PUCRS

FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

PAULO RICARDO KNOB

MODELING AN EMPATHETIC EMBODIED CONVERSATIONAL AGENT

Porto Alegre
2022

Pontifícia Universidade Católica
do Rio Grande do Sul

Pontifical Catholic University of Rio Grande do Sul
Faculty of Informatics
Computer Science Graduate Program

MODELING AN EMPATHETIC EMBODIED CONVERSATIONAL AGENT

PAULO RICARDO KNOB

Dissertation submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fullfillment of the requirements for the degree of Ph. D. in Computer Science.

Advisor: Profa. Dra. Soraia Raupp Musse

Porto Alegre
2022

Ficha Catalográfica

K72m Knob, Paulo Ricardo

Modeling an Empathetic Embodied Conversational Agent / Paulo Ricardo Knob. – 2022.
119 p.
Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Soraia Raupp Musse.

1. Agentes virtuais. 2. Simulação. 3. Reconhecimento de Faces. 4. Detecção de Emoções. 5. Bate-papo. I. Musse, Soraia Raupp. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Loiva Duarte Novak CRB-10/2079

Paulo Ricardo Knob

MODELING AN EMPATHETIC EMBODIED CONVERSATIONAL AGENT

This Doctoral Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on August 22nd, 2022.

COMMITTEE MEMBERS:

Prof. Dr. Emerson Cabrera Paraiso (PPGIA/PUCPR)

Prof. Dr. Isabela Gasparini (CCT/UDESC)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Prof. Dr. Soraia Raupp Musse (PPGCC/PUCRS - Advisor)

“Twenty years from now you will be more disappointed by the things that you did not do than by the ones you did do.”
(Mark Twain)

Acknowledgments

I would like to express, at first, my deepest thanks to my family, for their unwavering support and belief. Also, I would like to thank all the friends I have made in the Virtual Humans LAB (VHLAB), both the ones who are not there anymore and the ones who are still conducting amazing research projects. In addition, i would like to extend my thanks to my co-advisor, Catherine Pelachaud, for her patience, time and endless suggestions which enriched this work (it is a shame that COVID impeded me from visiting you and your team in Paris!). Finally, but not less important, I would like to thank my advisor Soraia Musse, for putting up with me for all those last six years (between mastery and PhD) and for being a true leader in a world of bosses.

Modelando um Agente de Conversação Personificado Empático

RESUMO

A empatia pode ser vista como um comportamento sócio-emocional complexo, que é resultado da interação entre tanto dispositivos cognitivos como afetivos e é responsável pelo fato de uma pessoa ser capaz de identificar e imitar emoções de outras pessoas, por exemplo. Além disso, a memória humana é uma ferramenta poderosa que permite a cada pessoa armazenar e recuperar informações sobre quase tudo o que acontece em sua vida. Equipar um agente conversacional incorporado (ECAs) com a capacidade de empatia, assim como outros recursos como memória, podem ajudar a tornar a interação com seres humanos mais fácil e natural. Este trabalho tem como objetivo propor e construir um agente conversacional empático dotado de uma memória similar à humana. Além de poder conversar com uma pessoa, é também capaz de mostrar certa extensão de empatia por essa pessoa. Além disso, este modelo dota o agente virtual com algumas outras habilidades, como reconhecer a pessoa com quem está conversando pela face e armazenar/recuperar informações com base em um modelo de memória humana. Alguns experimentos foram realizados para coletar informações quantitativas e qualitativas, as quais mostram que o modelo proposto funciona como pretendido. Finalmente, alguns caminhos para futuros trabalhos também são apresentados, esclarecendo o que está planejado ser feito para melhorar a qualidade deste trabalho.

Palavras-Chave: Agente virtual, Simulação, Reconhecimento de faces, Detecção de emoções, Bate-papo, Agente empático, Agente conversacional incorporado, Memória.

Modeling an Empathetic Embodied Conversational Agent

ABSTRACT

Empathy can be seen as a complex socio-emotional behavior, which is a result from the interaction between both cognitive and affective devices and is responsible, for instance, for one person being able to identify and mimicry others emotion. Moreover, the human memory is a powerful tool which allows for each person to be able to store and retrieve information about almost everything that happens. Equipping an embodied conversational agent (ECAs) with the ability of empathy, as well other features like memory, can help to make the interaction with humans smoother and more natural. This work aims to propose and build an empathetic conversational agent endowed with a human-like memory. Besides being able to chat with a person, it is also able to show some extent of empathy by such person. Also, this model endow the virtual agent with a few other abilities, like recognizing the person it is talking to by its face and store/retrieve information based on a human memory model. Some experiments were conducted to gather both quantitative and qualitative information, which show that the proposed model works as intended. Finally, a few avenues for future work are also presented, elucidating what it is planned to do aiming to improve the quality of this work.

Keywords: Virtual Agent, Simulation, Face Recognition, Emotion Detection, Chat, Empathetic Agent, Embodiment Conversational Agent, Memory.

List of Figures

List of Tables

List of Acronyms

AM-ART	Autobiographical Memory-Adaptive Resonance Theory
ECA	Embodied Conversational Agent
ECC	Ensemble Classifier Chain
ECM	Emotional Chat Machine
EIP	Emotion Interaction Patterns
ELU	Exponential Linear Unit
ESK	Event-Specific Knowledge
GRU	Gated Recurrent Unit
HOMER	Hierarchy of Multi-label Classifiers
LSTM	Long-Short Term Memory
LTM	Long-Term Memory
MRFN	Multi-Representation Fusion Network
NLP	Natural Language Processing
NLPCC	Natural Language Processing & Chinese Computing
NLTK	Natural Language ToolKit
NPC	Non-Playable Character
OCEAN	Openness Conscientiousness Extraversion Agreeableness Neuroticism
PAD	Pleasure Arousal Dominance
STM	Short-Term Memory
SRU	Simple Recurrent Units
SVM	Support Vector Machine

Contents

1. INTRODUCTION

Human beings are the only known species that use spoken language to communicate, having a skill developed in communication that uses factors other than speech, such as, for example, body expressions and gazing [CSCP00]. A concept that is studied as relevant during communication is empathy, which is the sharing of emotions between individuals, as well as the behavior of adopting another person's point of view [dWP17]. For example, if someone is talking with a person who just lost a beloved relative, he/she can perceive this person is truly sad and also feel sadness as well. Therefore, facial expressions are linked to the content of speech, emotion and personality, as well other behavioral variables, being even able to replace sequences of words, accompany them and be used to help disambiguate what is being said when the acoustic signal is degraded [CPB+94]. In fact, facial expressions may be used to communicate and influence other's behavior [AC76, CC14]. Thus, it seems important to detect and understand facial expressions in every interaction between people.

Still talking about empathy, it is known that it can evoke altruistic and prosocial behavior, besides being able to have a positive effect on relationships, enhance communication and help to mitigate aggressive behavior [Omd14]. In addition, it was already argued that emotions have both cognitive and social functions, which are important when developing an intelligent system [Min91]. Considering all these finds, the study of empathy for virtual agents seems important, both to evoke empathy on users and to act in an empathetic way towards said user.

Regarding the facial expressions subject, one topic which deserves attention is human eyes. It is common belief that the eyes are the window to one's soul, which elucidates the importance of the eyes in face-to-face communication. Besides the expressions eyes can borrow, their movement is also important. In addition to voluntary movement, the involuntary movement performed by human eyes is known as saccade, where they go from one gaze position to another [LZ15]. To minimize the time in transit, as well the time used to make some corrective movements, this phenomena must find a balance between two conflicting demands: speed and accuracy [LBB02].

Leaving the facial expressions topic, another important subject concerning human behavior is human memory. In general terms, the human memory is divided in three parts: Sensory Store, Short-term Store and Long-term Store [LL19]. The

Sensory Store is where a given new information (for example, a car seen by the eyes) is stored. Despite it being able to store a large amount of information, it can not do so for too long; it takes around just one second for such information to vanish. Some of this information is transferred to the Short-term Store, so it is not lost so quickly. This store has a small capacity, but can keep the information for around fifteen seconds before it vanishes. While the information is inside the Short-term Store, it can be transferred to the Long-term Store, which is a virtually unlimited store that has all the information which is always available, such as our own names, the ability to speak, the days of the week and so on [LL19].

Embodied Conversational Agents (ECAs) are virtual agents which are able to interact and talk with humans in a natural way. In the last years, many research was made to improve the quality of the communication abilities of such ECAs, both verbal and non-verbal [Yal20, BWM+19, SHCK19]. A fair amount of effort is being directed on ECAs which can help people to have a healthier life [KtSM+19, SCCG20, DBODAH19], for clinical interviews [PDA+20, MMMR+19] and the training of some skill [CW19, AHS19].

Following this line of research, in this thesis it is aimed to propose an empathetic Embodied Conversational Agent (ECA) with general purpose endowed with many abilities. We developed both a 2D and a 3D model for our ECA, called Arthur (2D) and Bella (3D). We did so both to give more options for the user to choose and to investigate the difference in perception by the users. Besides a conversational module, using text and voice, this ECA is able to recognize the person he/she is talking to, as well to assess the user emotional state through his/her facial expressions. Also, Arthur/Bella is able to demonstrate different levels of emotion through his/her facial expressions, being also endowed with an Empathy Module. Lastly, it is equipped with a memory module, which tries to replicate the behavior of human memory and, thus, allows for Arthur/Bella to learn information with and from the user, while interacting; and to remember it later in the conversation or, even, in a different interaction. The empathy is built in the communication with the user in mainly three parts of our model: firstly, through a pre-defined module of communication where Arthur asks questions about the user (demonstrating interest in the conversation); then, in the module of memory once the user feels that Arthur remember him/her; and finally with simple facial expressions that Arthur applies as a result of detecting facial expression module of Arthur. Next, it is presented the research problem, as well its relevance on the field.

1.1 Research Problem

The problematic addressed in this doctorate thesis aims to develop an ECA endowed with an empathy behavior and a human-like memory, among other features. Such features should help this virtual agent to present a more natural interaction behavior, as observed in interactions between human beings.

In the literature, there are many work proposing ECAs for different goals. Some work evaluate the use of personality on ECAs [SHCK19], while others aim to measure the impression of the user while interacting with the virtual agent][BWM+19], and so on. In general, the main applications found for ECAs in last years are medical purposes [KtSM+19, DBODAH19, SCCG20] and skill training [AHS19, CW19]. Work which try to model empathy in ECAs are still uncommon, specially because of the challenges it involves [Yal20].

The motivation for this research is related with the need for modeling better socially-aware virtual agents. Despite the existence of numerous proposals for ECAs and virtual agents in general [Yal20, BWM+19, SHCK19, CRP19, ZMP+18], each one with its own contribution to the area, there are still many avenues to explore concerning human behavior. Besides empathy, another example that could be explored is the human memory behavior. A virtual agent endowed with such tool would possess the ability to store/retrieve information in a similar way that human brain does and, thus, would be able to behave in a more human-like way. On this matter, there are several work which try to model some level of memory [KRK13, WTM16, EMJ18], each of them having different applications for the memory model. Despite all the important work made so far, no model was found which was able to deal with both memory and empathy behaviors. Such composed model could be very useful to improve the social and emotional behavior of an Embodied Conversational Agent, because many events that happen in people lives have some kind of feeling/emotion attached and, thus, such events could be remembered by the agent with this emotion together, allowing it to react in an empathetic way.

The applications of such ECA are many. As for personal assistants, a human-like memory module can help the agent to learn things about the person and give more personified answers/suggestions. There are several ECAs models aimed at medical clinical interviews, which could take advantage of an empathetic behavior. Moreover, the area of games could use and empathetic agent with a human-like memory, so Non-Playable Characters (NPCs) could interact with the player in a more natural way.

To deal with these problems, it is proposed the development of an Embodied Conversational Agent which has both human-like memory and an empathetic behavior, among other features. Next, are presented the goals (main and specifics) of this thesis.

1.2 Goals

The main goal of this thesis consists on the **proposal of a model of a multi-purpose empathetic Embodied Conversational Agent (ECA) endowed with several abilities**. Such abilities are used to improve the communication skills of the virtual agent and provide a smoother and more natural interaction with people. For example, an empathetic behavior can make people more comfortable to interact with the agent. Moreover, the use of a human memory model should help the virtual agent to store/retrieve information in a more human way, which can help it to deliver meaningful responses to the user. Next, the specific goals are presented.

1.2.1 Specific Goals

In order to achieve the main goal, some specific goals are proposed, as follows:

- **Virtual agent**. Proposal of the interface of the virtual agent, with its basic configurations and interfaces;
- **Basic Chatbot**. Proposal of the basic chatbot, allowing for human users to interact with the virtual agent at some extent;
- **Face Recognition**. Proposal of a module which allows the virtual agent to recognizes a person it already seen before. Alongside with the memory module (listed below), it should allow the agent to meet new people, recognizes and learn things about him/her;
- **Emotion Detection**. Proposal of a module which allows the agent to infer the emotion that a person is experiencing, based on his/her facial expression.

Alongside with the memory module (listed below), it should allow the agent to learn the emotions attached to events learned (e.g. the death of a close relative should be attached to a sad emotion). Also, it can help the agent to attune with an empathetic behavior;

- **Agent Memory**. Proposal of a memory module for the virtual agent, based on human memory models. It should allow the agent to store/retrieve information and, lastly, deliver meaningful responses to the user;

- **Empathy**. Proposal of an empathy module for the virtual agent, based on literature models. It should allow the agent to behave in an empathetic way towards the user;

- **Natural Language Processing**. Proposal of a module which deals with the processing of the sentences written/spoken by the user. Such sentences can be divided into tokens and stored at the memory. Also, the classification of each token (e.g. noun, verb) can be useful to find the topic of the conversation;

- **Appearance and Expressiveness**. Refinement of both appearance and expressiveness of the virtual agent. It was chosen a cartoon-like appearance, while the expressiveness deals with the facial expressions shown by the virtual agent given an emotion or behavior;

Next, the textual structure of this work is presented.

1.3 Text Structure

This work is divided into five chapters. This chapter presented an introduction about the subject of this thesis, presenting the research problem and its relevance, as well the goals of this work.

The Chapter 2 aims to present many related work concerning the subject of this thesis. Such works involve concepts related with chatbots, human and agents memory, embodied conversational agents and its appearance/expressiveness.

The Chapter 3 presents the proposed model. This model describes how each feature is built and works separately, as well how they are assembled together into a virtual agent.

The Chapter 4 presents and discuss the preliminary results achieved by this work so far. Such results involve tests about each feature, as well to discover if the virtual agent is working as expected.

Finally, The Chapter 5 concludes the work and presents the final considerations.

2. RELATED WORK

This Chapter presents many work which both borrow a theoretical foundation for this thesis and are related with the goals this work aims to achieve. Section 2.1 aims to present theory and work concerning chatbots, it means, intelligent interfaces which can talk with humans. Section 2.2 aims to explain the behavior of human memory, as well to present existing work in modeling such complex behavior into virtual agents. Section 2.3 discuss Embodied Conversational Agents (ECA) and the manner that a virtual agent presents itself, talking about its appearance and expressiveness.

2.1 Chatbots

As commented by Mathur et al. [ML19], a chatbot, or conversational agent, is a "computer program that can hold a conversation through text or speech". Also, the authors state that, for some tasks, these conversational agents do not need to deliver high-quality responses. Therefore, they propose a model that consumes fewer resources and is able to augment conversation data without increasing the size of the vocabulary. They use a modified version of the GRU (Gated Recurrent Unit) instead of the LSTM (Long short-term memory) to encode and decode sequences of text, which reduces the need for computational resources. GRUs are a gating mechanism in recurrent neural networks: in short, they work just like a LSTM with a forget gate. As one would expect, training and validating such chatbot models require large dataset containing dialogs between people. For their work, they choose to work with the English version OpenSubtitles 2011 dataset [Tie09], which is an aggregation of subtitles organized by genre and year of release of the respective movie. Such dataset is filtered to generate a conversational corpora suitable for experiments, based on characteristics of the dataset (for example, each punctuation mark is considered as a word token). To solve the problem, the authors propose to use an attention-based bidirectional GRU decoder, with ELU (Exponential Linear Unit) as activation function and a dropout value of 0.5 between the internal layers. For the output of the neural network, a softmax function is applied. The results achieved suggests that the proposed model can generates acceptable responses for common questions. It could be applied in various fields, including health care, finance, e-commerce and manufacturing.

Zhou et al. [ZHZ+18] focus their work on propose the Emotional Chat Machine (ECM), a chatbot which can generate content relevant responses which are, also, emotionally consistent. Such situation arises three main challenges: obtain a high quality emotion-labeled data in a large scale corpus, consider emotions in a natural and coherent way, and embed such emotion information in a neural model. The work tries to solve the problem using the following definition: given a sentence and an emotion category of the response to be generated, the main goal is to generate a response which is coherent with the chosen emotion category. So, the model is trained to learn not only the post/answer pair, but also the emotional category of this pair. The model was implemented using TensorFlow[1]. In order to train, validate and test the model, the authors built an emotion classifier using the NLPCC (Natural Language Processing & Chinese Computing) emotion classification dataset, comprised of six final emotions: Angry, Disgust, Happy, Like, Sad and Other. Such dataset was used in challenging tasks of emotion classification in both NLPCC2013[2] and NLPCC2014[3]. Both an automatic and manual evaluation were conducted. The automatic evaluation was made assuming that the emotion accuracy would be the concordance between the expected emotion category and the predicted emotion category. The manual evaluation was made with three human annotators, where they should answer if the generated response was appropriate for a post, as well if it was natural enough to have been produced by a human being. The results achieved by their work show that ECM can generate appropriate responses, if the emotion category of the response and the emotion of the post both belong to one of the frequent Emotion Interaction Patterns (EIP). EIP is defined as the pair: categories of emotion of the publication and its response. In fact, ECM was able to generate responses appropriate not only in emotion but also in content.

Swanson et al. [SYF+19] also deal with the problem of applying chatbot responses in production level. In their work, they propose a dual encoder architecture, which is optimized to select among as many as 10,000 responses within a couple tens of milliseconds. Such architecture uses a fast recurrent network and multi-headed attention. Instead implementing a LSTM, they use a SRU (Simple Recurrent Units) approach, which uses light recurrence and makes it highly parallelizable and delivers a better performance on training and inference. As for the

[1] https://github.com/tensorflow/tensorflow

[2] http://tcci.ccf.org.cn/conference/2013/

[3] http://tcci.ccf.org.cn/conference/2014/

dataset, the authors chose to work with a proprietary help desk chat dataset, which consists of 15 million utterances from 595.000 conversations. Such data is split into three categories: training (80%), validation (10%) and testing (10%). Their results include both quantitative metrics and human evaluation, where both of them deliver good results. The authors, in fact, state that the proposed model is suitable for use in a production conversational system, being able to achieve over a 4.1x inference speedup when compared with traditional encoders, such as LSTM.

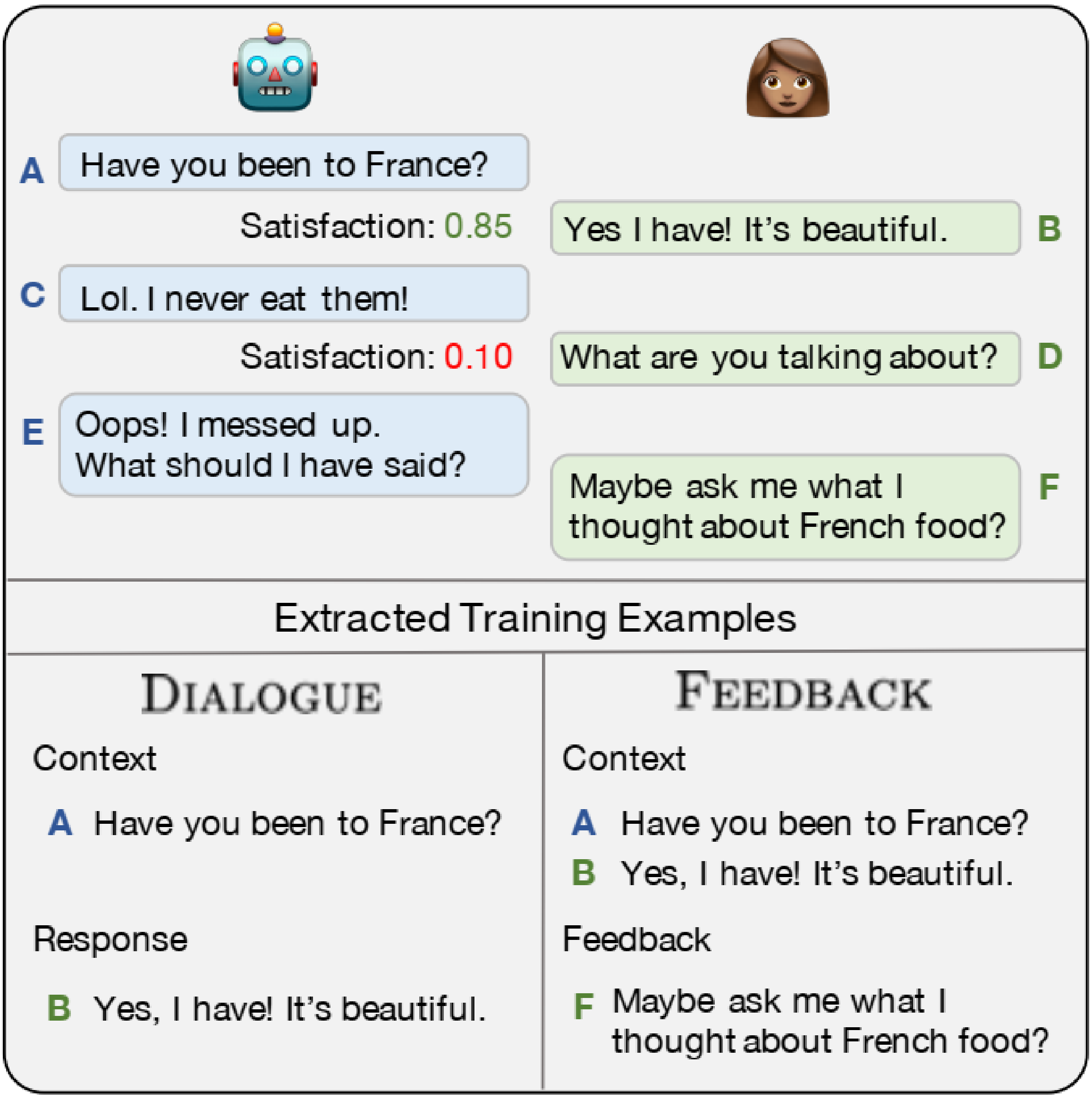
Tao et al. [TWX+19] address the challenge of multi-turn response selection in retrieval-based chatbots. Such challenge can vary both from informal language and typos to context awaring. To solve this, the authors propose a multi-representation fusion network (MRFN) for context-response matching, focusing on a method to fuse the representations in matching and how different types of representations contribute to the performance of matching. The proposed solution of the authors goes as follows: Suppose that we have a data set $D = \{c_i, r_i, y_i\}_{i=1}^{N}$, where $c_i = \{u_{i,1}, u_{i,2}, ..., u_{i,m_i}\}$ represents a conversational context with $u_{i,k}$ the k-th utterance; r_i denotes a response candidate; and $y_i \in \{0, 1\}$ is a label where $y_i = 1$ means that r_i is a proper response for c_i, otherwise, $y_i = 0$. Therefore, the main goal is to learn a matching model $g(.,.)$ from D. To do so, the authors consider three different representations (word representations, contextual representations, and attention-based representations), which both encode semantic information and capture the relationship between them in a given utterance. The tests were run using GRU with 1000 neurons, over two data sets: Ubuntu Dialogue Corpus [LPSP15] and Douban Conversation Corpus [WWX+16]. The results achieved suggests that the proposed model achieves new state-of-the-art performance on both data sets.

