3	20	31	59

6. Discussion

Participants in the Adaptive group had more stress than the Neutral group. Nine percent of participants had moderate to extremely severe psychological stress in the Neutral group, but this amount was 37% in the Adaptive group. In general, females had more stress than males. Twenty-three percent and 17% of females and males had moderate to extremely severe levels of stress, respectively. Considering that in the Adaptive group there are eight more females than the Neutral group and females appear more stressed, this could also contribute to better reduced study stress in the Neutral group. Further analyses of expected counts revealed that participants in the Empathic group were less stressed, depressed and anxious, and more conscientious than expected, in contrast with the other two groups.

To answer the first research question, we consider changes in reported study stress levels for each group. The difference between score1 and score 2 (i.e., score1 minus score 2) is significantly different between Neutral and Adaptive groups. This means that interacting with Sarah was beneficial for the two groups. The finding is in line with the literature that regardless of the type of the character's dialogue (i.e., Empathic or Neutral) the study stress before and after interaction with the character are significantly different [13,36]. The Adaptive group did not receive any significant benefit from the interaction. Thus, while we did not find any significant differences between the three groups for baseline score (i.e., score 1), for final study stress (i.e., score 2), there is a significant difference between Adaptive group and the other two groups. Even though there are no significant differences in baseline study stress between groups, the baseline study stress score was highest in the Adaptive group. We believe that the reason for that could be the timing of experiment, which was close to exam time. Many of the tips provided concern good study habits that need to be applied throughout the semester. Since students were soon to do their exams at the end of the semester, it was probably too late to change these behaviours. The higher baseline study stress score and result of DASS 21 show that students in the Adaptive group experienced more stress and anxiety than the other two groups. Higher study stress score provides a greater opportunity to show improvement, but perhaps the higher levels of psychological stress and anxiety made these students less able to gain benefits from the tips. The dialogue in the Adaptive group was a mix of sentences from the dialogues experienced by Neutral and Empathic groups. In adaptive dialogues from 10 relational cues, zero to eight of the cues were triggered for participants. However, our analyses revealed that the accuracy of our rules for the Adaptive group was poor, and as a result, the adapted dialogue was not suitably tailored to most individuals. We conclude that poor tailoring, higher stress levels and receiving the tips too late, could be the reasons why the Adaptive group did not gain significant benefit from the interaction.

To answer the second research question, although participants in all groups reported varying levels of rapport ranging from low to high, with the majority establishing medium levels of rapport, we did not find any significant differences between the three groups for rapport score. Participants in the Neutral group who interacted with Neutral Sarah reported the highest levels of rapport. The results of rapport in adaptive Sarah were slightly less than neutral Sarah. Empathic group had the lowest rapport score. The higher reported rapport by the Neutral group is consistent with findings that participants who feel less emotional intensity about problems they are facing will build more rapport with a character that uses neutral language, whereas participants with high emotional feelings will build more rapport with an empathic character [13]. This also explain the low rapport score in the Empathic group, as they reported lower study stress and significantly lower DASS21 results. Perhaps if participants had been more highly stressed, the benefit of using empathic language (either fully or partially/tailored empathic) would have been evident. We further conjecture, that the Empathic group might have built more rapport with neutral Sarah, since they were least stressed. Since our Adaptive group was highly stressed but only received around four empathic cues on average, resulting in a dialogue more similar to the Neutral Group than the Empathic group, we believe the tailoring was not conducive for building rapport with this cohort of participants.

Another explanation may be related to the study design. The duration of the study was around 20 min for empathic and 15 min for neutral and varies for adaptive Sarah based on the triggered cues. Time might influence the sense of rapport students feel towards the character. Returning to the first research question, time might also impact on students' study stress level. We hypothesise that visiting adaptive Sarah for more than one session over multiple days and in shorter sessions may result in a better outcome. Finally, since no participants selected "not applicable" to any of the rapport questions, we conclude that participants considered Sarah to have social qualities and did not consider her to be merely a computer or machine.

To answer the third research question, there is no significant relationship between the rapport score and participants' study stress score after or before the interaction.