Zhang et al. [ZGL+19] propose to solve the problem of consistency on chatbot responses, concerning both context and personas (casual speaker). In their work, they present a self-supervised approach that uses the natural structure of conversational data to learn and leverage both topic and persona features. To do so, they use a discriminatory feature extraction mechanism which is able to seize conversational topics and personas in a self-supervised manner. It allows the model to use massive unlabeled data sets while protecting sensitive user information, since the model does not require the speaker identity. Also, their model is able to generate responses which adhere to high-level features, such as topic and persona. For the tests, two data sets are used: Twitter FireHose, collected from 2012 until 2016; and Maluuba dataset. For the neural network, a Long-Short Term Memory (LSTM) neural network is used, with a hidden layer of size 500. Adam is used as the optimizer, while the learning rate is set to 0.00001. Adam is a well known optimization algorithm largely used for training deep neural networks. Its name is derived from adaptive moment estimation. The results achieved indicate that the proposed model is able to capture meaningful topics and personas features. Also, the incorporation of the learned features helps to significantly improve the quality of generated responses on both data sets, even when comparing with models which explicit persona information.

Li et al. [LGB+16] tackle the problem of consistency in neural response generation, where an inconsistent chatbot can respond similar questions differently. They propose to solve this problem using the concept of Persona: a composite of elements of identity, language behavior and interaction style. They explore two persona models: a single-speaker Speaker Model and a dyadic Speaker-Addressee Model, within a sequence-to-sequence (Seq2Seq) framework [SVL14]. While the Speaker Model integrates a speaker-level vector representation into the target part of the Seq2Seq model, the Speaker-Addressee Model encodes the interaction patterns of two interlocutors by constructing an interaction representation from their individual embedding and incorporating it into the Seq2Seq model. Tests were run over Twitter Persona Dataset, extracted from the Twitter FireHose for a six-month period, and scripts from the American television comedies Friends and The Big Bang Theory, available from Internet Movie Script Database (IMSDb). Twitter database was used for the Speaker Model and the scripts were used for the dyadic Speaker-Addressee Model. The results achieved shows that their model was able to capture personal characteristics such as speaking style and background information. Moreover, in the Speaker Addressee model, the evidence suggests that there is benefit in capturing dyadic interactions.

Hancock et al. [HBMW19] proposes a chatbot model which can learn from its own mistakes. In their work, they present a dialog agent which has the ability to extract new training parameters from the conversations engaged. To do so, they estimate the satisfaction of the user based on the responses it gives. If the conversation seems to be going well, the responses of the user are used as new training data. Otherwise, the agent asks for guidance to help it improve. Figure 2.1 shows an example of their chatbot, which estimates user satisfaction to know when to ask for feedback and improve the dialogue.

Their dialog agent model is built using Transformer architecture [VSP+17]. In general, the chatbot performs a Dialogue task which has a Satisfaction value. If that Satisfaction value is lower than a certain threshold value, a Feedback task is



A Have you been to France?

Satisfaction: 0.85

Yes I have! It's beautiful. B

C Lol. I never eat them!

Satisfaction: 0.10

What are you talking about? D

E Oops! I messed up.
What should I have said?

Maybe ask me what I
thought about French food? F

Extracted Training Examples

DIALOGUE

Context

A Have you been to France?

Response

B Yes, I have! It's beautiful.

FEEDBACK

Context

A Have you been to France?

B Yes, I have! It's beautiful.

Feedback

F Maybe ask me what I
thought about French food?

Figure 2.1: Example of Hancock's et al. chatbot. It estimates user satisfaction to know when to ask for feedback. From the satisfied responses and feedback responses, new training examples are extracted in order to improve the dialogue. Source: [HBMW19].

asked and processed in order to improve future interactions. From time to time, the agent is retrained using all available data. Results achieved by their model indicate that learning from dialogue with a self-feeding chatbot improves performance in a significant way, independently of the amount of supervised examples.

A recent work conducted by Croes et al [CA21] aimed to discover if a human being can build a relationship with a chatbot, as well which set of traits can help in such interactions. In order to conduct their research, they used the chatbot

Mitsuku (https://www.pandorabots.com/mitsuku/). The set of traits measured were: social attraction, self-disclosure, intimacy, interaction quality, empathy, communication competence and feelings of friendship. The results achieved show that all these social processes diminish as time passes by, but intimacy. It suggests that the more people interacted with the chatbot, the worse the evaluation was. It seems to be reinforced by another discovery of the authors: after multiple interactions, people did not consider the chatbot as their friend.

One interesting concept presented by the work is the process of relationship formation. According Levinger [Lev80], such process occurs in stages and follows an ABCDE sequence of relationship development, where A stands for Attraction, B for Build-up, C for Continuation, D for Decline and E for Ending. In another topic, since chatbots deal only with words, it can be interesting to extract different information from such words, like sentiments. Almeida et al. [ACP+18] focus their work in the emotion identification of short texts. In their work, they classify several methods based on the emotion classification problem, like Ensemble Classifier Chain (ECC) and Hierarchy of Multi-label ClassifiERs (HOMER). Considering multi-label learning methods, the results achieved suggest that the best method among the tested was ECC. Following this topic of sentiment analisys in text, Ferreira et al. [FDNP15] tackle the imbalanced database problem. The authors argue that Support Vector Machine (SVM) Classifier is largely used to identify emotions in text, specially due to their good generalization capability, as well as its robustness with data with high dimensions. However, such generalization capability suffers from imbalanced databases, because the method is going to assign most of the texts to the majority class. Therefore, the authors propose a genetic algorithm which aims to balance a database or corpus. The method is applied in two experiments: evaluating the six possible emotions and evaluating only positive/negative value. The results achieved suggest that balancing an imbalanced corpus could be an interesting alternative for emotion identification in texts, specially for multi-emotion identification.

The main difference between the work presented in this section and ours is that Arthur/Bella does not rely only on a trained neural network, but have defined interaction behaviors and is able to learn from previous interactions. More details are given in Section 3.2.

2.2 Human and Agent Memory

Concerning human memory, Loftus et al. [LL19] aims to describe the way that memory is conceived. The authors say that each person is constantly taking information from the environment, proceeding to store, manipulate, and record portions of this information in a succession of memory stages. Therefore, there are two main tasks involving the scientific investigation of memory: (1) identification of the memory stages themselves and (2) the investigation of what types of information each stage can process. Also, their work demonstrates the behavior of human memory, which can be divided into three stages: Sensory Store, Short-term Store and Long-term Store.

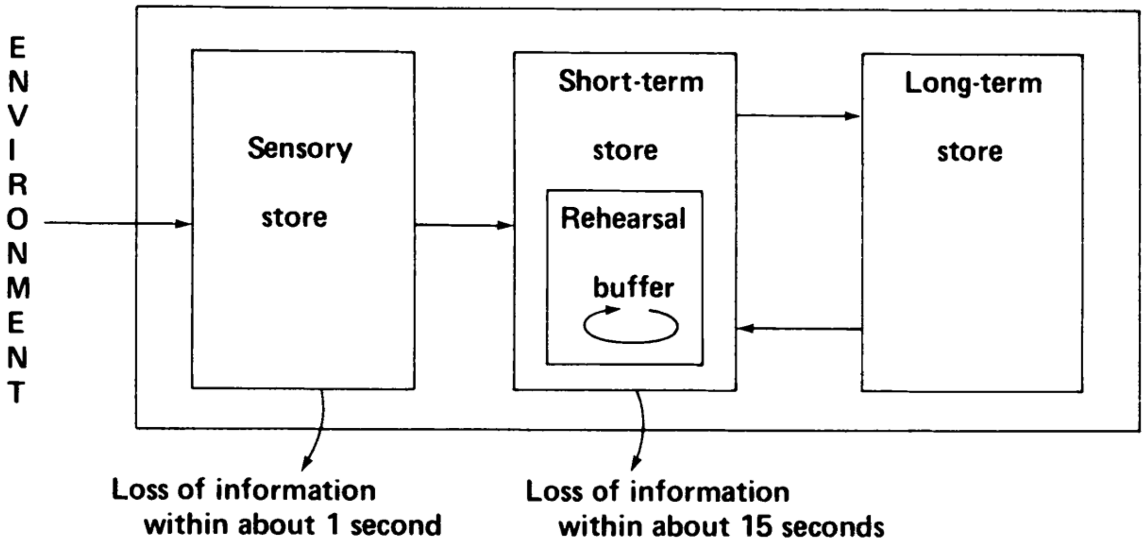

ENVIRONMENT
Sensory store
Short-term store
Rehearsal buffer
Long-term store
Loss of information within about 1 second
Loss of information within about 15 seconds

Figure 2.2: Schema of human memory. All information perceived by people enters by the Sensory Store. The Short-Term Store can maintain information for about 15 seconds, if it is not being rehearsed. The Long-Term Store is virtually unlimited and can keep information for a long time. Source: [LL19]

Figure 2.2 shows a simplified schema of human memory. When people receive external information, it enters their system through one of the sense organs (for example, vision). Such information is then placed into the Sensory Store, which can hold a lot of information for a very short period of time. Indeed, such information is usually lost within a mere second. Part of this information can be transferred to the Short-term Store, so it will not be lost so fast. Although, short-term memory has far less capacity than sensory, so just important information is transferred to it. It can maintain information for a time being of around 15 seconds. Also, Short-term

Store counts with a special structure known as Rehearsal Buffer, which can maintain information indefinitely using the rehearsal process, it means, repetition [LL19]. For example, one can remember a phone number just repeating it to itself. Finally, the Long-term store consists in the virtually unlimited capacity storage of information that each human being has, which is more or less permanently available to us. It is usually assumed that any information present in the Short-term Store can be copied (or transferred) to the Long-term. Basically, the longer some particular information stays in Short-Term Store, the more of such information can be transferred into Long-Term store. For example, let us assume that you have a phone number (e.g. 3313-7766) to remember. If you keep repeating “3313-7766” to yourself, you are maintaining it in Short-term Store, more specifically, into the Rehearsal Buffer. Additionally, during this time, information about the number may be transferred to Long-term Store. This process is a bit slower, thus, depending on the time the information was keep in Short-term memory, just a part of the information may be transferred to the Long-term memory. For example, instead to transfer all phone number, it is possible that just the first part (3313) or a random part (3?13-7??6) of the information enters into the Long Term memory, where the "?" represents the part of the information not transferred.

Assuming the information is stored into some level of memory, how can one retrieve such information if required? Using the same example as before, let us suppose one desires to call to the number 3313-7766. Firstly, a search is conducted in the Short-term Store: if the information is indeed there, good. Otherwise, another search is conducted in the Long-Term store. If the information is indeed found, it is transferred to the Short-term Store and used as needed. It is possible that the information is not complete, because there was not enough time to transfer it as a whole. So, just part of the information is remembered, like just the first part of a phone number [LL19].

One of the most accepted models concerning human memory cited on literature is known as Autobiographical Memory. As defined by Bluck et al. [BL98], autobiographical memory is "a system that encodes, stores and guides retrieval of all episodic information related to our person experiences". Also, according to Conway et al. [CPP00], autobiographical memory can be grouped in three levels: lifetime periods, general events and event-specific knowledge. So, such memories can be directly accessed if the cues are specific and relevant to the person. Otherwise, if the cues are too general, a generative retrieval process must be used to produce more specific cues for the retrieval of relevant memories. The authors say that the difference between them is that "the search process is modulated by control processes in generative retrieval but not, or not so extensively, in direct retrieve" [CPP00].

Following this definition of autobiographical memory, Wang et al. [WTM16] build a model to mimic such behavior. Their model, known as Autobiographical Memory-Adaptive Resonance Theory (AM-ART), is a three-layer neural network that encode lifetime periods, general events and event-specific knowledge, respectively, being, therefore, consistent with the model presented by Conway et al. [CPP00]. Also, it encodes the 5W1H schema, which represents *when* an event occurred, *where* it happened, *who* was involved, *what* happened, *which* pictorial memory was associated with the event and *how* was the person feeling during the event. The results achieved by their work show that AM-ART was able to perform better than the keyword-based query method, since the last can not deal with noisy cues in many existing photos or memory repositories. Also, the model was able to both encode and retrieve real-life autobiographical memory, creating pictorial snap-shots of the life experience of a person, together with the associated context.

It is believed that the Autobiographical Memory is a part that composes the Long-term Memory of each human being. In a similar way, the Working Memory is believed to exist inside the Short-term Memory. As defined by Baddeley [Bad92], such definition has indeed evolved from the concept of an unitary short-term memory system, referring to "a brain system that provides temporary storage and manipulation of the information necessary for such complex cognitive tasks as language comprehension, learning, and reasoning". According to the author, the Working Memory is comprised of three parts: the visuospatial sketch pad, the phonological loop and the central executive. This last one cited would represent the coordinator of such memory, so, one of its roles would be coordinate the information from the slave systems. Concerning these slave systems, the visuospatial sketch pad deals with the visual and spatial information, being related to the processes of visual perception and action. In its turn, the phonological loop works with the verbal communication, representing an evolution of the basic speech perception and production systems.

There are also other work which tackle the memory problem. Edirisinghe et al. [EMJ18] model an autobiographical memory for a robot that can store knowledge about users during friendly interactions, recalling them during future interactions. Autobiographical Memory was developed in a three-layer architecture. Once the robot interacts with a new person, it creates a user profile for that person. The results achieved show the potential of such memory mechanism for robots, which can improve the long-term interactions between humans and robots. Kasap et al. [KMT12] focus on the problem that people often lose interest on virtual agents or robots after the novelty effect disappears. In order to build a long-term interaction model which can keep the interest of users, they developed a robotic tutor called Eva endowed with many aspects, like emotion and memory. In fact, the results achieved by their work provide the first evidence that the use of a memory system, in a long-term interaction, can effectively help in keeping the attention of the users as time passes by.

Martinez et al. [MK20] considers the problem of interacting with multiple users at the same time. In order to do so, they argue that conversational agents should be able to distinguish between two classes of interactions: those that address a single person and those open to any group member. To solve this, the authors present a module which keeps a concurrent record of conversations, where each one of them can be explicitly marked as a group or individual interaction. Moreover, they include a memory module in their dialogue manager, which allows the virtual agent to reason about past interactions. Such module is stored in a database and used to keep track of what was already spoke about. For example, if the user already said to the agent that he likes hockey, the agent can ask "Do you still like hockey?" or drive the conversation toward this topic (e.g. "Let's talk more about hockey").

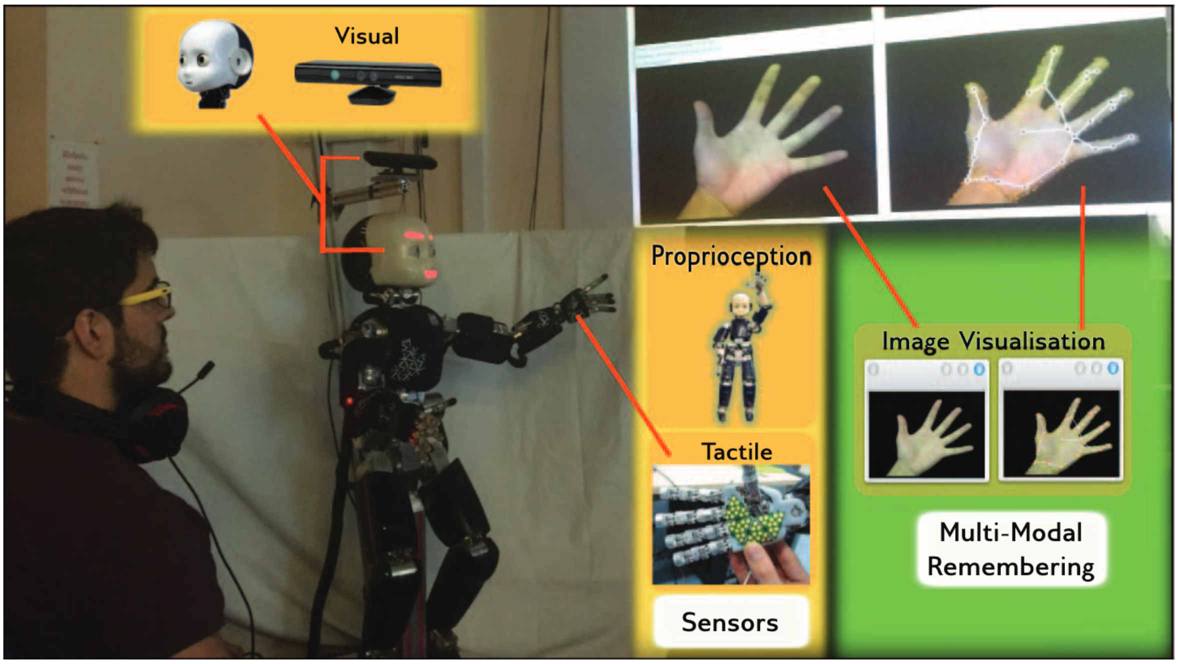


Figure 2.3: Setup of the ICub robot. All the data acquired can be stored in the autobiographical memory. The sensors of the robot are two eye-cameras, the state of the joints and tactile information. Source: [PFD15]

Petit et al. [PFD15] implement Autobiographical Memory in a robot, named ICub. Figure 2.3 presents the setup of the robot, alongside its sensors. All the information collected by the robot's sensors can be stored into its memory and used later. Memory data is stored in Postgres database and episodes are defined within semantic words. For example, "Can you remember the last time HyungJin showed you motor babbling?" is recognised using the grammar rule "Can you remember the <temporal cue> time <agent cue> showed you <action cue>?". This way, it is possible to know that the question is about an action "motor babbling" done by an agent called "HyungJin" for the "last" time. With this, a SQL query can easily search for the information inside the memory.

The main difference between the work presented in this Section and ours is that we do not train a neural network. We modeled our memory in a procedural way, similarly to [KRK13]. Also, our memory retrieval can be guided by an emotional state and, even, change the mood of the agent. More details are presented in Section 3.8.

2.3 Empathy, Appearance and Expressiveness: ECAs

While developing an interactive agent, there is a concern about how it presents itself. It is important that the interaction occurs in a way that the subject in contact with such agent does not feel uncomfortable and/or awkward. As shown in Dill et al. [DFH+12], when trying to present a character that pretends to be human-like, there is a certain eerie feeling (uncanny valley) when it looks human, but not as close as expected from a real one, especially when animation is poor made. It is pointed out that, considering a character's face, the elements that can cause more strangeness are eyes and mouth. The authors, therefore, show that people tend to have a better reception of cartoon-like or a very close to a expected human being character.

Real human eyes tend to not stay static for too much time, having some involuntary or voluntary movement at some point. Such phenomena is known as saccade, being defined by Leigh and Zee [LZ15] as rapid movements of both eyes from one gaze position to another. Lee et al. [LBB02] propose an algorithm for a virtual character eyes animation, more specifically, to produce a movement called saccade, a behavior in which rapid and discontinuous eyes movement from a position to another occurs. The authors obtained a statistical model through the analysis of eye-tracking images from a subject during a conversation with a software. In combination with literature data, they achieved a model capable of generate movements considering various aspects, like magnitude and duration, of a human saccade. As result, the model obtained a natural and friendly perception from subject into author's survey.

Concerning empathy behavior, Pereira et al. [PLM+10] presented a robot aimed to act as a social companion, able to express different kinds of empathetic behaviors, both with facial expressions and utterances. Its main task was to comment on the movements of two chess players, being empathetic towards one of them and neutral with the other. The results of the study suggest that the players with whom the robot was being empathetic perceived it more as a friend than the other players.

The work of Yalcin [Yal20] aims to model empathetic behavior on Embodied Conversational Agents (ECAs). As it is commented, empathy is "the ability to understand and react towards the emotions of others", being an important feature for smooth interpersonal interactions. Yet, the complexity of an empathetic model is related with the wide range of behaviors it arises, like mirroring, affective matching, empathetic concern, altruistic helping and perspective taking [CG11, dWP17]. The ECA built by Yalcin has three stages: listening, where the agent captures input from the person it is talking to; thinking, where the agent process the information; and speaking, where the agent gives a proper response, both with words and gestural behavior. Also, concerning the empathetic behavior, it should be able to allow the agent to give responses to the user in a verbal and non-verbal way. Since an empathetic behavior relies on the emotion of the subject, an emotion recognition module is used alongside the video input for the agent. Concerning this emotion, the audio of the person speaking is also used to help determine the overall emotion.

Biancardi et al. [BWM+19] proposed an ECA model which can measure and manage the impressions of the person it is talking with. In other words, their work is focused on the impressions of the user itself, concerning the ECA. Such impressions are measured with warmth and competence dimensions, which, according the authors, are considered as the most fundamental dimensions in social cognition. Their main goal is to adapt the behaviours presented by the ECA based on the impressions and reactions of the user to user, such as their facial expressions.

In general, their model is comprised by two main modules:

- User's Impressions Detection: to capture the impression of the person. In order to achieve this, a few tools are user, like the EyesWeb framework (which extracts the Action Units of the face of a person) to get the contraction of different muscles of the face, and the Microsoft Speech Platform to get the person's voice.

- Agent's Impressions Manager: which decides the behavior of the ECA based on the impression found for the user.

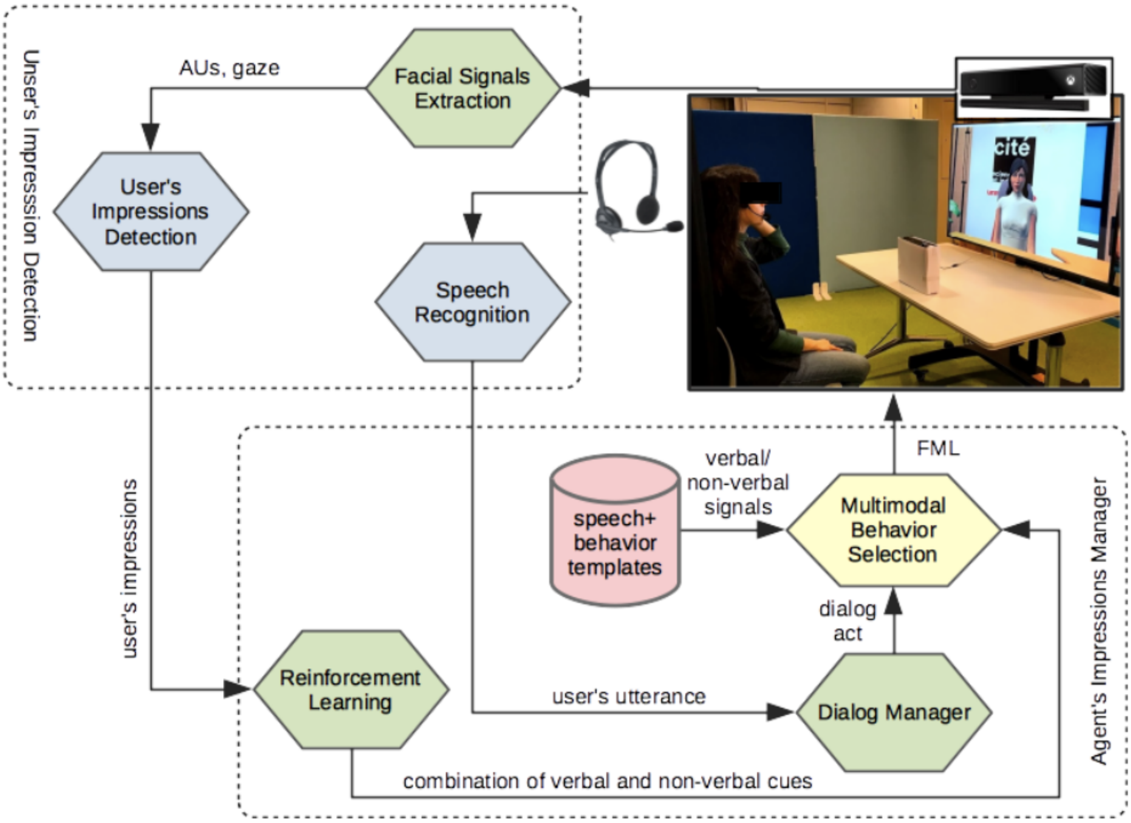

Unser's Impresssion Detection
AUs, gaze
Facial Signals Extraction
User's Impressions Detection
Speech Recognition
cité
user's impressions
verbal/ non-verbal signals
FML
speech+ behavior templates
Multimodal Behavior Selection
dialog act
Reinforcement Learning
user's utterance
Dialog Manager
combination of verbal and non-verbal cues
Agent's Impressions Manager

Figure 2.4: System architecture of Biancardi et al. ECA. It is mainly comprised by two modules: the User's Impressions Detection and the Agent's Impressions Manager. Source: [BWM+19]

An overview of their model can be seen in Figure 2.4. The results achieved by their work show that the participants rated higher the ECA when it was able to adapt its behavior due to user's impressions, which seems to confirm that the authors were successful on their endeavour. On the other hand, many participants were a bit disappointed with some features of the ECA, like its appearance, voice and animations, which were described as "creepy" and "disturbing". One of their future works is to improve such features.

Sajjadi et al. [SHCK19] conducted an experiment which aimed to investigate the effect of a person interacting with a personality-driven ECA. In their work, they hypothesize that an ECA which has its non-verbal behavior governed by a personality-driven behavioral model would both increase the level of social presence of the person and provide a better game experience. They also discuss that, in order to increase such level of social presence, immersion and involvement are very important. To test they hypothesis, the authors built a prototype of an ECA with a personality-driven model. This model is responsible of the expression of its emotional state, providing a variety of non-verbal cues. At each interaction, the ECA called Linda can re-calculate its emotional state based on its actual state and the interaction itself. If the users provides a given stimuli, Linda can change its emotional state towards a given state. On the other hand, if the user does not provide any stimuli while talking, Linda can become bored. Figure 2.5 shows two examples of Linda. On the left, it is showing an angry emotion, while it shows a sad emotion on the right. An experiment was conducted with 41 participants in order to evaluate the initial hypothesis. The results achieved seem to validate them. As the authors comment, it was observed that an emotionally-personified ECA with an extrovert-based personality generates a higher sense of behavioral involvement in human users, when compared to a less emotionally-personified agent with no non-verbal behavior. Therefore, they were able to conclude that, as observed in the experiment, higher levels of incorporated personality on the ECA induce a higher level of involvement by the users.

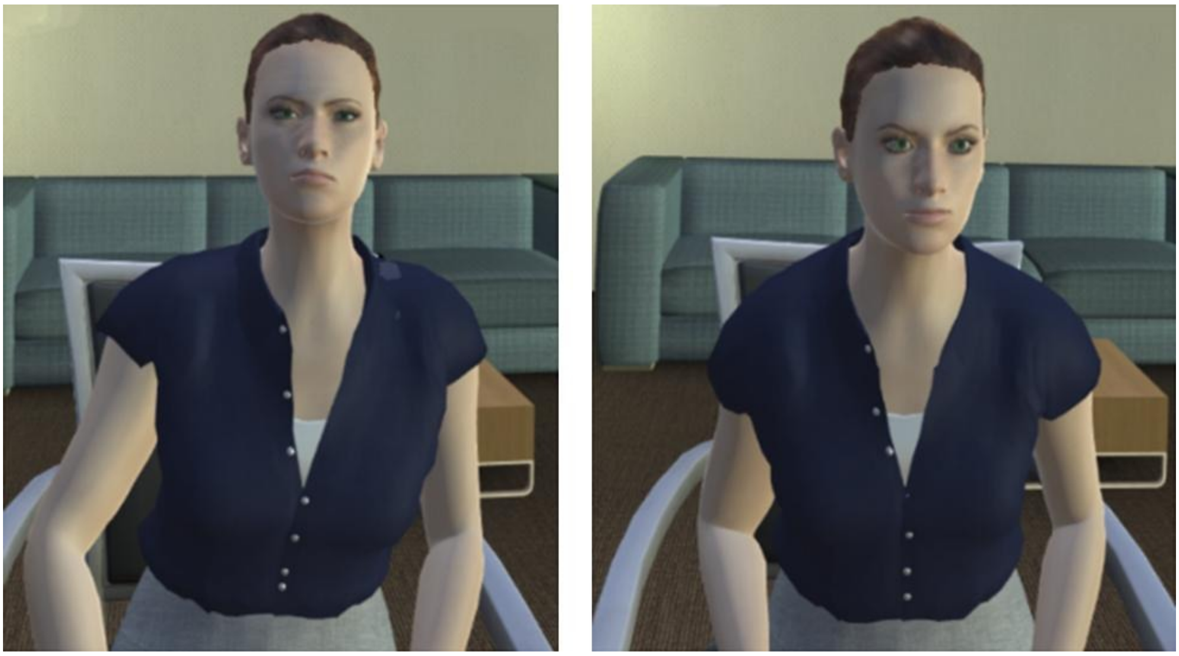


Figure 2.5: Personality-driven model Linda. On the left, it is showing an angry emotion. On the right, a sad emotion. Source: [SHCK19]

In an interaction between two humans, one can interrupt the other for some given reason. For example, one of them may be bored about the conversation, or tired, or even remembered something about the subject. Cafaro et al. [CRP19] focused their work on modeling such interruptions in an ECA. Based on nonverbal reactions presented by the user and driven by an evolutionary algorithm, the authors propose a technique which allows to build a ECA able to manage interruptions of users.

As one may already know, a genetic/evolutionary algorithm is a heuristic search method used to find optimized solutions to search problems, based on the well accepted theory of natural selection and evolutionary biology [And05]. In such algorithms, a genome represents the sequence of genes that is evolving at each iteration. In the work of the authors, each genome is a nonverbal reaction to a conversational interruption, while the genes which compose it are the specific nonverbal behaviors, as follows: Head tilt, Nod/Toss, Lids close, Eyebrows, Eye squeeze, Smile, Shoulders up and Gesture phase freeze. An user study was developed to evaluate the behavior of the virtual agent. The task of the participants was, essentially, to find out if the ECA reacted to an interruption with a friendly of hostile attitude. The average satisfaction level of the participants was above 4 out of 5 points, which shows that their work is promising, especially on a field rarely studied in ECAs.

Dermouche et al. [DP19] proposed to endow an ECA with the ability to adapt its own behavior according the behavior of the user. For example, when two people interact, one can nod to show concordance, gaze to an appointed object or smile when the other person smiles. The main novelty of their work resides on the ability of the virtual agent to present nonverbal behaviors as a function of both person's and its own behaviors. Also, the generation of such behaviors occurs in real-time, which enhances the quality of the interaction. In order to do so, the authors modify a LSTM (Long-Short Term Memory) neural network adding a "user-in-the-loop" approach, to constantly predict the behavior of the agent in response to the behavior of the user. For this prediction, the model uses as input both the agent's and the person's past behavior, portraited as nonverbal behaviors like smile, head movements and gaze. That way, the agent is not only able to communicate its intention, but also its engagement in the interaction. In order to evaluate the model, an interactive experiment was designed, where the virtual agent plays the role of a virtual guide which describes an exhibition about video games in a science museum. It is assumed that both user's satisfaction and engagement are going to increase due to the adaptive behavior of the virtual agent. The results achieved show that users were, indeed, more satisfied interacting with the ECA when it was able to adapt its behavior. However, these results were significantly positive only when the virtual agent adapted its smile to user's behavior, which can imply that a bias, from the user part, could have prevented the ECA from having more better results for the other signals.

Concerning the goal of an ECA, Ochs et al. [ODMP+17] proposed a virtual agent who acts as a patient and is used to train physicians in breaking bad news. Authors discuss about the way that medics deliver bad news to patients, which has a significant importance in the therapeutic and recovery process. Many experienced clinicians and medical trainees face hardship with such task, which justifies that medics should be trained and develop skills in communication. Real-life training simulations are performed with actors, who assume the patient role, but such training solution is costly and time consuming, which justifies the development of a virtual agent.

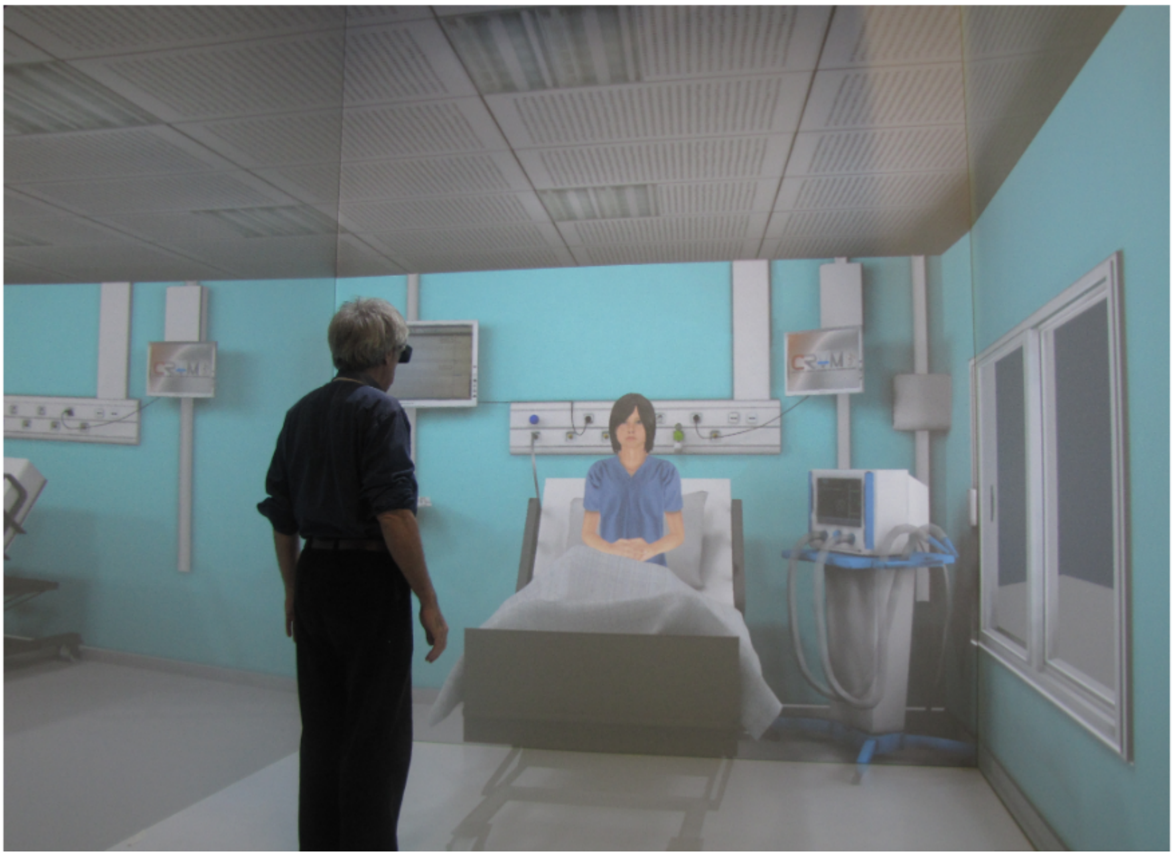


Figure 2.6: User interacting with the virtual patient in the virtual reality room. Source: [ODMP+17]

Their solution was developed for PC and virtual reality (both headset and room). Figure 2.6 shows a user interacting with the ECA patient in the virtual reality room (CAVE). The behavior of the virtual patient was modeled analyzing many videos of real-life training simulations, where the actor who plays the role of the patient follows a pre-determined scenario. The dialog of the ECA was modeled using OpenDial [LK16], which is a java-base toolkit used in the development of dialog systems. In the stage of the work, experiments were still being run. The authors comment that they had plan to analyze the experience of the users, in a subjective way, through questionnaires, but also to compute other objective measures through verbal/non-verbal behaviors, such as amount of gestures and the length of sentences used.

Finally, when it comes up to assembling an ECA from scratch, a considerable amount of work is put into assembling multiple complex components into a virtual agent. To address this issue, Beinema et al. [BDR+21] proposed the Agents United Platform, where developers have access to a set of integrated components which work together in the building of Multi-Agent Conversational Systems. Such components are compounded of a sensor framework (in which information provided by the user is collected), a memory component (which stores information), a topic selection engine (which chooses the topic and ensures that the conversation is relevant), an interaction manager (which controls the flow of the interaction), dialogue execution engines (which builds the dialogues) and behavior realisers (which builds non-verbal behaviors).

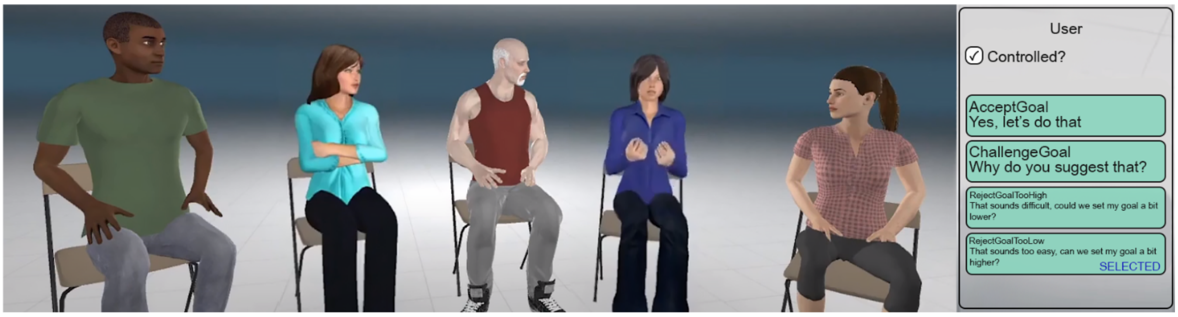


Figure 2.7: A group of interacting agents. The user can select an answer in the right to participate in the conversation. Source: [ODMP+17]

Figure 2.7 presents a scenario, built in Unity, of a group of agents. The user can interact with them by selecting an answer from the menu in the right of the image. The authors comment that one of the main contributions of their work is the integration of expertise, models, and insight from many state-of-the-art components for both argumentation and social conversation. Moreover, it makes it easier for researchers to build their own tailored virtual agents. Finally, the authors also comment on possibilities for future work. One example would be the implementation of a group-aware semi-autonomous nonverbal agent behaviour, where the virtual agents would be able to show group behaviors like gazing and turn-taking during interactions.

In this work, we chose to model our virtual agent in a cartoon manner to avoid the strange feeling discussed by Dill et al. [DFH+12]. Also, we included saccade eyes movement, following the algorithm proposed by Lee et al. [LBB02]. More details are presented in Section 3.9.

2.4 Chapter Considerations

This Chapter presented many work related with what is being proposed in this thesis. A literature revision was made in order to search for the most important and modern work concerning ECAs, memory modeling and virtual agent features.

The main contributions of this work when compared with others presented on this Chapter is related with Embodied Conversational Agents (ECAs). In the last years, many research was made to improve the quality of the communication abilities of such ECAs, both verbal and non-verbal [Yal20, BWM+19, SHCK19]. Also, a fair amount of effort is being directed on ECAs which can help people to have a healthier life [KtSM+19, SCCG20, DBODAH19], for clinical interviews [PDA+20, MMMR+19] and the training of some skill [CW19, AHS19]. Although, very few work tackle the problem of empathy and its relationship with the memory. Such feature can be very useful to create more immersive and customized interactions.

The next chapter presents the proposed model, describing its many features and how they are assembled to work together in order to achieve the goals of this work.

3. PROPOSED MODEL

This chapter aims to present the work proposed in this thesis. The proposed model is divided into several modules, as it can be seen in Section 3.1. Section 3.2 explains how the chat with the virtual agent works, as well the Voice Detector module. Section 3.4 shows how the virtual agent is able to recognizes a person he/she is seeing, while Section 3.5 describes how our ECA recognizes the user's emotion. Section 3.8 deeply discuss the operation of the virtual agent memory, showing how it works, how it is organized and how it can be used in the conversation. Finally, Section 3.9 presents the facial reactions that the virtual agent can assume, depending on its actual feelings.

3.1 Overview

This section aims to present a brief overview of the proposed model. The overview of our model is illustrated in Figure 3.1. This work was mostly developed using Unity3D [Tec20] in the C# language. Some modules were developed using Python, e.g. the Face Recognition module. Finally, there is a minor part developed in Prolog, concerning the Beliefs module.

As it can be seen in Figure 3.1, our model is divided into several modules. In blue we highlighted the two main Controllers, which are responsible of controlling the interplay between many modules. The Behavior Control is responsible to define the appropriate behavior of the virtual agent, according all data available (i.e., person who is talking to it, agent memory, emotion detected, and so on). In other words, it allows the virtual agent to react to a given input provided by the user. Therefore, it is connected with all other modules and controllers. The Memory Control is responsible for managing the memory of the virtual agent and is linked with all the memory features (i.e., Memory Learning, Memory Retrieval, Memory Consolidation, General Events and ESK). It is going to be explained in Section 3.8. We use a common method to define artificial memories, which is to use Short and Long-term memories, where information is transferred from one to the other (STM and LTM, respectively) [LL19]. So, during the interaction between Arthur and the user, the information is stored in STM. Then, a phase called Memory Consolidation Module deals with the consolidation of the Long-Term Memory (LTM) of the agent. There-

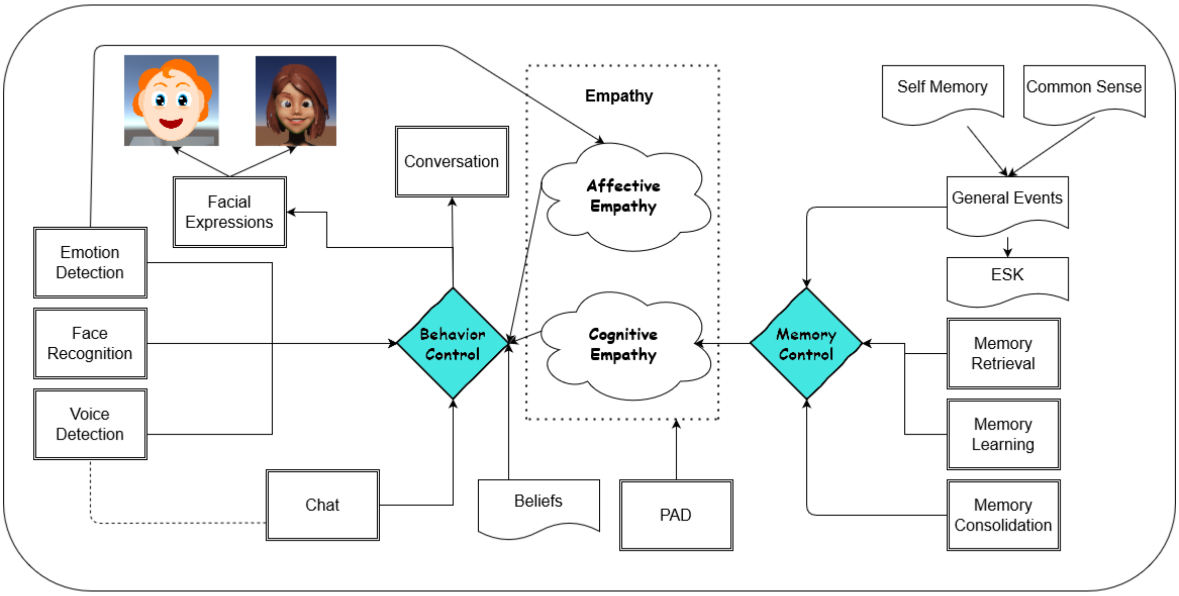


Figure 3.1: Overview of the proposed model. In blue, we highlight the two main Controllers. The Behavior Control is responsible to define the appropriate behavior of the virtual agent, while the Memory Control is responsible of store and retrieve memories.

fore, saved memories can be forgotten by Arthur if their importance is too low or if they are not used for too long. Such process is detailed in Section 3.8.5.