This study sought to demonstrate the benefit of a tailored interaction with an IVA by adapting the agent's dialogue based on rules learnt from interactions with previous individuals. However, analysis of the tailoring revealed poor accuracy. The accuracies of TP and FP are 95%, whereas the accuracies of TN and FN are 9%. This indicates that overall, 95% of the triggered RCs had been correctly triggered and 5% should not have been triggered. Furthermore, 9% of the non-triggered RCs had correctly not been triggered but 91% were incorrectly not triggered. Moreover, from the survey, participants found minimum 6 and maximum 10 cues helpful/empathic, but the rules generated 2, 3, 4, or 5 cues for them. To generate the adaptive rules, we tried different models (i.e., C5.0, CHAID) and selected the best accurate ones. To achieve the best decision tree, where the predictors were not generated by default values in misclassification matrix of SPSS Modeler, misclassification cost was used to include information about the relative penalty associated with incorrect classification. The accuracy of our trees was less than 75%. The decision trees were often over-fitting, and therefore, sometimes they needed to be pruned. As a result, the decision tree and corresponding rules could have been overfitting to the data at hand. In some rules, misclassification costs were used to obtain the decision tree with predictors. These two reasons could result in having 357 FNs. Although the rules were successful in selecting some individuals accurately, the rules had low coverage and failed to identify the majority of participants who would have preferred to receive the empathic cue. Finally, Table 10 shows the correct and wrong predictions for each triggered RC. Rules learnt on previous users did not cover Adaptive group unseen cases. We had to relax certain constraints in order to generate rules. The prediction accuracy for True Positive (TP) was 95% and for False Negative (FN) was 91%.

The number of participants who received the right predictions in RC3, RC8 and RC9 is more than the ones who have not received the correct prediction. We believe that since in RC surveys, participants were only asked to respond to two samples of each RCn, that might not be a valid indicator to measure if the participant likes all the RCn statements. This is also reflected in our analyses that shows that not everyone who found one sentence helpful/empathic also considered the other sentence to be helpful/empathic.

This was our first attempt of deploying an adaptive agent who adapts to the user on the fly based on the IVA's expertise gained from earlier interactions with other individuals. Our results indicate that we have to improve all the RCs rules, especially the rules for those cues where the number of triggered RCs is far less that the number of desired RCs.

7. Conclusions, Limitation and Future Work

In this paper, we reported our design, development and evaluation of our first generation of adaptive virtual advisor. Our proposed extended architecture includes a repository of User Profiles and the agent's Expertise module. These two components along with the Decision Making and Knowledge Base modules of FAtiMA formed an Adaptive Engine in the agent architecture. We conducted an evaluation of the Adaptive Engine by conducting an experiment where participants randomly interacted with Neutral Sarah (no empathic dialogue cues), Empathic Sarah (fixed empathic cues) or Adaptive Sarah (empathic cues tailored to the features of the user). According to the results of our experiments, the benefits of tailoring were not apparent and participants in the Neutral group received the most benefit for reducing study stress.

The current expertise module contains the rules, which were derived by using machine learning methods on our previous studies data. While the rules would identify whether a user likes a social dialogue (like RC1), it does not distinguish which type of social dialogues the person prefers. Currently, if RC1 is triggered, all the 'social dialogue' sentences will be uttered by Sarah. In the future, by using relative feedback, the agent can receive the feedback from user's preferences on each empathic cue. We believe that by collecting larger sample size data, we will have a better understanding of how to improve the system.

In FAtiMA, what is called personality of the character is defined by a set of goals, emotional reaction rules and the character's action tendencies, which all are in deliberative layer and based on OCC model of emotion. In future, we can take those into account to create adaptive agents with different personalities.

As an alternative to evaluate the value of an Adaptive agent with tailored dialogue, we are currently running another experiment that seeks to avoid the cold start and rule accuracy problem we faced in this study. In this new study, we directly ask participants what types of relational cues they would like the character to use. Based on the users' choices, the agent adapts accordingly on the fly. In the future, we aim to improve the Adaptive Engine by receiving feedback from the user in real time to find out if they like or do not like certain language and update the user profile and expertise engine accordingly by using incremental knowledge acquisition methods, such as ripple-down rules [44].