The Chat and Voice Detection modules are responsible for the interaction between the user and the agent, in the scope of verbal and textual behavior, where the Chat Module allows for the user to type some sentence and the Voice Detector Module is able to transform the voice of the user into words, which are used as input to the agent. The Face Recognition module allows Arthur/Bella to recognize the person he/she is talking to, while the Emotion Detection Module allows him/her to identify the person's perceived emotion. The Facial Expressions module is responsible to animate the facial expressions of Arthur or Bella, such as emotions and eyes movement, while the Conversation module delivers the verbal response of the agent.

The Self Memory and the Common Sense modules give our virtual agent some previous knowledge about many things before it can start to interact with people. While the Self Memory module gives the agent knowledge about himself/herself, the Common Sense module gives the agent knowledge about several things about the world and the environment. The Beliefs module allows Arthur or Bella to reason about the knowledge he/she has. Finally, the Empathy module endows Arthur and

Bella to demonstrate an empathetic behavior towards the person he/she is talking with.

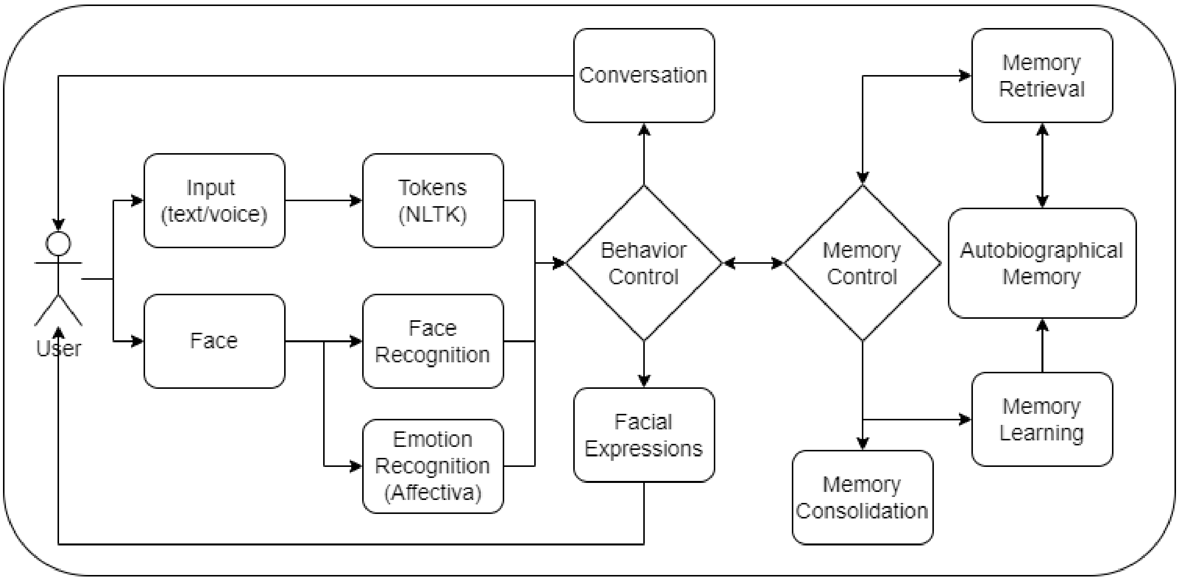


Figure 3.2: Flowchart of the proposed model. The information of the user is sent to the Behavior Control, both verbal and non-verbal. The verbal information is translated into Tokens, while the face of the user is used for both Face Recognition and Emotion Recognition. The Behavior Control receives this input from the user and uses it to access the Autobiographical Memory of Arthur/Bella, communicating with the Memory Control. In its turn, the Memory Control is able to both store new information into the Autobiographical Memory (Memory Learning) and retrieve memories (Memory Retrieval). Additionally, it can consolidate the memory of the virtual agent (Memory Consolidation). Finally, the Behavior Control decides how Arthur/Bella should answer/behave and outputs verbal (Conversation) and non-verbal (Facial Expressions) behaviors.

Additionally, we also want to present how the data flows through our model. Figure 3.2 shows the flowchart of the proposed model. Firstly, the information of the user is sent to the Behavior Control, both verbal (i.e., text or voice) and non-verbal (i.e., face). The verbal information (i.e., what the user spoke/wrote) is translated into Tokens, while the face of the user is used for both Face Recognition and Emotion Recognition. The Behavior Control receives this input from the user and uses it to access the Autobiographical Memory of Arthur/Bella, communicating with the Memory Control. In its turn, the Memory Control is able to both store new information into the Autobiographical Memory (Memory Learning) and retrieve memories (Memory Retrieval). Additionally, it can consolidate the memory of the virtual agent (Memory Consolidation), from the Short-Term Memory to the Long-Term Memory. Finally, the Behavior Control decides how Arthur/Bella should answer/behave and outputs verbal (Conversation) and non-verbal (Facial Expressions) behaviors back to the user. The user receives the output and provides a new input, starting the cycle all over again.

Next Sections are going to present more details of the method and of each one of the modules.

3.2 Chat and Voice detector

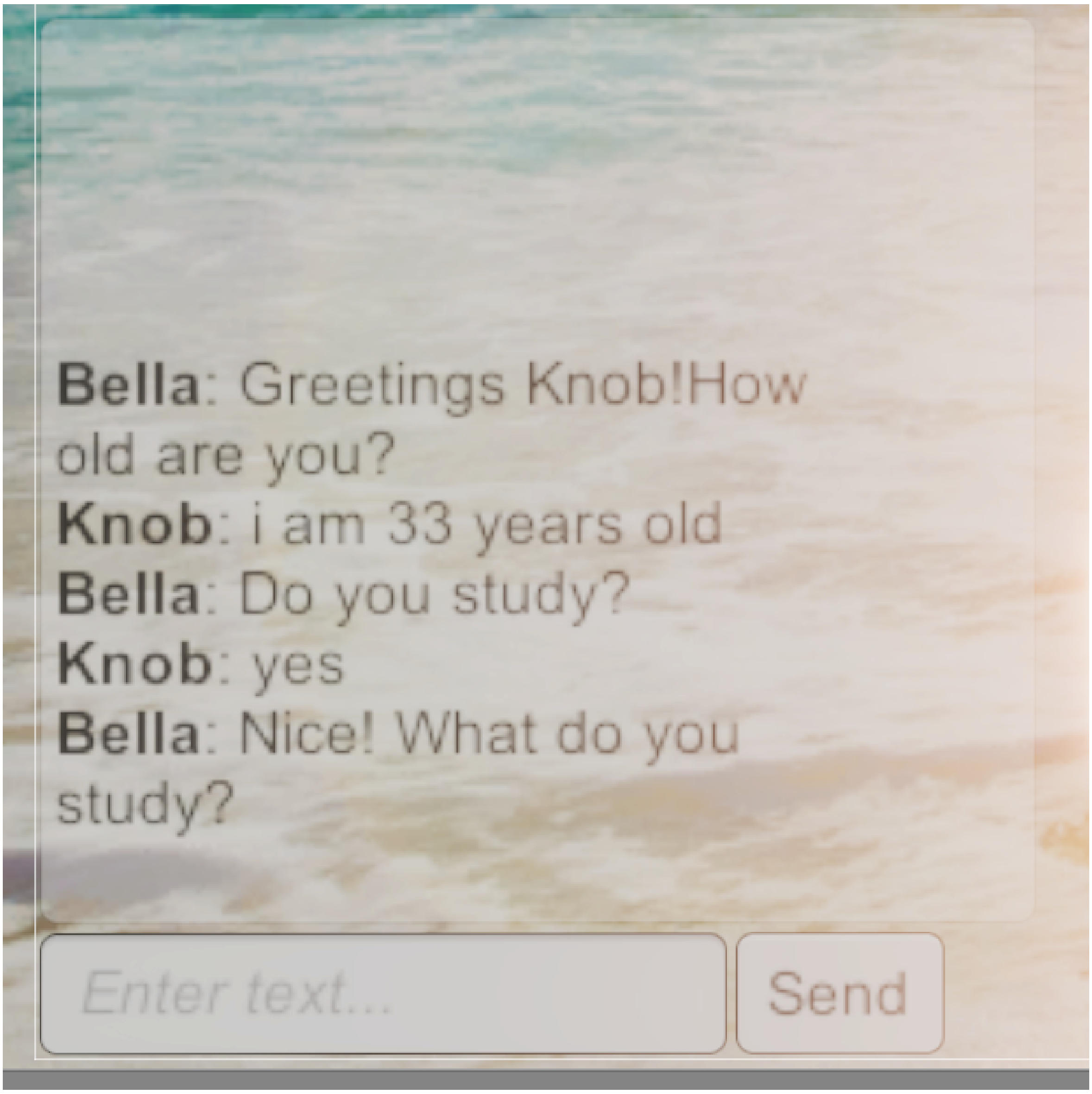
The chat module allows for the user to talk with the virtual agent, writing words and sentences. To do so, a simple text input is used. The user can, also, use its own voice to interact with the virtual agent. The voice detection module is able to transform the voice input into words, which can be understood by the agent. To do so, the DictationRecognizer[1] class was used. It is available for C# and, thus, for Unity3D, allowing to access the system scripts which deal with voice translation to text. Since it deals with Windows scripts, it has a clear downside: it can not be used with other operational systems. Although, we keep the Dictator script modularized in our method; therefore, any other method able to translate voice into text can be used. Also, using voice or not can be easily activated/deactivated in the interface.

Figure 3.3 shows an example of a chat between the user and Bella. The "Enter text" field is used for the user to write things to the virtual agent, while the "Send" button sends the message. If the Voice Detector module is active, it is not necessary to write the messages, only to speak. The text area shown above the field and the button keeps track of all the conversation, both from the agent and the user. Additionally, a chat log is stored in the end of the interaction, with everything that was spoken, so it is possible to consult it even after the interaction is over.

3.3 Conversation

The Conversation module is responsible for what Arthur and Bella speak to the user. It depends on many factors, like what the user said to the agent, what the agent is able to remember, and so on. At the beginning of the conversation, we modeled some initial questions that our agent can ask the person it is talking to.

[1] https://docs.unity3d.com/ScriptReference/Windows.Speech.DictationRecognizer.html



Bella: Greetings Knob!How old are you?

Knob: i am 33 years old

Bella: Do you study?

Knob: yes

Bella: Nice! What do you study?

Figure 3.3: Example of a chat between the user and Bella. The "Enter text" field is used for the user to write things to the virtual agent. The "Send" button sends the message. The text area above both of them keeps track of all the conversation.

These questions serve both as an ice-breaker and to know some basic information about that person. To the extent of this work, the modeled questions are:

- How old are you?
- Do you work?
- Do you study?
- Do you have children?

Each question can, also, have another question(s) linked with them. If the person says that he/she works, Arthur asks: "I see. What is your job?". In a similar way, if the person says that he/she studies, Arthur asks: "Nice! What do you study?". Finally, if the person says that he/she has children, Arthur makes two more questions: "Good! How many kids do you have?" and "What are their names?". All the information provided by the user when answering Arthur questions are stored in a database and constitute Arthur's memory. Since such memory is stored at a database, it is fairly easy to change/increase/decrease it, as the need arises. Also, it is possible to exchange it for another available information, just altering this information at the database. All the information provided by the user is stored in the memory of the agent, as further explained in Section 3.8. If, at any time, the user asks something about a given subject which the virtual agent already knows about, he/she is able to answer. For example, supposes that John tells to the agent that he is 42 years old. Then, in another conversation, someone asks to the virtual agent "how old is John?". In this case, the agent is able to answer that John is 42 years old. Finally, if the agent is not sure about what to talk, the sentence written/talked by the user is sent to a chatbot API[2] and the answer is received and shown to the user. Since we are using an API for the chatbot responses, it is fairly easy to use another chatbot model, if needed.

In person to person interactions, it is fairly common that, at some point, the conversation's topic comes to an end and an uncomfortably silence follows. In such situations, someone usually says or ask something to the other person, in order to break the awkward silence and keep the conversation flowing (who never asked the famous question about the weather?). Thus, besides the "icebreaker" questions introduced above, we also believe that it is interesting that Arthur and Bella are able to break the silence and propose some topics for conversation, if the need arises (but, of course, do not need to ask necessarily about the weather!). In order to solve this, we propose to model some small talks.

As defined by the Cambridge Dictionary, small talks have the basic definition that it is a "conversation about things that are not important, often between people who do not know each other well"[3]. The main advantage of bringing this concept to Arthur and Bella is that it allows to build a more approximate relation between virtual agent and human, especially when it is for a Long-term Interaction [Mor05]. The main idea is to allow both Arthur and Bella to have a conversation about topics

[2] https://rapidapi.com

[3] https://dictionary.cambridge.org/dictionary/english/small-talk?q=small+talk

which people usually talk about when they do not have a specific task or problem to solve. For example, asking another person what is his/her favorite band may not have much relevance to reach a specific goal, but is important to establish a wealthy and friendly relation between two human beings. While interacting, we make use of small talks quite often, sometimes even to create closer relationships with other people.

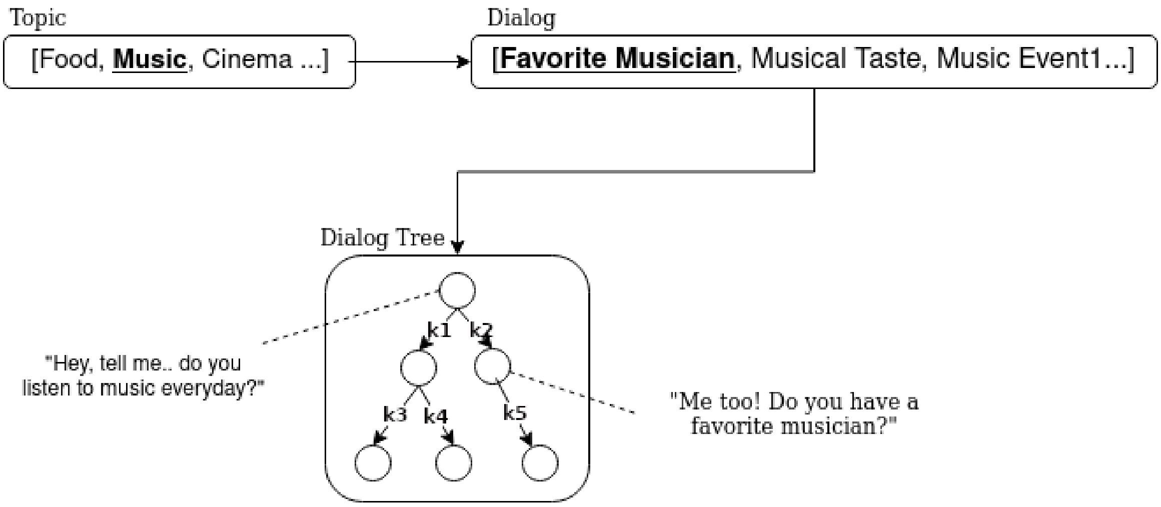

Topic
[Food, **Music**, Cinema ...]
Dialog
[**Favorite Musician**, Musical Taste, Music Event1...]
Dialog Tree
k1
k2
k3
k4
k5
"Hey, tell me.. do you listen to music everyday?"
"Me too! Do you have a favorite musician?"

Figure 3.4: Small talks structure. Topics define the broad subject, Dialogs refer to a given Topic and each Dialog has a Dialog Tree with different utterances. In the example, the selected Topic was "Music" and the selected Dialog was "Favorite Musician". Inside this Dialog, the first thing that Arthur or Bella says is the utterance present in the root of the Dialog Tree (i.e., "Hey, tell me... Do you listen to music everyday?"). Each "k" (k1, k2, k3, etc.) represents a set of keywords. If the answer of the user has many words which are present in k1, the path is traveled downleft. Otherwise, if the answer has many words which are present in k2, it is traveled downright (i.e., "Me too! Do you have a favorite musician?").

In order to build our small talks, we must first define its structure. There are many research about dialogue systems [MB03, Csá, JBD+18] and its advantages to make the interaction as more natural as possible. Since the main goal of this work is not to create a new dialog system, we chose to create a simple conversational structure based on a Decision Tree [CGRFK20]. We chose to work with a Decision Tree because it is a simple and robust way to model a dialog flow. Thus, our small talks are divided into three parts: Topics, Dialogues and Dialog Tree. Figure 3.4 presents an overview of it. Topics define the broad subject that Arthur or Bella can talk about. Different Topics could be Music, Food, Sports and so on. Each Topic has a set of Dialogues which can be chosen to be used. For example, in the Topic

"Music", a set of Dialogues which could be defined is "Favorite Musician", "Musical Taste" and so on.

For each of the Dialogues defined inside a certain Topic, a Dialog Tree is built. This Dialog Tree is composed of branches and nodes which define what Arthur or Bella can speak to the user. Each node represents an utterance that the agent can speak, while each branch represents the possibility to travel between one node and another. In Figure 3.4, it is possible to see that each branch has a "k" associated (e.g. k1, k2, k3, and so on). Each "k" represents a set of keywords. So, if the answer of the user has many words which are present in k1, the path is traveled downleft. Otherwise, if the answer has many words which are present in k2, it is traveled downright (i.e., "Me too! Do you have a favorite musician?"). Figure 3.5 presents an example of this process. In the example, the Topic chosen is "Food" and the Dialogue is "Favorite Food". Inside this Dialog, there is a Dialog Tree with three nodes (which means, three possible utterances for the agent to speak). Arthur or Bella begins speaking the utterance associated with the root node n1 (Do you like food?). If the answer of the user contains words like "Love", "Yes" or "Food", which are the keywords associated with the branch connecting n1 to n2, the next agent's utterance will be n2 (I love pizza, do you?). Otherwise, if the answer of the user contains words like "No", "Hate" or "Food", which are the keywords associated with the branch connecting n1 to n3, the next agent's utterance will be n3 (Oh ok, but i assume that you eat every day, right?).

It is important to point out some things here. First, how are the keywords associated with each branch? For example, it is possible to see in Figure 3.5 that the keyword "Food" is present in both branches (to n2 and to n3). The keywords are chosen based on frequency. Let us take the example presented in Figure 3.5. Depending of how people answer the first question n1 (Do you like food?) and to where it goes (n2 or n3), the keywords are assigned to each set. For example, if someone answers n1 with something like "I hate food" and travels to n3, the words "hate" and "food" are added to the keywords set of its respective branch.