Author Contributions: Conceptualization, D.R., H.R.; software, H.R., C.K., S.M.; validation, H.R., D.R.; Methodology, D.R., H.R.; investigation, H.R., D.R.; formal analysis, H.R., D.R., A.A.B.; data curation, H.R., D.R., A.A.B.; writing original draft preparation, H.R., D.R.; writing review and editing, D.R., A.A.B., C.K., S.M., H.R.; supervision, D.R., C.K., S.M.; resources, Macquarie university provided us with the resources such as Qualtrics survey and participants for the studies. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Thanks to Meredith Porte for technical assistance and to all participants involved in the studies.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morency, L.P.; Stratou, G.; DeVault, D.; Hartholt, A.; Lhommet, M.; Lucas, G.; Morbini, F.; Georgila, K.; Scherer, S.; Gratch, J.; et al. SimSensei Demonstration: A Perceptive Virtual Human Interviewer for Healthcare Applications. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
2. Mancini, M.; Niewiadomski, R.; Bevacqua, E. Greta: A SAIBA compliant ECA system. In Proceedings of the Troisiéme Workshop sur les Agents Conversationnels Animés, Paris, France, 27–28 November 2008.
3. Lisetti, C.; Amini, R.; Yasavur, U.; Rishe, N. I can help you change! An empathic virtual agent delivers behavior change health interventions. *ACM Trans. Manag. Inf. Syst. (TMIS)* **2013**, *4*, 19. [CrossRef]
4. Dias, J.; Mascarenhas, S.; Paiva, A. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion Modeling*; Springer: Cham, Germany, 2014; pp. 44–56.
5. Traum, D.; Marsella, S.; Gratch, J. Emotion and dialogue in the MRE virtual humans. In *Tutorial and Research Workshop on Affective Dialogue Systems*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 117–127.
6. Gratch, J.; Marsella, S. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, QC, Canada, 28 May–1 June 2001; pp. 278–285.
7. Lim, M.Y.; Dias, J.; Aylett, R.; Paiva, A. Creating adaptive affective autonomous NPCs. *Auton. Agents Multi-Agent Syst.* **2012**, *24*, 287–311. [CrossRef]
8. Bickmore, T.; Gruber, A.; Picard, R. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Educ. Couns.* **2005**, *59*, 21–30. [CrossRef] [PubMed]
9. Sandbank, T.; Shmueli-Scheuer, M.; Herzig, J.; Konopnicki, D.; Shaul, R. Ehctool: Managing emotional hotspots for conversational agents. In Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion, Limassol, Cyprus, 13–16 March 2017; pp. 125–128.
10. Brave, S.; Nass, C.; Hutchinson, K. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Int. J. Hum. Comput. Stud.* **2005**, *62*, 161–178. [CrossRef]
11. Leite, I. Using adaptive empathic responses to improve long-term interaction with social robots. In *User Modeling, Adaption and Personalization*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 446–449.

12. Bickmore, T.W.; Picard, R.W. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact. (TOCHI)* **2005**, *12*, 293–327. [CrossRef]
13. Ranjbartabar, H.; Richards, D.; Bilgin, A.; Kutay, C. First Impressions Count! The Role of the Human's Emotional State on Rapport Established with an Empathic versus Neutral Virtual Therapist. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
14. Gratch, J.; Lucas, G.M.; King, A.A.; Morency, L.P. It's only a computer: The impact of human-agent interaction in clinical interviews. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, Paris, France, 4–8 May 2014; pp. 85–92.
15. Ochs, M.; Sadek, D.; Pelachaud, C. A formal model of emotions for an empathic rational dialog agent. *Auton. Agents Multi Agent Syst.* **2012**, *24*, 410–440. [CrossRef]
16. Dupuy, L.; De Sevin, E.; Ballot, O.; Cassoudesalle, H.; Dehail, P.; Aouizerate, B.; Cuny, E.; Micoulaud-Franchi, J.-A. A Virtual Patient to Train Semiology Extraction and Empathic Communication Skills for Psychiatric Interview. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, Glasgow, UK, 19–23 October 2019; pp. 188–190.
17. Shea, S.C. *Psychiatric Interviewing E-Book: The Art of Understanding: A Practical Guide for Psychiatrists, Psychologists, Counselors, Social Workers, Nurses, and Other Mental Health Professionals, with Online Video Modules*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2016.
18. McQuiggan, S.W.; Lester, J.C. Modeling and evaluating empathy in embodied companion agents. *Int. J. Hum.-Comput. Stud.* **2007**, *65*, 348–360. [CrossRef]
19. Stal, S.T.; Kramer, L.L.; Tabak, M.; den Akker, H.O.; Hermens, H. Design Features of Embodied Conversational Agents in eHealth: A Literature Review. *Int. J. Hum. Comput. Stud.* **2020**, *138*, 102409.
20. Caridakis, G.; Raouzaiou, A.; Bevacqua, E.; Mancini, M.; Karpouzis, K.; Malatesta, L.; Pelachaud, C. Virtual agent multimodal mimicry of humans. *Lang. Resour. Eval.* **2007**, *41*, 367–388. [CrossRef]
21. Paiva, A.; Leite, I.; Boukricha, H.; Wachsmuth, I. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2017**, *7*, 11. [CrossRef]
22. Bevacqua, E.; Mancini, M.; Pelachaud, C. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 262–269.
23. Dörner, D. The mathematics of emotions. In Proceedings of the Fifth International Conference on Cognitive Modeling, Bamberg, Germany, 10–12 April 2003.
24. Kopp, S.; Krenn, B.; Marsella, S.; Marshall, A.N.; Pelachaud, C.; Pirker, H.; Thórisson, K.R.; Vilhjálmsson, H. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 205–217.
25. DeVault, D.; Artstein, R.; Benn, G.; Dey, T.; Fast, E.; Gainer, A.; Georgila, K.; Gratch, J.; Hartholt, A.; Lhommet, M.; et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems; 2014; pp. 1061–1068.
26. Zintgraf, L.M.; Roijers, D.M.; Linders, S.; Jonker, C.M.; Nowé, A. Ordered preference elicitation strategies for supporting multi-objective decision making. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 1477–1485.
27. Kingsley, D.C.; Brown, T.C. Preference uncertainty, preference learning, and paired comparison experiments. *Land Econ.* **2010**, *86*, 530–544. [CrossRef]
28. Verhagen, T.; Feldberg, F.; van den Hooff, B.; Meents, S.; Merikivi, J. Understanding users' motivations to engage in virtual worlds: A multipurpose model and empirical testing. *Comput. Hum. Behav.* **2012**, *28*, 484–495. [CrossRef]
29. Chen, A.T.; Wu, S.; Tomasino, K.N.; Lattie, E.G.; Mohr, D.C. A multi-faceted approach to characterizing user behavior and experience in a digital mental health intervention. *J. Biomed. Inform.* **2019**, *94*, 103187. [CrossRef] [PubMed]
30. Mascarenhas, S.; Dias, J.; Prada, R.; Paiva, A. A dimensional model for cultural behavior in virtual agents. *Appl. Artif. Intell.* **2010**, *24*, 552–574. [CrossRef]
31. Marsella, S.C.; Johnson, W.L.; LaBore, C. Interactive pedagogical drama. In Proceedings of the International Conference on Autonomous Agents, Barcelona, Spain, 3–7 July 2000; Volume 3, pp. 301–308.
32. Aylett, R.S.; Louchart, S.; Dias, J.; Paiva, A.; Vala, M. FearNot!—An experiment in emergent narrative. In *International Workshop on Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 305–316.

33. Rodrigues, S.H.; Mascarenhas, S.F.; Dias, J.; Paiva, A. "I can feel it too!": Emergent empathic reactions between synthetic characters. In Proceedings of the 2009. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–7.
34. Aylett, R.; Vannini, N.; Andre, E.; Paiva, A.; Enz, S.; Hall, L. But that was in another country: Agents and intercultural empathy. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, Budapest, Hungary, 10–15 May 2009; Volume 1, pp. 329–336.
35. Bickmore, T.; Schulman, D.; Yin, L. Maintaining engagement in long-term interventions with relational agents. *Appl. Artif. Intell.* **2010**, *24*, 648–666. [CrossRef] [PubMed]
36. Ranjbartabar, H.; Kutay, C.; Richards, D.; Mascarenhas, S. Towards an Adaptive System: Users' Preferences and Responses to an Intelligent Virtual Advisor based on Individual Differences. *Inf. Syst. Dev.* **2018**. Available online: https://aisel.aisnet.org/isd2014/proceedings2018/HCI/6/ (accessed on 15 August 2020).
37. Goldberg, L.R. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personal. Psychol. Eur.* **1999**, *7*, 7–28.
38. Henry, J.D.; Crawford, J.R. The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* **2005**, *44*, 227–239. [CrossRef]
39. Tickle-Degnen, L.; Rosenthal, R. The nature of rapport and its nonverbal correlates. *Psychol. Inq.* **1990**, *1*, 285–293. [CrossRef]
40. Astrid, M.; Krämer, N.C.; Gratch, J.; Kang, S.-H. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Comput. Hum. Behav.* **2010**, *26*, 1641–1650.
41. McCroskey, J.C.; Hamilton, P.R.; Weiner, A.N. The effect of interaction behavior on source credibility, homophily, and interpersonal attraction. *Hum. Commun. Res.* **1974**, *1*, 42–52. [CrossRef]
42. McCroskey, J.C.; McCain, T.A. The measurement of interpersonal attraction. *Speech Monogr.* **1974**, *41*, 261–266. [CrossRef]
43. Kim, H.-Y. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restor. Dent. Endod.* **2017**, *42*, 152–155. [CrossRef] [PubMed]
44. Kang, B.; Compton, P.; Preston, P. Multiple classification ripple down rules: Evaluation and possibilities. In Proceedings of the 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, 26 February–3 March 1995; Volume 1.