Second, it is possible to see in Figure 3.5 some numbers associated with each keyword (for example, "Love" has 6 and "Food" has 5). These numbers represent the amount of times that such keyword appeared during interactions and were used to reach their respective nodes. For example, during many interactions between the agent and people, "Love" was used 6 times to go from n1 to n2. This number can be used to calculate the frequency that each keyword appears in interactions, both alone and relative to other keywords of the set. This way, we can know

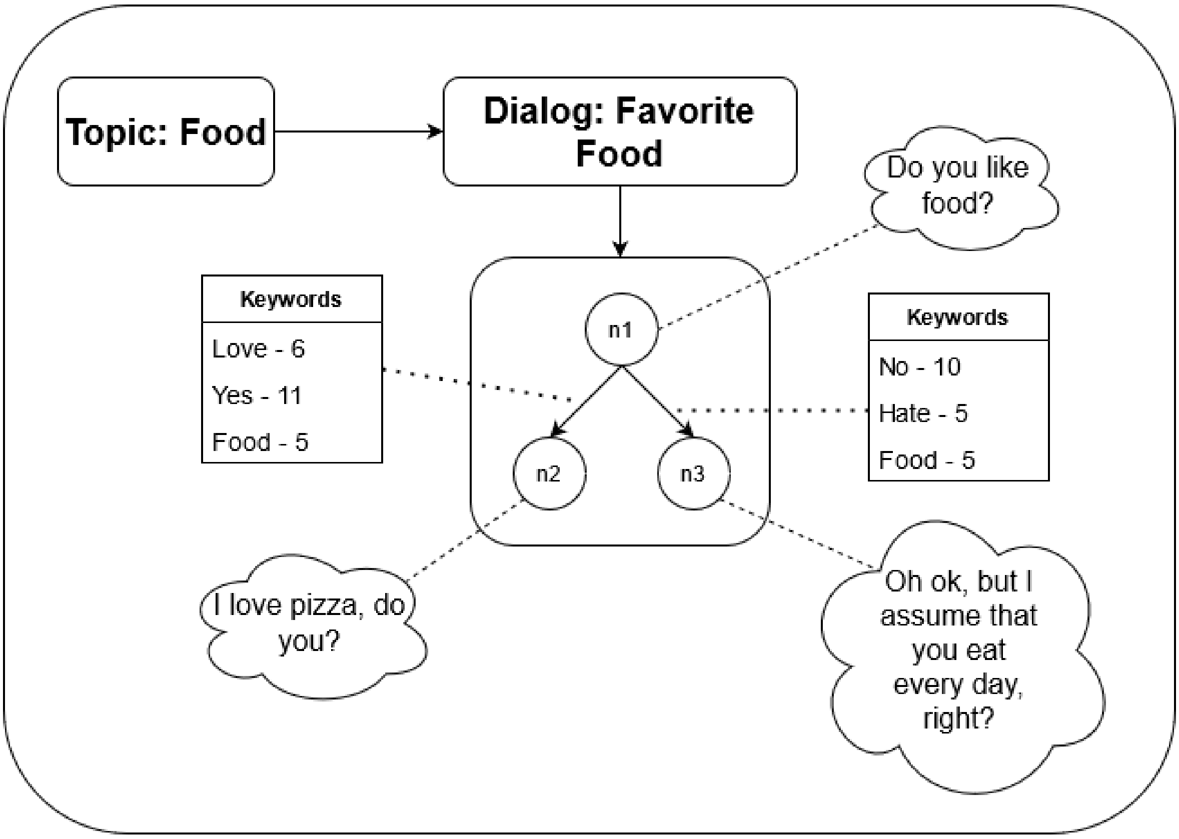


Figure 3.5: Small talk example. Inside the Topic: Food and the Dialog: Favorite Food there is a Dialog Tree with three nodes (which means, three possible utterances for the agent). The dialog starts with the root node n1 (Do you like food?). If the answer of the user contains words like "Love", "Yes" or "Food", the next agent's utterance will be n2 (I love pizza, do you?). Otherwise, if the answer of the user contains words like "No", "Hate" or "Food", the next agent's utterance will be n3 (Oh ok, but i assume that you eat every day, right?).

which keywords are more "important" to go from a node to another (for example, if many people answer n1 with the keyword "Love", its frequency is high and, therefore, we can give more weight to it when deciding to which node the agent should travel). It give us two advantages. Firstly: it does not matter that the same keyword is present in different sets (like "Food", in the example of Figure 3.5), because it can have different weights, as well other keywords can have much higher weights. Secondly: it allows Arthur and Bella to, based on the occurrence of words, learn from previous interactions and improve their decision-making when traveling down the Dialog Decision Tree.

We chose to work with two different frequencies, which we called Simple Frequency (*FSimple*) and Sibling Frequency (*FSibling*). The simple frequency is nothing more than the number of times *nt* that a given keyword *k* appears in node *n* in comparison with all other keywords which appear in this node (nt_{all}^{n}). The formulation goes as follows:

$$FSimple_k^n = \frac{nt_k^n}{nt_{all}^n}, \tag{3.1}$$

where $FSimple_k^n$ is going to be a value lying between 0 and 1. In its turn, the sibling frequency is the number of times nt that a given keyword k appears in node n in comparison with the number of times that k appeared in other nodes of the same tree level (nt_k^{level}). The formulation goes as follows:

$$FSibling_k^n = \frac{nt_k^n}{nt_k^{level}}, \tag{3.2}$$

where $FSibling_k^n$ is also going to be a value lying between 0 and 1. Finally, the final frequency is simply the mean value between the simple and the sibling frequencies, as follows:

$$F_k^n = \frac{(FSimple_k^n + FSibling_k^n)}{2}, \tag{3.3}$$

where F_k^n will be the frequency that keyword k appears in node n.

Finally, since Arthur or Bella only makes use of small talks when the interaction seems to "cool down", a timer was defined. We empirically defined that if the user says nothing to the virtual agent for 30 seconds, but stays in the web cam, Arthur or Bella randomly selects a pair Topic/Dialog and initiates a small talk conversation.

3.4 Face Recognition

The Face Recognition module allows for the virtual agent to recognize the person he/she is talking to. To do so, this work is based on the library developed in Python for face recognition[4]. In short, this library is a trained neural network which allows to recognize faces of people, comparing a given image with other images saved on a Data directory. For more information, please consult the link at the footnote.

[4] https://github.com/ageitgey/face_recognition

The pipeline is pretty simple, as shown in Figure 3.6. On Unity side, the data is captured from the webcam and saved as a PNG image, with the face of the person. Then, this information is sent to the face recognition module, on Python side. In its turn, this module returns a list with the names of the people who can be the person in the image, ordered by a discrepancy value which lies between 0 and 1. The less value of discrepancy, the more probable that the person in the list is the one in the image. If no person is found, the face recognition module returns an empty list. Finally, on Unity side, such list is taken and the person with the lower value of discrepancy is chosen. If the list is empty, we assume that there is a new person to be met, so Arthur or Bella asks the name of the person.

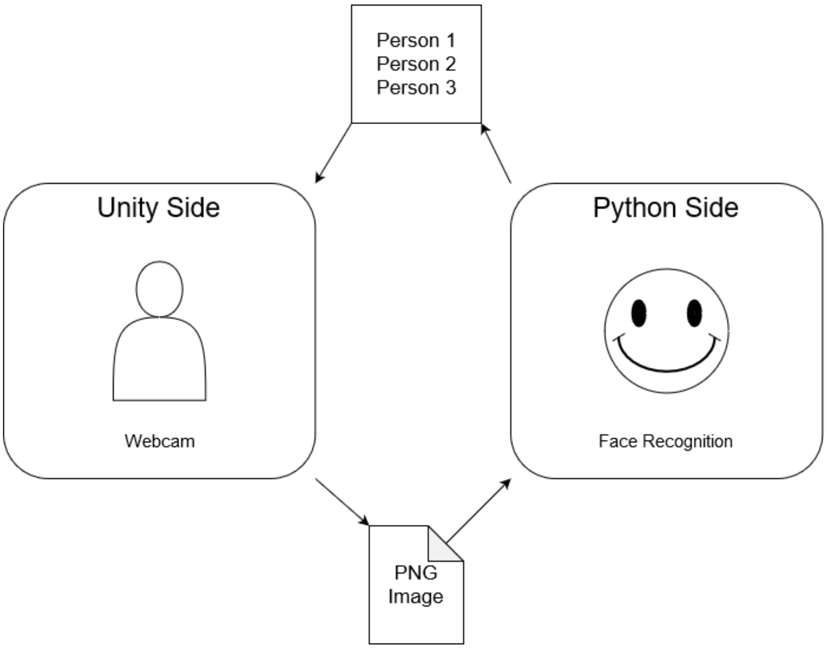


Figure 3.6: Pipeline of the Face Recognition module. On Unity side, the data is captured from the webcam and saved as a PNG image. On Python side, the face recognition loads this image and returns a list with the names of the people which it thinks can be the person in the image.

3.5 Emotion Detection

The Emotion Detection module is able to detect the emotion of the person who is talking with the virtual agent and appearing at the webcam. In order to do so, the Affectiva plugin[5], available for Unity, was used. It uses a trained neural network

[5] https://affectiva.com/

to identify many points in the face of the person. Depending on the disposition of such points, the network is able to predict the emotion that the person is feeling. For more information, please refer to the link on the footnote.

When a face is found in the webcam, the plugin immediately tries to determine the emotion of the person. Since emotion can vary very rapidly, it is checked at regular intervals of time (i.e., each second). The detected emotion can be used for two things: First, to determine the valence of a given information. For example, if a person talks about the death of its beloved pet, he/she is probably going to present a sad face, which is going to reflect in a sad memory for the agent. Second, to model empathy. For example, if the virtual agent sees a happy person, it can also become happier. We also use the Valence attribute, returned from Affectiva, to update the mood of Arthur or Bella. As explained by Affectiva, Valence is "a measure of the positive or negative nature of the recorded person's experience". More information is going to be provided in Section 3.10, where we present the Empathy Module.

3.6 ECAs' Beliefs

The Beliefs Model was built in the sense of having a manner for our virtual agent to be able to reason regarding different pieces of information. We, as human beings, are able to make connections in a natural and instantaneous way. For instance, if a person says that he/she has two children, named John and Mary, we automatically assume that John and Mary are siblings. So, this kind of reasoning has to be coded for Arthur and Bella.

In order to encode such feature into our ECA, we proposed a knowledge-based system of statements, which are incorporated into Arthur and Bella. The statements can be produced in two different ways. First, they can be manually defined: this way, anyone can define tailored PROLOG statements that can attend to their own goals. As would be expected, in order to manually define such statements, one must have knowledge in PROLOG. Since not everyone has such knowledge, we also defined a second way to produce statements: they are automatically created based on the memory of the agent. As defined in Section 3.8, Arthur is endowed with a human-like memory model, where much information is stored while the agent interacts with people. We take such memories and use them to create PROLOG statements.

Using the previous example, if a user tells Arthur/Bella that he/she has two children, named John and Mary, the ECA is going to store this information on its memory and we can create two statements: parent(user,john) and parent(user,mary). Such statements can, also, be used for more complex relationships: if a siblings statement is created (i.e., sibling(X,Z) :- parent(Y,X), parent(Y,Z), X!=Z), Arthur/Bella can infer that John and Mary are siblings. We included two tailored beliefs, as follows:

- Sibling: sibling(X,Z) :- parent(Y,X), parent(Y,Z), X!=Z
- Grandparent: grandparent(X,Z) :- parent(X,Y), parent(Y,Z)

3.7 ECAs' Self Memory and Common Sense

When we talk with someone else, it is common to ask some questions about the other person. Moreover, it is perfectly natural that the person talking with the virtual agent would also ask these kinds of questions and, therefore, the ECA needs to have such knowledge about itself in order to answer. The Self Memory Model was built to endow Arthur/Bella with some knowledge about themselves, aiding in making the interaction more natural. To do so, we manually included some information in its initial memory concerning some casual topics that could be chosen by the user (i.e., name, age, tastes, etc.). At any given time, when the user asks about such topics, Arthur/Bella can search its own memory and give a proper response.

In a similar way, as we, human beings, grow up, we learn many and many things about an infinity of topics. Since how to walk to why the sky is blue, we have encoded in our memory a vast knowledge about things that, sometimes, we do not even know how to explain [vHO87]. The Common Sense module aims to give Arthur/Bella this kind of knowledge. In order to do so, we chose to work with Wordnet [Mil95], a large lexical database of English terms which express the concept of many nouns, verbs and so on. It is comprised of about 150.000 words, alongside their respective description. We ordered all these words by their sense-number[6], as provided by the database, and included the first 10.000 pairs of words/descriptions of this database into our agent's memory. Thus, when Arthur/Bella is asked about any of these terms, he/she can answer with the proper description. Finally, since the

[6] A way to represent the word relevance included in the dataset.

Wordnet's terms are inserted into the agent's memory, the Common Sense module is directly related with the memory of the agent. The pairs of words/descriptions are translated into the autobiographical memory, becoming part of the knowledge of the ECA.

3.8 Agent Memory

The memory of the virtual agent is used to register different data collected during the interactions with people. To do so, it was adopted a model known as Autobiographical Memory. In short, as defined by Bluck et al. [BL98], autobiographical memory is "a system that encodes, stores and guides retrieval of all episodic information related to our person experiences". Therefore, it is able to store different kind of information (e.g. text, episodes, images) and retrieve it if a certain cue (or cues) are informed. For example, when one meets a new person, usually both of them greet each other and tell its respective names. If the memory serves this person well, he/she is able to remember the name of the other person when they meet again.

According to Conway et al. [CPP00], Autobiographical Memory can be grouped into three levels: Lifetime Periods, General Events and Event-Specific Knowledge. Lifetime Periods serve as an index to cue General Events and can be linked to different periods across the lifespan of a person. For example, the first job of someone can be seen as a lifetime period. General Events are events which occur inside a given lifetime period. Using the first job example given before, some general events inside this could be the first day at job, meeting a new colleague or a happy hour with people of the office. Finally, Event-Specific Knowledge (ESK) includes a pool of resources which form a memory of a given event. Such information can be stored in different types, like images, grammatical or audio. Figure 3.7 shows the modeling of the autobiographical memory knowledge base, as proposed by Conway et al. [CPP00].

As it is possible to see in Figure 3.1, we model our memory using General Events and ESK. As the intent of our virtual agent is to interact with people, we chose not to model Lifetime Periods. General Events represent the events which occur during the interaction between Arthur/Bella and the user. For example, when Arthur/Bella meets someone new, a new General Event is generated (i.e., Meet new person). More details about General Events are given in Section 3.8.1. Moreover,

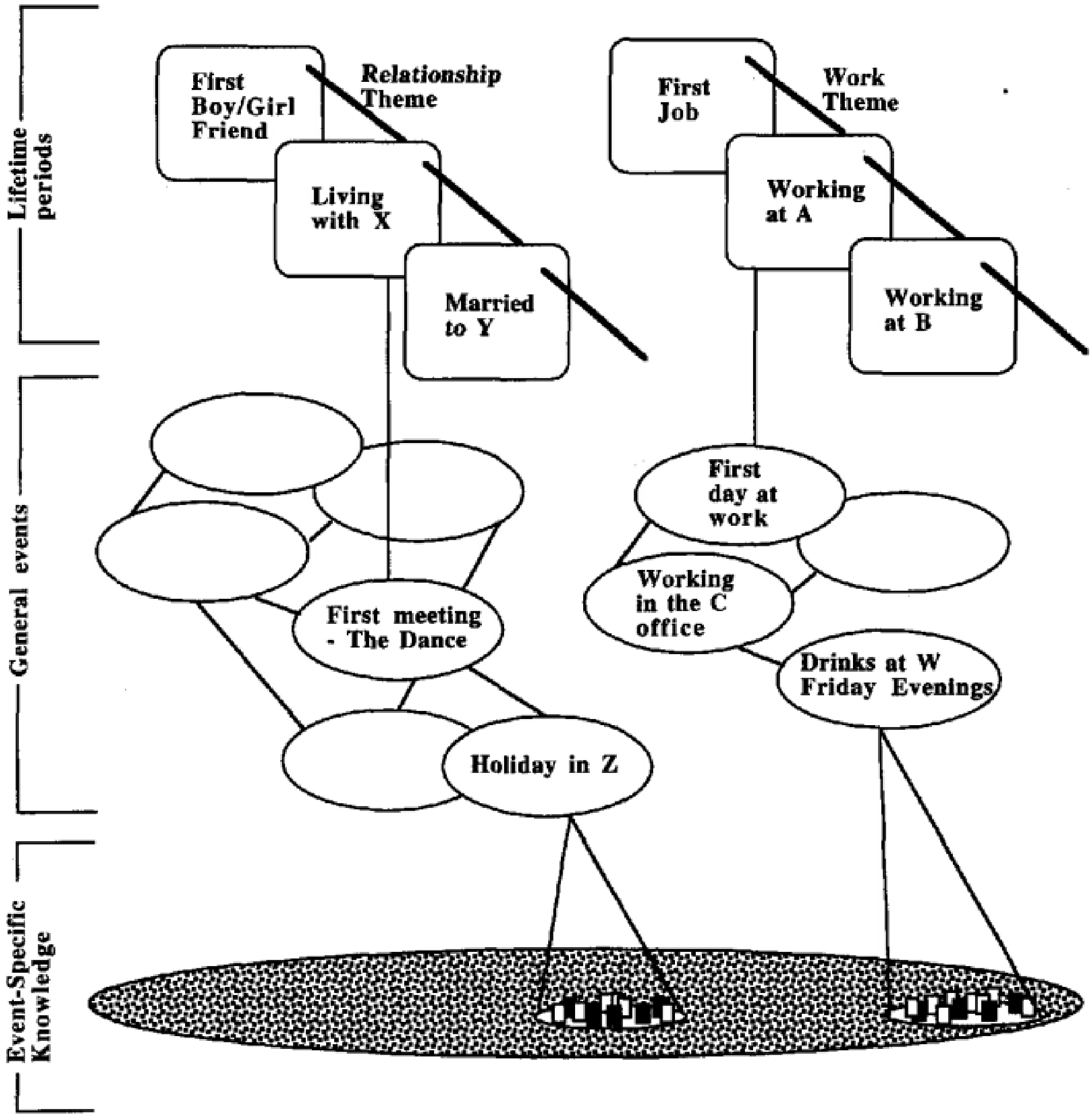


Figure 3.7: Autobiographical memory knowledge base. While general events model different events that has happened, event-specific knowledge (ESK) model the details of such events. Therefore, ESK is shown as an undifferentiated pool of resources which can be activated if certain cues are given. Source: Conway et al. [CPP00].

following the Autobiographical Memory model, we also model a pool of information which represent the Event-Specific Knowledge (ESK) of the memory model. We call each piece of information (e.g., each word/term, each image) a resource. Finally, we store both General Events and resources in two levels: Short-Term Memory (STM) and Long-Term Memory (LTM). More details are given in next sections.

3.8.1 General Events

In our model, General Events are comprised of:

- Timestamp: the moment this event was created or updated.
- ID: an unique ID which refers to this event.
- Type: refers to the type of the event and there are currently three possible types: *i)* Belief: when the event is related with some belief of the virtual agent; *ii)* Person: when the event is related with something about the user; and *iii)* Agent: when the event is related with something about Arthur or Bella.
- Information: A brief explanation about what the event is about.
- Polarity [-1,1]: the polarity of the sentence, it means, if it is a positive or negative sentence. For example, the sentence "I am feeling good" would be a positive sentence, while the sentence "I am not feeling good" would be a negative one. When we divide the sentence into tokens, we also use the sentiment library[7], provided by NLTK, to calculate the polarity of the given sentence. To do so, we chose to work with Vader method. It returns a float value lying between -1 and 1, where negative values reflect a negative polarity and positive values reflect a positive polarity.
- Resources: a list with the resources (i.e., ESK) associated with this General Event. Using the previous example, some resources associated with this event could be a picture of the deceased pet and some grammatical information like the pet name, death and when it passed away.

3.8.2 Event-Specific Knowledge

As commented before, we model a pool of information which represent the Event-Specific Knowledge (ESK) of the memory model. We call each piece of information (e.g., each word/term, each image) a resource. Each resource has the following information:

[7] https://www.nltk.org/howto/sentiment.html

- Timestamp: the moment this resource was created or updated.
- ID: an unique ID which refers to this resource.
- Type: refers to the type of the resource. We classify each resource as a 5W1H (who, what, when, where, why, how) information, so, the possible types are: Person, Location, Time, Activity, Emotion and Imagery. Additionally, we included a seventh possible type: Object.
- Information: refers to the information itself that the resource should reflect. For example, if the resource type is Imagery, the information is going to contain the path to such image.
- Activation [0,1]: As described in Loftus et al. [LL19], an information present in the short-term memory is rapidly forgotten (around 15 seconds), unless it is rehearsed, it means, repeated over and over. The activation represents this rehearsal process. When a new resource enters the STM of Arthur, activation is set at its maximum value (i.e., 1). As time passes by, a logarithmic-based decay function is applied for each resource, diminishing their activation value. It is defined as follows: $A^{*}_{ID_{STM}} = Log(A_{ID_{STM}} + 1)$, where $A^{*}_{ID_{STM}}$ is the new activation value of STM for resource ID, $A_{ID_{STM}}$ is the current activation value and $Log()$ is a logarithmic function. If any resource in the memory is rehearsed (e.g. remembered, seen again), its activation value is set back to 1. This attribute exists only for STM.
- Weight [0,1]: represents the importance of the resource. For example, meeting a new person can be considered more important than talking about the weather. We empirically defined the initial importance of each resource (at STM). Core memories (like the agent's beliefs and knowledge about itself) have weight = 1 and can not be forgotten. Important memories (like the things that the agent learns from the user) have weight = 0.9. Finally, non-important memories have weight = 0.1. At LTM, this attribute is not initialized but indeed affected by the values of $W^{*}_{ID_{LTM}} = f(A^{*}_{ID_{STM}}, W^{*}_{ID_{STM}})$ at STM, in the consolidation of the memory, as explained in Section 3.8.5.

In order to store grammatical resources, we divide a given sentence in significant tokens and keep each of them separately. To do so, we use the Natural Language ToolKit (NLTK)[8], a platform for building Python programs to work with

[8] https://www.nltk.org/

human language data. Such tool is able to "tokenize" a text, it means, split it in sentence or word tokens. It is also able to remove "stop words" from the text, it means, words which have low or no meaningfulness for the context of the sentence. We built a script which is able to remove the "stop words" from the text and "tokenize" it word by word. For example, assuming the user told to the virtual agent "I am going on vacation with my dad to Glasgow", the tokens returned by the NLTK script would be "vacation", "dad" and "Glasgow".

3.8.3 STM and LTM

As commented before, we store both General Events and resources in two levels: STM and LTM. According to Loftus et al. [LL19], STM is used to store important information for a short period of time (i.e., at most, 15 seconds), while LTM is an information storage with virtually unlimited capacity that each human being has. In our work, we model STM and LTM separately. The STM is comprised of two lists: one for General Events and other for resources. Both lists can have, at most, seven items, as defined by Miller's Law [Mil56]. If a new resource/General Event should enter the STM, the less important information is forgotten. For the resources, we check the weight value (i.e., the resource with the lower weight is removed). For the General Events, we check the weight of the resources which are connected with each event: the lowest average value is removed. It is important to note that this process is rarely triggered for the General Events: it is most likely for the STM to have more resources than General Events, because each event is usually connected with two or more resources.

Figure 3.8 shows the organization of both Short and Long-Term Memory. The same type of resources/General Events can be stored in STM and LTM, but the LTM is theoretically unlimited, it means, can maintain an unlimited number of resources. To keep track of the content of the Long-term memory, we save all its information in a database (both General Events and ESK), so we can retrieve it at any moment. Finally, both General Events and resources can be transferred from the STM to the LTM by a consolidation process, as detailed in Section 3.8.5.

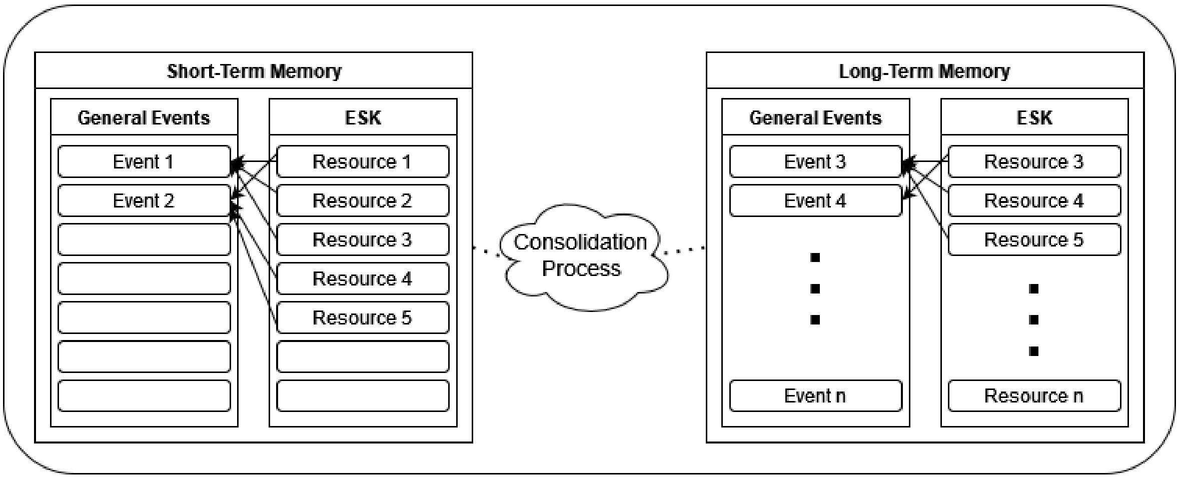


Figure 3.8: Organization of both Short and Long-Term Memory. For the Short-Term Memory (STM), each list can have at most seven itens. In its turn, the Long-Term Memory (LTM) is virtually unlimited. The Consolidation Process is responsible to consolidate the memory from STM to LTM, as detailed in Section 3.8.5.

3.8.4 Memory Retrieval

Assuming the information is stored, how can one retrieve such information if required? The Memory Retrieval module deals with the retrieval of the information from the storage. According to Conway et al. [CPP00], there are two types of memory construction: Generative Retrieval and Direct Retrieval. Generative retrieval is a method guided by cues, it means, the retrieval depends on some "hint" in order to find some information. For example, the word "dog" can make some people remember of their beloved pet. Thus, the word "dog" acted like a cue to the generative retrieval method, constructing a memory of a special dog. In its turn, Direct Retrieval method refers to spontaneously recalled memories, it means, memories that are recalled automatically, with no apparent cue. According to Berntsen [Ber96], such process occurs between two to three times each day. In this work, the memory retrieval is developed following the Generative Retrieval method. When the user interacts with the virtual agent, the information provided can be used as cue(s) to the Retrieval method.

The retrieval process can occur both on STM and LTM. Firstly, the process is done on the STM. If it finds a general event there with the given cue(s), such event is returned and the process is finished. Otherwise, the process is repeated on the LTM. Again, if it finds a general event there with the given cue(s), such event is returned and moved to the STM, finishing the process. Otherwise, nothing is returned, meaning Arthur or Bella have no memory about this/these cue/cues. Figure 3.9 shows a simplified example of the grammatical memory organization after the sentence "I am going on vacation with my dad to Glasgow" is informed to the virtual agent. So, for example, if the user tells "I went fishing with my dad", the words "fish" and "dad" are used as cues for the retrieval method. Supposing Arthur or Bella has the General Event represented in Figure 3.9 in his/her memory, the cue "dad" is found; thus, that general event is retrieved as a memory of the user's dad. With this information, the virtual agent can tell the user that it remembers that the user went on a vacation in Glasgow with its dad. Finally, supposing that the virtual agent's memory has many general events with the same cue (e.g. "dad"), the general event with the most cues is selected. For example, assuming that Arthur or Bella has two general events in his/her memory, as follows: "vacation-dad-Glasgow" and "fish-dad". Now, the cues "dad" and "Glasgow" are informed. In this case, the first general event ("vacation-dad-Glasgow") is returned, because it has both of the cues informed ("dad" and "Glasgow"), while the second general event has only one ("dad"). On other had, if the cues informed are "dad" and "fish", the second one is selected. If a tie occurs (for example, informing only the cue "dad"), the first one found in the memory is selected.

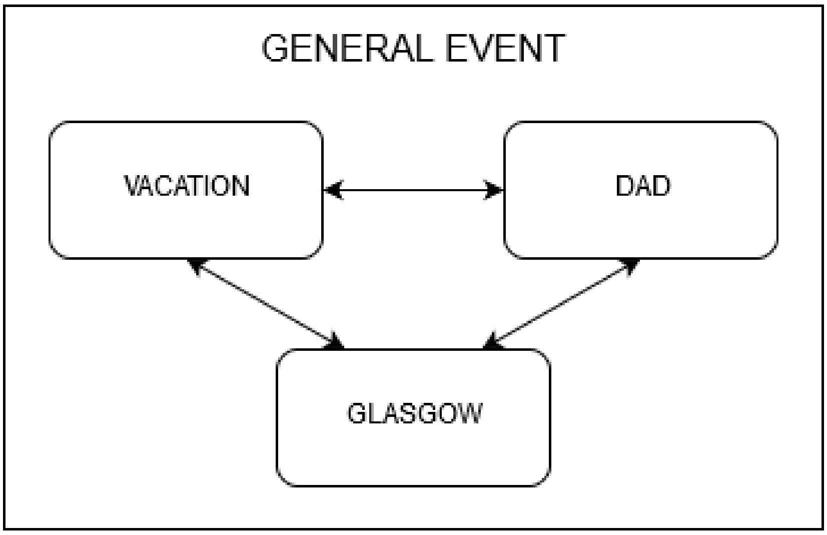


Figure 3.9: Simplified example of grammatical memory organization (i.e., just words), after the sentence "I am going on vacation with my dad to Glasgow" is informed to the virtual agent. The general event would contain the three more significant words: "vacation", "dad" and "Glasgow".

3.8.5 Memory Consolidation

This module is responsible for consolidating the Long-Term Memory of the virtual agent based on data available at STM. According Klinzing et al. [KNB19], the formation of such LTM is a major function of sleep. As their work affirms, the authors "consider the formation of long-term memory during sleep as an active systems consolidation process that is embedded in a process of global synaptic downscaling". In short, the consolidation process prioritizes important memories over mundane ones. For example, emotional memories are more important and have more impact than neutral memories. Therefore, less important information can be placed in a second plan or, even, forgotten.

In this work, it was developed a module to simulate such process, which uses two information from the STM's resources: Activation and Weight. If the Activation value A_{ID} of a given resource ID is below an empirically defined threshold (i.e., $A_{ID} < 0.2$), the Weight attribute of this resource is reduced at STM, as follows:

$$W^{*}_{ID_{LTM}} = Log(W_{ID_{STM}} + 1), \tag{3.4}$$

where $W^{*}_{ID_{LTM}}$ is the new Weight of this resource, $W_{ID_{STM}}$ is the current weight at STM, and $Log()$ is a logarithmic function. Such process is repeated for all stored resources at STM. Then, resources which have low importance are wiped out from the STM, while the other resources are transferred to the LTM. We empirically define that a resource has low importance if its weight drops below 0.2. In its turn, General Events are not forgotten by the virtual agent, unless all the resources belonging to a given event are also forgotten. We do so to make possible for the agent to forget just parts of the information, just like it happens with a real person. In addition, please notice that during the memory consolidation, all data at STM is erased.

3.9 Facial Expressions

It is important that all features, described so far, work as intended in order to make the virtual agent reach his/her goals. However, since an ECA is being modeled, it is equally important that such agent presents itself in a believable and natural way.

In this work, we built two different embodiments: Arthur and Bella. Arthur portraits a 2D model of a man, while Bella portraits a 3D model of a woman. Concerning Arthur, in order to model the virtual agent facial characteristics, the following approach was used: the various parts of agent's face (such as eyes, mouth, etc) where created using Unity3D sprites. Then, after placing each of them into their position, it was defined a set of constraints in order to avoid further distortions, e.g., when the agent's head moves, the movement is followed by each part. Finally, the different expressions were modeled using a plug-in called Anima2D [9].

The process of animation of Anima2D consists in the usage of skeletal animation. For a given mesh in the face that should suffer some kind of deformation, such as the eyebrows and mouth, a set of bones is defined and linked to this mesh. After that, it is possible to animate the agent through the time just by applying transformations into the bones. Therefore, after animating a certain facial expression (which can be either animated or a static one), it is possible to save it for further usage. The current expressions of the agent are the six basic emotions defined by Ekman [Ekm92], namely happiness, fear, disgust, anger, surprise and sadness. Also, there is a neutral expression and one for sleeping (in order to indicate that the memory consolidation process was triggered, as explained at the previous section). Figure 3.10 presents Arthur portraying all the six basic emotions, from top-left to bottom-right: happiness, sadness, anger, disgust, surprise and fear.

In order to give some expression to the eyes, it was implemented saccade movements in agent's eyes, according to Lee et al [LBB02]. As commented in Section 2.3, human eyes are not static for much time, tending to have involuntary movements, even when the person is focusing his/her vision on something. Such phenomena is known as saccade, from the french word *saccade*, which means *jerk*. As defined by Leigh and Zee [LZ15], saccades are rapid movements of both eyes from one gaze position to another. Finally, a lip sync tool was developed in order to control the mouth movement when Arthur speaks.

Concerning Bella, we acquired a 3D rigged model of a cartoon female head [10] and have made some modifications, especially concerning her expressions. Based on the work of Melgare et al. [MMSQ19], we have also modeled the six basic emotions defined by Ekman [Ekm92], namely happiness, fear, disgust, anger, surprise and sadness. The modeling followed the blendshapes already provided by the rigged model. Although the original model provided some possibilities for hair,

[9] https://assetstore.unity.com/packages/essentials/unity-anima2d-79840

[10] https://www.turbosquid.com/pt_br/3d-models/rigged-female-head-face-morphs-3d-max/917863

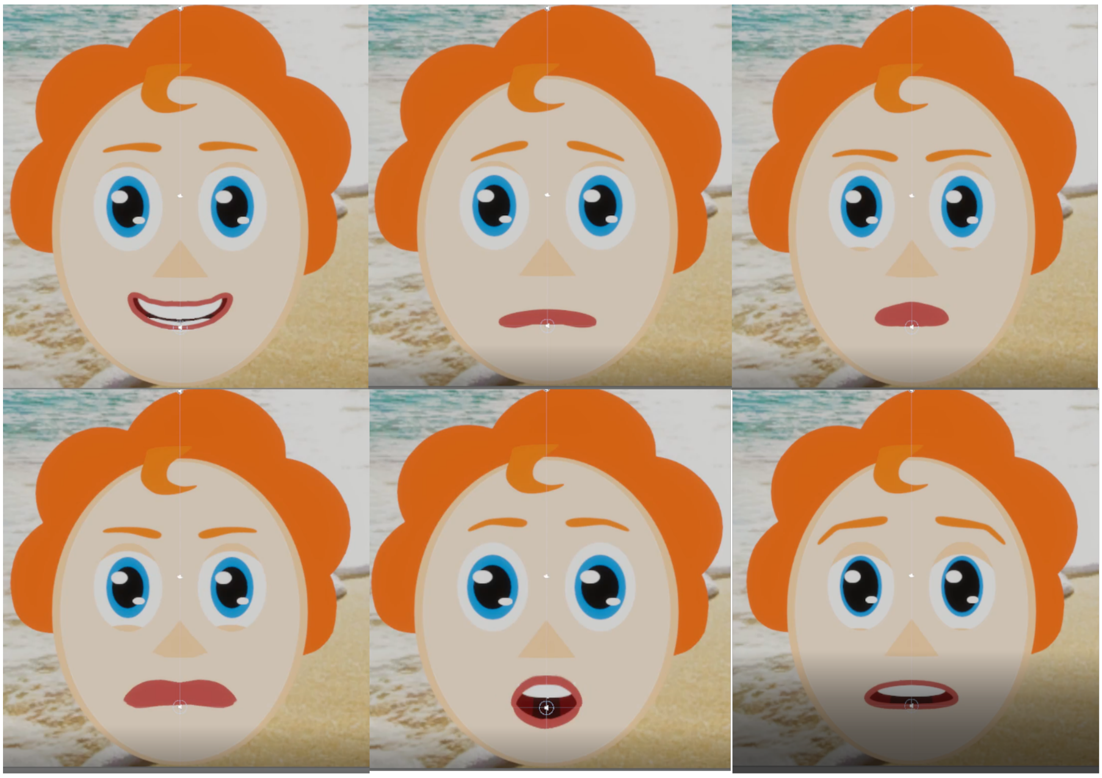


Figure 3.10: Arthur's basic emotions. From top-left to bottom-right: happiness, sadness, anger, disgust, surprise and fear.

they were too complex for our purpose. Thus, we used the bald model and modeled a simpler hair for Bella. Saccade eye movements and lip sync were also added for her, as it was done with Arthur. Figure 3.11 presents Bella portraying all the six basic emotions, from top-left to bottom-right: happiness, sadness, anger, disgust, surprise and fear.

Even following the literature and developing facial expressions for the six basic emotions, how can we be sure that people are going to be able to perceive such expressions? In order to see if people can correctly perceive the emotions of Arthur and Bella, we conducted a research to evaluate the perception of people concerning the six basic emotions modeled. Fifty eight volunteers participated in the experiment with Bella, where 22 were men and 36 were women. Twenty six people informed that they had some familiarity with graphical computing, while another 26 informed that they had not and 6 preferred not to answer the question. Their task was to watch video sequences where Bella expressed her emotions and answer which emotion they were able to identify (if any). The results show that the participants were able to identify all six emotions, being Happiness the easiest to identify (98.8%) and Anger/Fear the hardest ones (82.7%, both). For the experiment

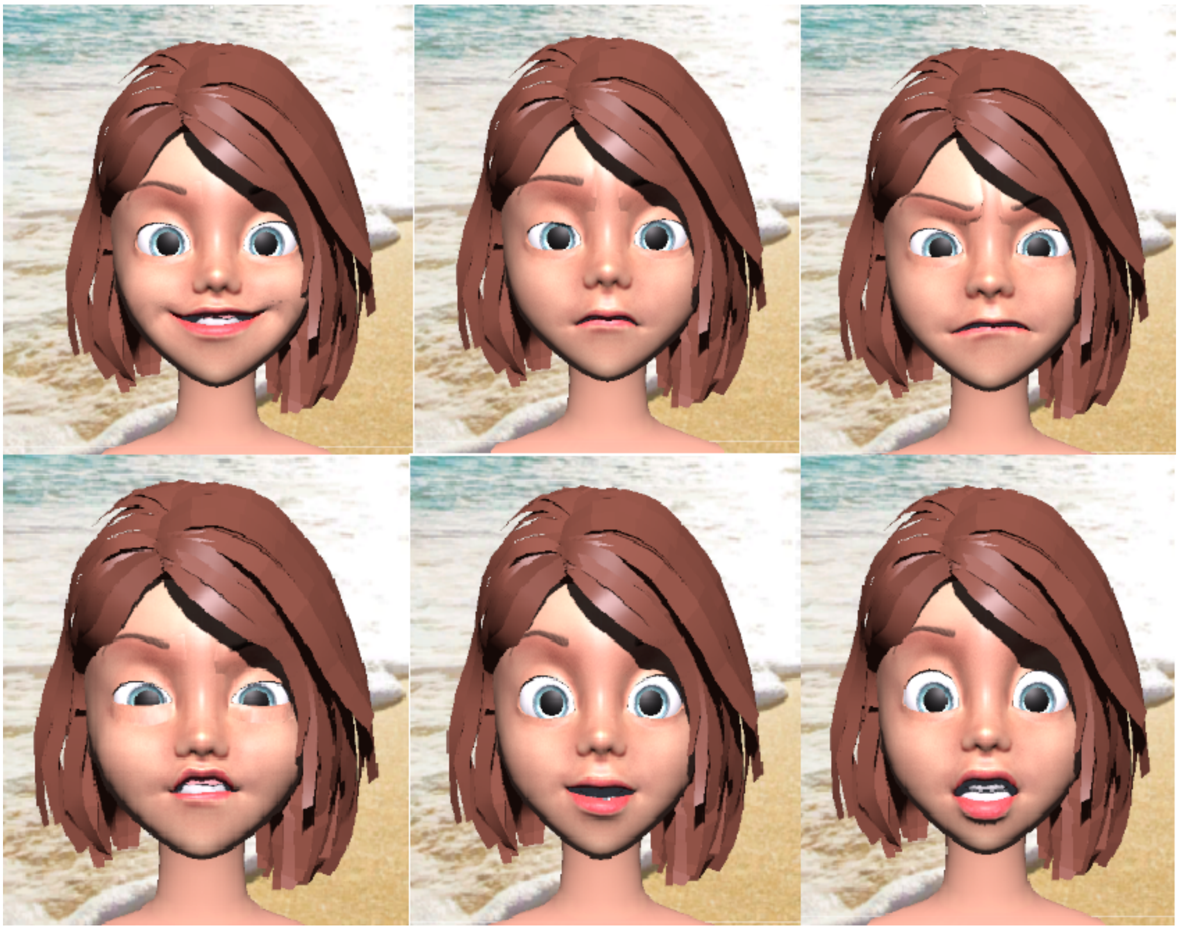


Figure 3.11: Bella's basic emotions. From top-left to bottom-right: happiness, sadness, anger, disgust, surprise and fear.

with Arthur, another 58 volunteers were selected, being 32 women, 25 men and 1 "Other'. Thirty eight people informed that they had some familiarity with graphical computing, while another 16 informed that they had not and 4 did not answer the question. The results show that the participants were able to identify the majority of the emotions, being Surprise and Happiness the easiest to identify (98.2% and 96.5%, respectively). However, participants had problems to identify Anger (34.8%). In short, people were able to identify the majority of the emotions expressed by both Arthur and Bella. Table 3.1 presents the results.

3.10 Empathy

Empathy can involve cognitive attributes or affective attributes which also can be combined [GM85]. Cognitive attributes of empathy involve cognitive reason-

Table 3.1: Emotion perceived by people for both Arthur and Bella.

Emotion	**Arthur**	**Bella**
Happiness	96.5%	98.8%
Sadness	91.3%	90.1%
Anger	34.8%	82.7%
Fear	53.4%	82.7%
Surprise	98.2%	91.4%
Disgust	51.7%	88.9%

ing used to understand another person's experience [Hoj07]. Emotional or affective attributes involve physiological enthusiasm and spontaneous affective responses to someone else's display of emotions [PMI05]. Moreover, it is known that both verbal and non-verbal communication can be useful when emulating empathy [TM07]. Concerning our empathy model, Arthur/Bella were endowed with three main human-like characteristics: personality, emotion and mood. Personality is understood as "characteristics of a virtual human that distinguishes him/her from the others", while emotion can be seen as "a state of mind that is only momentary" and mood as "a prolonged state of mind, resulting from a cumulative effect of emotions" [Ksh02]. While the personality of our agent is defined based on the OCEAN model [Gol90], we model the emotional states of Arthur and Bella with PAD dimensions [RM77]. Endowing our virtual agent with the ability to change its emotional state dynamically allows it to behave in an empathetic way towards the user. Next section details such characteristics.

3.10.1 PAD

In order to endow Arthur/Bella with mood valence, we chose to work with PAD space. The PAD space was introduced by Russel and Mehrabian [RM77] and stands for Pleasure, Arousal and Dominance. Each of these dimensions range from -1 to 1 and represent an axis in a three-dimensional space. Figure 3.12 presents a simple example of the organization of PAD three-dimensional space. According to the authors, this three-dimensional space is a good alternative to define and represent many emotional states. Moreover, they also suggest 151 different emotional states represented inside the PAD space. In this work, we decided to use 13 emotional states, as defined by Russel and Mehrabian [RM77] (with the exception of the Neutral emotional state, which we defined as a starting point at the intersection of the three PAD dimensions) and shown in Table 3.2. These emotions were chosen based on the six basic emotions used in this work (i.e., Happiness, Sadness, Disgust, Anger, Surprise and Fear), alongside Neutral and Bored emotional states. The initial PAD state of the virtual agent can be updated and used to change the agent's emotion. More details are provided in Section 3.10.3.

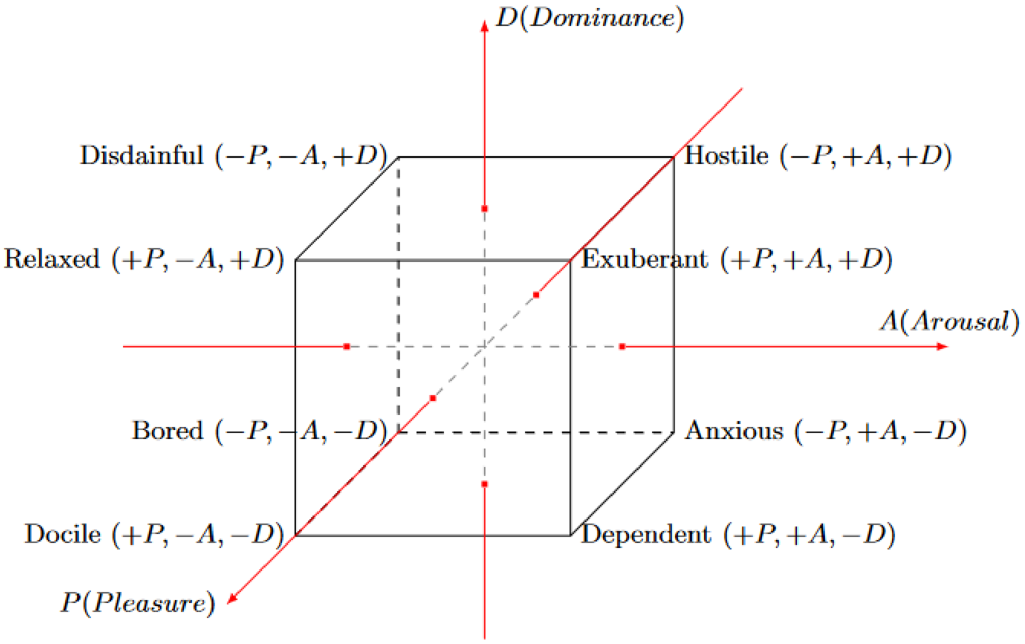


Figure 3.12: A simple example of PAD space. Each of these dimensions range from -1 to 1 and represent an axis in a three-dimensional space. Source: Tarasenko et al. [Tar10]

3.10.2 Personality

In order to define the personality of our agent, we chose to work with the OCEAN model, also known as Big Five, proposed by Goldberg [Gol90]. Based on the work of Sajjadi et al. [SHCK19], we assigned a personality profile to our virtual agent focused on the extrovert/introvert trait, limited by the Extraversion (E) trait. The agent is considered introvert if $E = [0, 0.5)$, and considered extrovert if $E = [0.5, 1]$. This profile is transferred to our PAD three-dimensional space, generating a default emotional state based on the personality of Arthur/Bella, as follows: By default, the extrovert personality profile is translated to the PAD space with the values $PAD_E = (0.8, 0.5, 1)$, being it Pleasure, Arousal and Dominance, respectively. On the other

Table 3.2: Emotional states of our ECA, adapted from Russel and Mehrabian [RM77]. P stands for Pleasure, A stands for Arousal and D stands for Dominance.

Emotional state	P	A	D
Neutral	0	0	0
Friendly	0.69	0.35	0.3
Happy	0.81	0.51	0.46
Surprised	0.4	0.67	-0.13
Angry	-0.51	0.59	0.25
Enraged	-0.44	0.72	0.32
Frustrated	-0.64	0.52	0.35
Fearful	-0.64	0.6	-0.43
Confused	-0.53	0.27	-0.32
Depressed	-0.72	-0.29	-0.41
Bored	-0.65	-0.62	-0.33
Sad	-0.63	-0.27	-0.33
Disgust	-0.60	0.35	0.11

hand, the introvert personality profile is translated to the PAD space with the values $PAD_I = (-0.8, 0.3, -1)$. Both of those default values were set based on the work of McCrae et al. [MCM05].

Moreover, besides the Extraversion dimension used by Sajjadi et al. [SHCK19], we also include the Neuroticism dimension to define the default emotional state. If Arthur/Bella has a Neuroticism value above 0.5 (values lie between 0 and 1), we assume that he/she is a bit paranoid and may not be feeling in the control of its own emotions. Therefore, if he/she is extrovert (i.e., $PAD_E = (0.8, 0.5, 1)$) and has a Neuroticism value above 0.5, we can reduce his/her Dominance, resulting in $PAD_E = (0.8, 0.5, 0.5)$. We chose to change Dominance based on its own definition. As defined by Mehrabian [Meh80], the Dominance space can be seen as a level of controlling/submissive feelings (for example, anger can be seen as a dominant emotion, while fear can be seen as a submissive emotion). Otherwise, if he/she is an introvert (i.e., $PAD_I = (-0.8, 0.3, -1)$) with a Neuroticism value lower or equal 0.5, we can increase his/her Dominance, resulting in $PAD_I = (-0.8, 0.3, -0.5)$.

PAD_E and PAD_I are used to define the initial emotional state of the agent, depending on the personality given. But those values also define a comfort zone for the virtual agent, it means, an emotional state in which he/she feels comfortable. More details are going to be provided in Section 3.10.3.

3.10.3 Emotional States

In this section, we discuss the modeling of the emotional states of the virtual agent, focusing on the Boredom and on the Emotional State Update.

Boredom

When two or more people are interacting, things can become weird if they stay some time without saying anything. This awkward silence can make people uncomfortable and even bored. In fact, Moreno et al. [MM07] state that interactions that experience an "under-loading" of information can lead to boredom and disengagement. In order to mimic such behavior, we chose to use the method proposed by Sajjadi et al. [SHCK19]. Based on the work of the authors, we included a Boredom value in our virtual agent, as follows: $Bor = [-1, 0]$, where -1 is the maximum value of boredom and 0 represents no bored at all. By default, the initial value of Boredom is set to 0, starting to increase (it means, towards -1) if the user stays 15 seconds (empirically defined) without interacting with Arthur/Bella. At any time an interaction occurs, the Boredom value is reset to 0.

The Boredom increasing occurs in a linear way. To do so, we chose to work with the personality profile of the virtual agent, focused on three of the OCEAN dimensions: Openness, Conscientiousness and Agreeableness. People with high values of Openness tend to be more curious and more creative. Also, according to Ambridge [Amb14], these people can present a lack of focus which, in our view, can lead to boredom. Conscientiousness can be described as a tendency for self-discipline. According with Toegel et al. [TB12], low values of this trait can be perceived as sloppiness which, in our view, could also lead to boredom. Finally, Agreeableness can be seen as a degree of social harmony. In general, low values represent people who put their own interest above others [BD19]. According Bartneck et al. [BVDHMAM07], such people can be seen as unfriendly and uncooperative which, in our view, can also lead to boredom. While Openness is directly related, Conscientiousness and Agreeableness are inversely related with Boredom: the lower their values, the more Boredom increases. We define the formulation as follows:

$$Bor = Bor - ((\frac{K}{2}.(O)) + (\frac{K}{2}.(1 - C)) + (K.(1 - A))), \tag{3.5}$$

where O, C and A stand for the three OCEAN dimensions cited before and K are constants that weight the Agreeableness (A) dimension as more important than the other two for the calculation. For this work, we defined $K = 0.5$. This value was empirically defined, so A impacts twice as much than the other two OCEAN traits used (i.e., C and O). Finally, the Boredom value *Bor* is also used for the emotional state update, as it is going to be explained in Section 3.10.3.

Emotional State Update and Empathetic Behavior

Empathy can involve cognitive or affective attributes, which also can be combined [GM85]. Cognitive attributes of empathy involve cognitive reasoning used to understand another person's experience [Hoj07]. Emotional or affective attributes involve physiological enthusiasm and spontaneous affective responses to someone else's display of emotions [PMI05]. When we talk with someone else, our emotions can change many times depending on how the interaction flows. In a similar way, an Embodied Conversational Agent endowed with emotion should be able to change its emotional state as interactions occur. In order to do so, we update the PAD state of Arthur/Bella in three specific situations, during the interactions with the user:

1. When the user says something;
2. When an emotion is recognized in the face of the user;
3. When something is remembered.

The update of the PAD values are done based on the work of Becker et al. [BKW04], adapted to our model as follows:

$$P = \frac{(P + Pol)}{2}, \tag{3.6}$$

where P is the Pleasure dimension of PAD and *Pol* is a value that lies between -1 and 1. In updating situation (1), the *Pol* value stands for the polarity of the sentence, meaning how positive or negative the sentence is. For example, if someone says "I woke up feeling really great today", it can be seen as a positive sentence. On the other hand, if someone says "I woke up feeling so bad today.", it can be seen as a negative sentence. In situations (2) and (3), *Pol* stands for the valence of the emotion, meaning how positive or negative the emotion is, being it recognized in the face of the person (situation 2) or associated with a given memory (situation 3).

For situation (2), the Affectiva plugin[11] is used to capture this information, which is also used for the detection of the emotion in the face of the user, as explained in Section 3.5. For situation (3), the emotional information is stored in the memory of the virtual agent, as explained in Section 3.8. Next, the Arousal dimension:

$$A = |Pol| + Bor, \tag{3.7}$$

where A is the Arousal dimension of PAD and Bor is the Boredom value explained in Section 3.10.3. The $|Pol|$ indicates that only the modulus of the Pol value is used. Finally, the Dominance dimension of PAD is a fixed value, so this value never changes. Its initial value can be changed depending on the personality of the agent (as explained in Section 3.10.1), but remains fixed during interactions. Also, as commented in Section 3.10.2, the emotion update is influenced by the comfort zone of the virtual agent. A bonus or penalty of 0.05 (empirically defined) is included in both Equations 3.6 and 3.7, being it a bonus if it is approaching the comfort zone (i.e., +0.05) and a penalty if it is distancing from the comfort zone (i.e., -0.05). Thus, the equations 3.8 and 3.9 replace equations 3.6 and 3.7 and are defined as follows:

$$P = (\frac{(P + Pol)}{2}) + Cz, \tag{3.8}$$

$$A = |Pol| + Bor + Cz, \tag{3.9}$$

where Cz stands for the comfort zone bonus or penalty. As commented before, the comfort zone acts as a magnet: if the emotional state update is approaching the comfort zone, it approaches faster. In a similar way, if the emotional state update is distancing from the comfort zone, it gets away slower.

At any given moment, when the PAD value is updated, the emotion of the virtual agent is updated as well to reflect this change. To do so, the closest emotion value (as defined in Table 3.2) is chosen. A simple distance function between two three-dimensional points is used. If the actual emotion is still the closest, no changes are made. Otherwise, if a different emotion is found to be closer than the current one, the new emotion is set and the respective animation is played by Arthur or Bella. When an emotion is identified in the face os the user (1), our virtual agent updates its own emotional state towards the identified emotion. As defined by Mehrabian et al. [ME72], we also define this sharing of emotions between the user and the agent

[11] https://affectiva.com/

as Affective Empathy. We define the other two situations (2) and (3) (i.e., when the user says something and when the agent remembers something) as Cognitive Empathy, which involve cognitive reasoning used to understand another person's experience [Hoj07], because both understanding the emotion implied in what the user says and in memories retrieved involve some level of cognitive reasoning.

3.11 Interface

So far in this chapter, we explained all our model and how Arthur/Bella behaves during interactions with people. In this section, we want to give attention to the interface built for our ECA. In Unity, we built two scenes: the initial menu and the interaction scene.

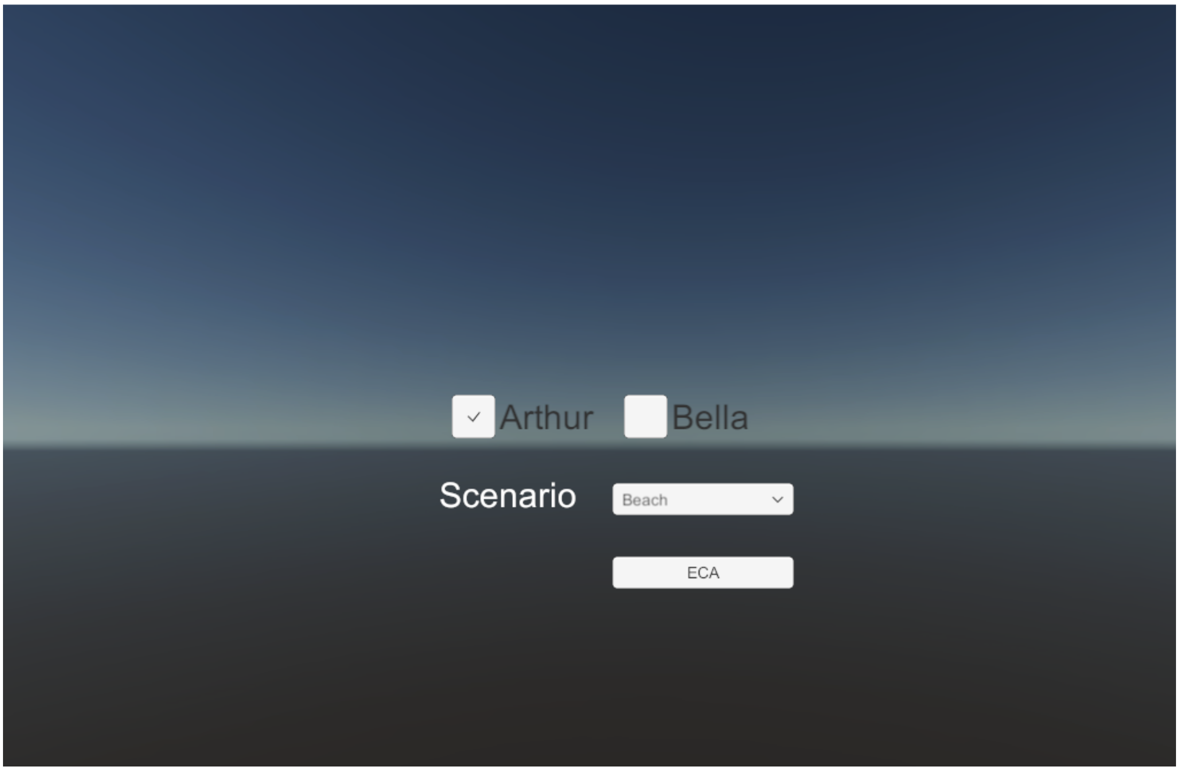
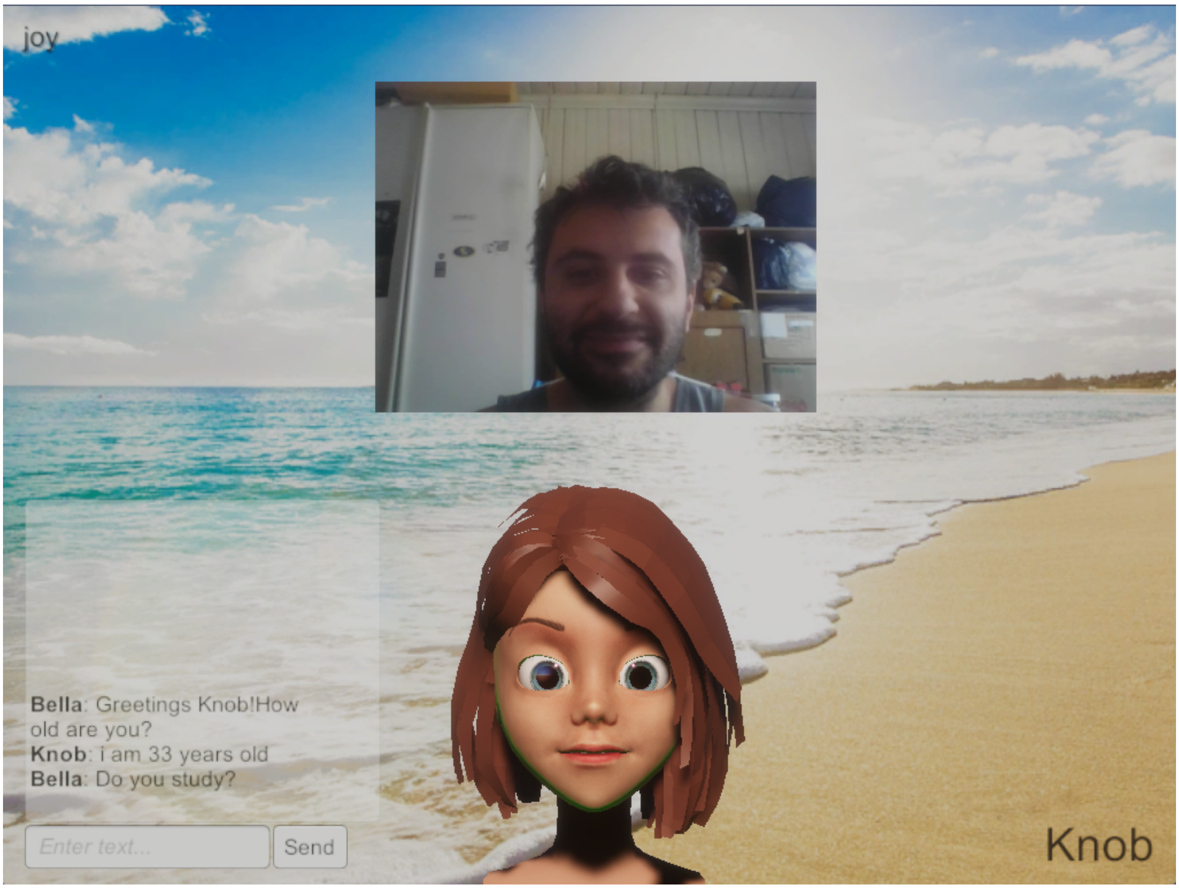


Figure 3.13: Initial menu. The user can choose if he/she wants to interact with Arthur or Bella, as well which background scenario he/she prefers (Beach, Office or Mountain).

Figure 3.13 shows the initial menu which is presented to the users. There, the user is able to select with which agent he/she wants to interact, Arthur or Bella. Moreover, the user can also select which scenario he/she wants to be shown in the background of the virtual agent. There are three options available: Beach, Office and Mountain. To start the interaction, the user just need to click on the "ECA" button.



Bella: Greetings Knob!How old are you?
Knob: I am 33 years old
Bella: Do you study?

Figure 3.14: Interaction scene. The user and Bella can be seen at the center, respectively at the top and the bottom. The name of the user can be seen at the bottom right, while the identified emotion can be seen at the top left. At the bottom left, the chat window is shown.

Figure 3.14 shows the interaction scene with our ECA. In the background, the "Beach" scenario can be seen. Right in the center of the figure, Bella and the person that she is talking with can be seen, respectively at the bottom and the top. The name of the person is shown at the bottom right of the figure, as well as the identified emotion for that person at the top left. The chat window can be seen at the bottom left, where the user can send messages to Bella and read everything that was already talked.

3.12 Chapter Considerations

This chapter presented the model proposed in this thesis to build an Embodied Conversational Agent (ECA). The main goal was to show how each part of the model was conceived and assembled together, as well to present how the model was built.

The next chapter is responsible to present the results achieved.

4. RESULTS

In this chapter, the results achieved by this work are presented. In order to test our method, we developed some experiments exploring the various features of Arthur and Bella. Mainly, we focus on the main contribution of our work: memory, empathy and the interplay between them. Firstly, Section 4.1 presents the results achieved by the memory feature of the virtual agent, both in scripted tasks and an experiment with subjects. Section 4.2 presents the results achieved in the empathy evaluation of the ECA.

For the experiments discussed in this section, the personality of the agent is set as the following OCEAN values: O = 0.9; C = 0.5; E = 0.9; A = 0.7; N = 0.5. The initial PAD value is, thus, set as follows: P = 0.8; A = 0.5; D = 1.

4.1 Memory Results

In order to evaluate the memory of Arthur and Bella, we conducted two different experiments. The first one involves scripted tasks, where we feed the virtual agent with some kind of information and check if he/she is able to store and remember this information. The second one involves an experiment with subjects and aims to check how people would evaluate the abilities of the virtual agent.

4.1.1 Scripted Tasks

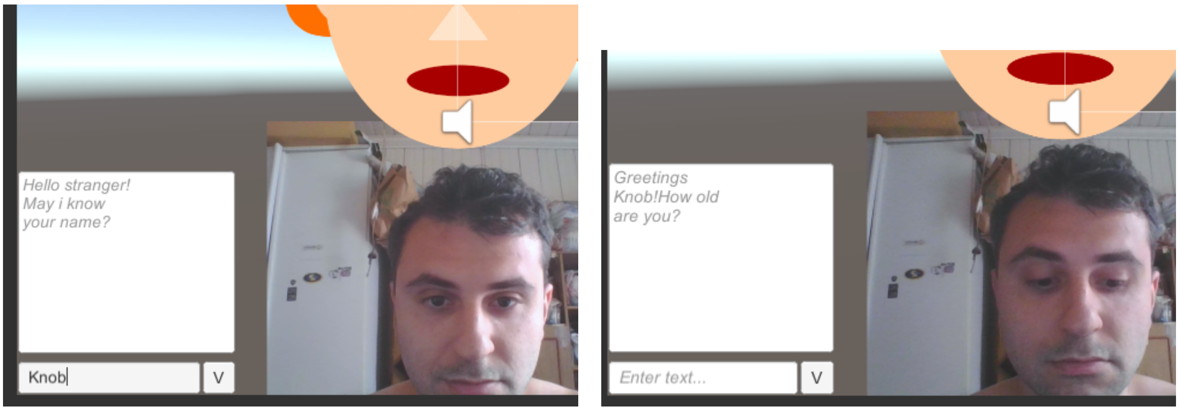
For the scripted tasks, we chose to work with two different scenarios:

- Introduction Scenario: Arthur meets a new person and starts to asking questions about him/her. After that, it is checked if Arthur can remember all information.

- Learning Scenario: Bella is fed with information about some objects. Later, it is checked if she is able to remember about such objects.

Introduction Scenario

In this scenario, Arthur begins with no knowledge about the person he is seeing on the webcam, thus, his first action should be to asking who it is, as illustrated in Figure 4.1. In the Figure 4.1 (a), Arthur is meeting a new person and, therefore, asks for this person's name with the statement "Hello stranger! May I know your name?". After the proper introduction (i.e., the person in question answers that his name is "Knob"), Arthur stores the name of this person on his memory. When they meet again (Figure 4.1 (b)), Arthur is able to remember both the face and the name of this person. Then, he proceeds with a cordial greeting ("Greetings Knob!"). Plus, Arthur tries to maintain the conversation asking questions about the person. In Figure 4.1 (b), it is possible to see Arthur asking the age of the person ("How old are you?"). When the answer is given, such information is also stored into the memory, increasing the information Arthur knows about this user.



Hello stranger!
May i know
your name?

Knob

Greetings
Knob!How old
are you?

Enter text...

(a) Arthur does not know this person.

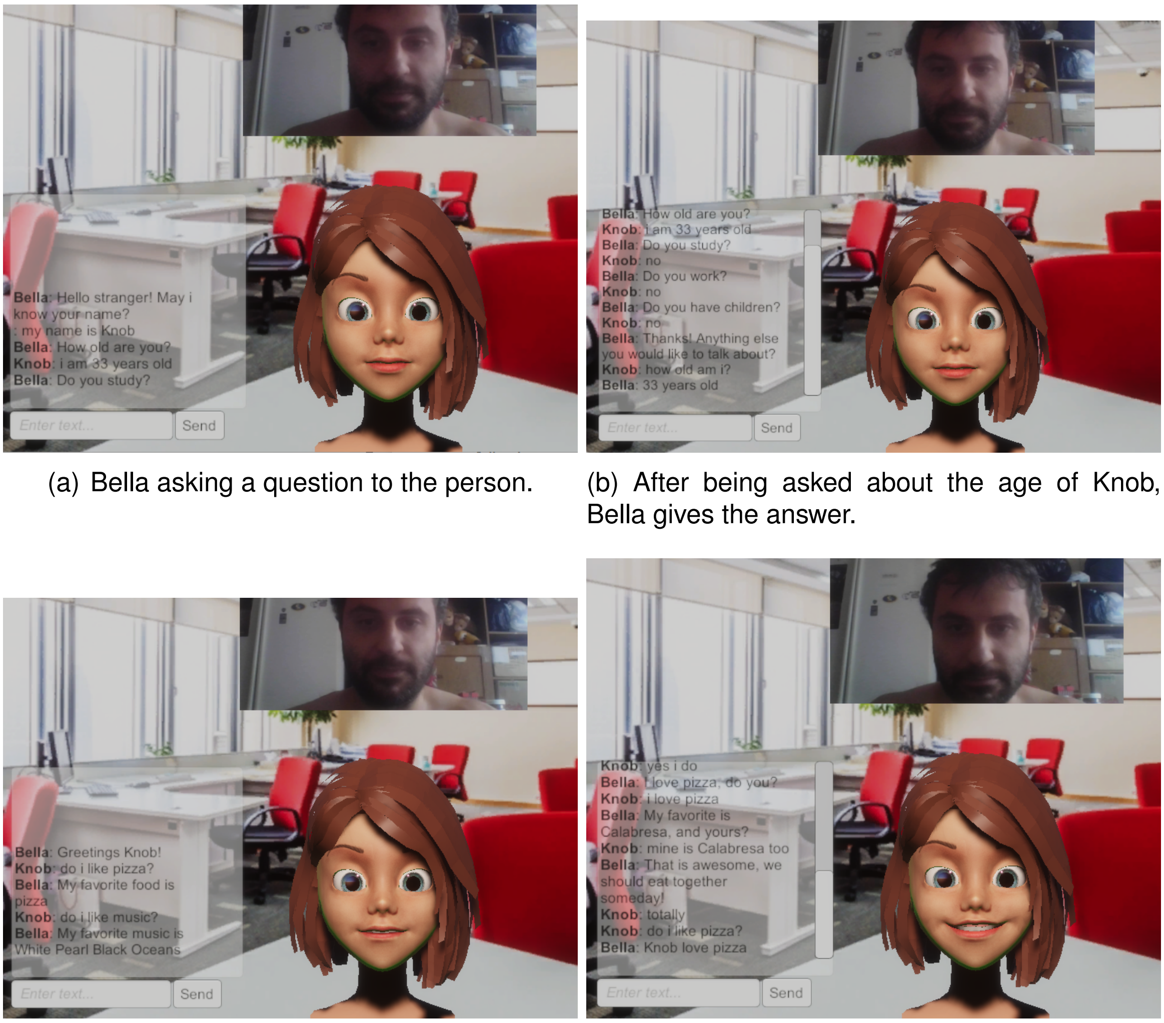
(b) After introduction, Arthur is able to remember this person.

Figure 4.1: Arthur meeting a new person. In (a), it asks for this person's name. After the proper introduction, Arthur stores the name of this person in its memory and is able to remember both the face and the name of this person, greeting it when they see each other again (b).

Remembering Scenario

As commented in Section 3.8, Arthur and Bella are able to learn information from the person he/she is talking with. Figure 4.2 shows Bella learning many information about Knob. As commented in Section 4.1.1, Arthur and Bella try to start or maintain a conversation with the person he/she is interacting, e.g., asking questions about him/her. Figure 4.2 (a) shows Bella asking if the person studies.

All possible questions are presented in Section 3.2. In Figure 4.2 (b), Bella answers the question about the age of Knob, saying he is 33 years old. She is able to do so because the person (i.e., Knob) already informed Bella about his age in a past interaction. In Figure 4.2 (c), Knob asks to Bella questions about his own preference (i.e., whether he likes pizza and music). Bella still do not have information about this in her memory, thus, she is not able to give these answers. In order to keep talking and stay in the topic, she answers about her own preference about the subject. Then, in Figure 4.2 (d), Knob tells Bella about his preference about pizza. When asked once again, she is able to properly answer the question.



Bella: Hello stranger! May i know your name?
: my name is Knob
Bella: How old are you?
Knob: i am 33 years old
Bella: Do you study?

(a) Bella asking a question to the person.

Bella: How old are you?
Knob: i am 33 years old
Bella: Do you study?
Knob: no
Bella: Do you work?
Knob: no
Bella: Do you have children?
Knob: no
Bella: Thanks! Anything else you would like to talk about?
Knob: how old am i?
Bella: 33 years old

(b) After being asked about the age of Knob, Bella gives the answer.

Bella: Greetings Knob!
Knob: do i like pizza?
Bella: My favorite food is pizza
Knob: do i like music?
Bella: My favorite music is White Pearl Black Oceans

(c) After being asked about some preferences of Knob, Bella does not know how to answer.

Knob: yes i do
Bella: I love pizza, do you?
Knob: i love pizza
Bella: My favorite is Calabresa, and yours?
Knob: mine is Calabresa too
Bella: That is awesome, we should eat together someday!
Knob: totally
Knob: do i like pizza?
Bella: Knob love pizza

(d) After Bella learns about some of Knob's preferences, she is able to answer properly.

Figure 4.2: Bella learning new things about the user.

4.1.2 Perceptive Study

In order to test our memory model with a qualitative evaluation, we conducted an experiment with subjects. Instead of interacting with the virtual agent, each person was presented to three video sequences. In the first video sequence, Arthur meets a new person and asks questions about he/she, as explained in Section 3.2, but the Memory Module is deactivated, thus, it does not retain information. In the second video sequence, the same interaction is conducted, but the Memory Module is activated, thus, Arthur can retain information and remember it further. Finally, in the third video sequence is specifically presented the process where Arthur learns a new information and remembers it further. Arthur is asked if he knows an object (i.e., cellphone), which he does not. Then, we teach him what a cellphone is and ask him again if he knows a cellphone. Since the Memory Module is activated, Arthur remembers what a cellphone is and answers the person. After each video sequence, the person is asked to answer some questions:

- Q1: How do you evaluate your comfort level regarding agent's appearance?
- Q2: How do you evaluate the agent's facial expressions realism level?
- Q3: How do you evaluate agent's information comprehension abilities?
- Q4: After watching the interaction, how do you evaluate agent's memory?

Subjects answered each question following a Likert scale, where:

- 1: Very low
- 2: Low
- 3: Moderate
- 4: Good
- 5: Very good

Regarding the expectations, the following hypothesis were formulated:

- H1: Since the facial expressions and the empathetic behavior are the same for all three videos, we expect little or no difference of how people will evaluate both their comfort level and the realism level of the agent's facial expressions (Q1 and Q2), throughout the video sequences.
- H2: We believe that the memory of Arthur is going to help him to understand the context of the interaction better. Therefore, in the videos where the Memory Module is activated, we believe the participants are going to rate Arthur's comprehension abilities better (Q3).
- H3: We expect that people rate Arthur's memory higher in Videos 2 and 3, when compared with Video 1 (Q4).

Concerning the subjects, 51 answers were registered, where 39 were male and 12 were female. Also, participants ranged from 17 to 61 years old. The average age was 31.49 with a standard deviation of 12.49. All participants agreed with the terms of the research and conducted the experiment until the end.

Table 4.1: Quantity of answers acquired by the experiment, separated by each video sequence. The head numbers (i.e., 1, 2, 3, 4 and 5) represent each value of the Likert scale.

Video 1	**1**	**2**	**3**	**4**	**5**
Q1	4	7	19	19	2
Q2	8	12	19	10	2
Q3	3	14	15	15	4
Q4	7	8	17	15	4
Video 2	**1**	**2**	**3**	**4**	**5**
Q1	6	9	15	19	2
Q2	5	12	20	12	2
Q3	4	4	14	23	6
Q4	5	4	11	22	9
Video 3	**1**	**2**	**3**	**4**	**5**
Q1	4	7	16	19	5
Q2	5	12	19	10	5
Q3	2	6	12	19	12
Q4	1	5	13	19	13

Table 4.1 shows the quantity of answers obtained per question, for each video sequence. In order to investigate the hypothesis variation of response of the users, we rely on the Kruskal test. The results reveal an uniformity in the participants answers concerning Q1 (p-value = 0.58) and Q2 (p-value = 0.59), thus, there is little

variation on the answers about the appearance and realism level of the agent (i.e., Q1 and Q2), which confirms H1. For Q1, most answers were between Moderate (3) and Good (4). For Q2, the most answers were Moderate (3). With this, we can also conclude that the appearance of Arthur has some space to improve, specially concerning the empathetic behavior and its facial expressions.

An unexpected behavior was observed on the answers about the comprehension abilities and memory of the agent (i.e., Q3 and Q4). Both questions had a significant amount of answers between Moderate (3) and Good (4) for Video 1. For example, if we take Q4, it was expected to have most answers lying between Very Low (1) and Low (2), since the agent has the Memory Module deactivated. However, we had 17 answers for Moderate (3) and 15 for Good (4), while having a total of 15 for the ones we were expecting (Very Low (1) and Low (2)). One possible explanation could be the order of the sentences that Arthur speaks. Since they are presented to the user in an orderly way, it may cause the impression that Arthur is indeed understanding and processing the information given by the user. Further investigation would be necessary to confirm or disprove it.

Even with such unexpected behavior, it seems that H2 and H3 were also confirmed. In Video 1, we had a total of 19 answers for Good (4) and Very Good (5), for both Q3 and Q4. In Video 2, with the Memory Module activated, we had a total of 29 answers, in the same range, for Q3 and 31 answers for Q4. In Video 3, we had 31 for Q3 and 32 for Q4. Also, if we look at the amount of answers Very Low (1) and Low (2), we can clearly perceive that it diminishes in Videos 2 and 3, when compared with Video 1. Finally, we conducted a Mann-Withney test, which suggested a change of perception (from Very Low (1) and Low (2) to Good (4) and Very Good (5)) when comparing Video 1 and Video 3, for both Q3 (p-value = 0.006) and Q4 (p-value = 0.001). With this, the results indicate that people felt a similar level of comfort and perceived a similar level of realism with the agent's facial expressions. On the other hand, the results indicate that the participants perceived that the agent had better comprehension abilities when the Memory Module was activated, also scoring higher values for Q4 (After watching the interaction, how do you evaluate agent's memory?) in Videos 2 and 3. Thus, we confirm H1, H2 and H3.

4.2 Empathy and Interplay with Memory

In order to evaluate the empathy of Arthur and Bella, we also conducted two different experiments. The first one involves Short-Term Interactions and Long-Term Interactions, where volunteers interacted with Arthur or Bella and answered a questionnaire about their impressions. The second one involves only Short-Term Interactions and it is focused on the relationship between Empathy and Memory present in Arthur and Bella.

4.2.1 Empathy Experiment

The main goal of this experiment is to evaluate the perception of people concerning the empathy of Arthur and Bella. This experiment was conducted with both LTIs and STIs. Concerning the LTIS, four persons, two men and two women, interacted daily with our virtual agent for ten days, resulting in a total of 40 answers for our survey, which has 7 questions. i.e., 280 answers. Each interaction lasted between 10-15 minutes. The details about each participant can be seen in Table 4.2. All participants read and agreed with the ethics term presented at the beginning of the questionnaire. Since empathy plays an important part, before starting the interactions, all participants were presented with a brief explanation about emotion and empathy. Moreover, we conducted the Toronto empathy questionnaire [SMML09] (TEQ) to measure the empathy level of the participants, also shown in Table 4.2. It is important to remember that the average score for men ranges from 43.46 to 44.45, while the average score for women ranges from 44.62 to 48.93, according to [SMML09].

Table 4.2: Participants of the LTI experiment.

Participant	**Age**	**Ed. Lvl.**	**Xp with ECAs**	TEQ
Man 1	27	Graduation	None	36
Man 2	21	Graduation	Low	48
Woman 1	27	Under-graduation	None	55
Woman 2	21	Under-graduation	Regular	62

Moreover, to conduct our STIs, eight persons, four men and four women, interacted with our virtual agents, only one time, and answered the same questionnaire. Each interaction also lasted between 10-15 minutes. The details about each participant can be seen in Table 4.3. All participants were also presented with a brief explanation about emotion and empathy and submitted to the Toronto empathy questionnaire [SMML09] (TEQ) to measure the empathy level of the participants, also shown in Table 4.2.

Table 4.3: Participants of the STI experiment.

Participant	**Age**	**Ed. Lvl.**	**Xp with ECAs**	TEQ
Man 1	21	Under-graduation	High	46
Man 2	23	Under-graduation	Low	52
Man 3	22	Under-graduation	High	60
Man 4	22	Regular	Low	37
Woman 1	20	Under-graduation	Low	45
Woman 2	25	Under-graduation	Regular	61
Woman 3	22	Under-graduation	Regular	53
Woman 4	21	Under-graduation	Regular	52

After each interaction, the participants answered a questionnaire compounded of:

- One question concerning Effectiveness, Efficiency and Satisfaction, as proposed by Santos et al. [dSKSM21].
- Adaptations of both Bartneck "Godspeed" questionnaire [BCK08] and Heerink questionnaire [HKEW09], in a total of two questions. Although both questionnaires were mainly used in the evaluation of robots, there are questions that can be adapted to virtual agent as well.
- Free text field, where the participant could freely write about his/her impressions about the ECA and the interaction.

Table 4.4: Questions asked to the participants of the experiment.

Question	**Likert Scores**
1) How do you evaluate your satisfaction with the agent's empathy?	[1;5]
2) I feel that the agent understands me.	[1;5]
3) Sometimes the agent seems to have real feelings.	[1;5]

Table 4.4 presents the questions made to the participants in the questionnaire. In order to conduct this evaluation, two hypothesis were raised:

- *H1*: We expect that the interactions with the empathetic agent are going to be more pleasant to the user than the interactions with the same agent without empathy.
- **H2**: We expect that the results achieved in the STIs are going to be represented by higher empathy values than the results achieved by the LTIs. Our hypothesis here is justified by the fact that the users may not perceive some issues (e.g., vocabulary, agent being unable to answer something, software errors, etc), with the STIs, that LTI users deal with due to the prolonged interaction time.

Next, we present the results achieved, both with Long-term Interactions and Short-term Interactions.

In order to test our hypothesis *H1* (We expect that the interactions with the empathetic agent are going to be more pleasant to the user than the interactions with the same agent without empathy), the volunteers conducted interactions with the virtual agent with and without empathy. To do so, Man 1 and Woman 1 interacted with the virtual agent with empathy, while Man 2 and Woman 2 did the same with the virtual agent without empathy. It is important to note that users did not know if the Empathy Module was activated on the agent or not.

Table 4.5: Average of the scores of the empathy assessment for the LTIs, referring to questions 1-3 in Table 4.4.

Participant	Empathy Module	Average
Man 1	True	2,9
Woman 1	True	2,2
Man 2	False	2,06
Woman 2	False	1,93

Figure 4.3 presents the scores of the four participants in the three evaluated questions, while Table 4.5 presents the average score of the three questions. It is possible to note in Figure 4.3 that the best scores were reached by Man 1 (3.3, 2.8 and 2.6), who uses the ECA with Empathy, while the worst scores were reached by the participants who had the Empathy Module deactivated (Man 2 for question 1 with 2.1, Woman 2 for questions 2 and 3 with 1.8 in both). Moreover, we explored the temporal evolution of the answers of the four participants during the ten days of interaction. Figures 4.4, 4.5 and 4.6 present this temporal evolution for questions 1-3 in Table 4.4, respectively. In Figure 4.4, concerning question 1, it is possible to

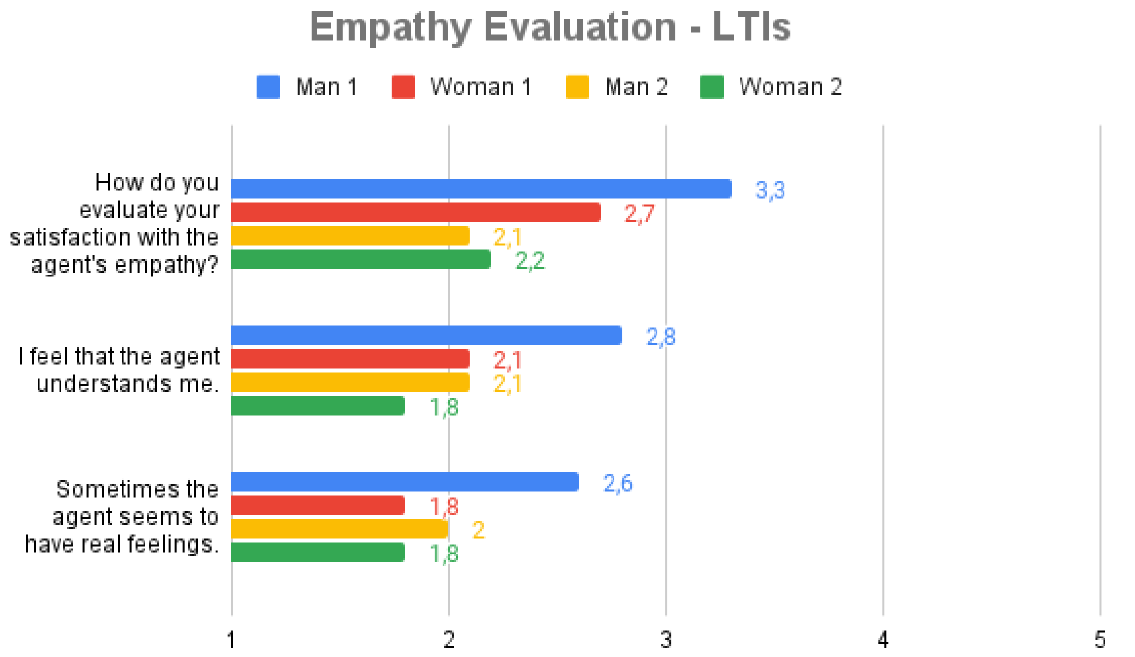


Figure 4.3: Scores of the empathy assessment for the LTIs, referring to questions 1-3 in Table 4.4. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5.

notice that Man 1 scored a 2 in his first day and alternated between 3 and 4 on the remaining days. Woman 1 presented a great variation of values, going from 1 to 5. Man 2 and Woman 2 presented a similar behavior: from day 3 onward, they both scored 2 until the end of the interaction. It seems to indicate that the group which had the Empathy Module activated (Man 1 and Woman 1) were more satisfied with the agent's empathy than the group which had the Empathy Module deactivated (Man 2 and Woman 2).

In Figure 4.5, concerning question 2, it is possible to notice that Man 1 scored a 3 in the first four days, alternating between 2 and 4 on the remaining days. Woman 1 presented a greater variation, starting with a 2 in her first day, 1 in her second day and passing through 3, 2 and 4 in the remaining days. Both Man 2 and Woman 2 varied from 1 to 3 in the ten days, never scoring 4 or 5. Although no interesting pattern could be perceived, it is interesting to note what happened on each day which caused a change of perception. For instance, we can see in Figure 4.5 that Woman 2 scored 1 in her first day of interaction, but raised her score to 3 in the second day. In the free text filed, she commented that "Arthur was funny today, he even made me laugh. He was also kinder and friendlier than yesterday.". It

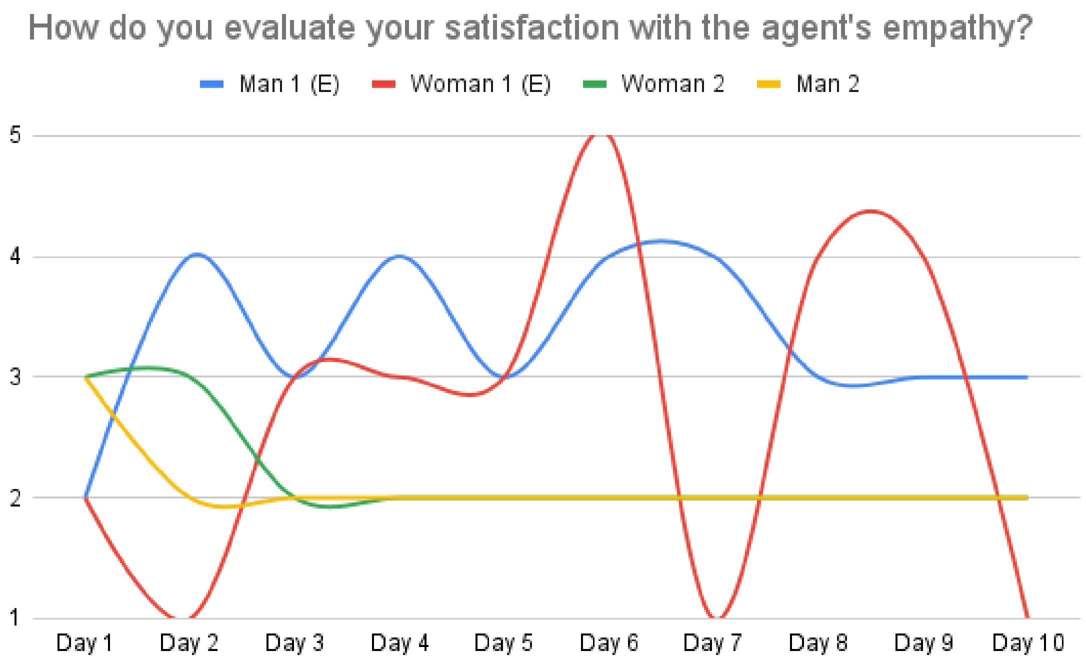


Figure 4.4: Temporal evolution of the results regarding question 1 in Table 4.4, for the ten days interaction and all four participants. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5.

is also possible to note that she dropped her score to 1 again in day 4, to which she commented that the agent was presenting some unexpected behavior, like mistaking her name.

In Figure 4.6, concerning question 3, it is possible to notice that Man 1 scored a 4 in the first day, 3 between days 2 and 5, and 2 in the remaining days. Woman 1 started with a 1, then scored 2 in the next two days and 3 in days 4 and 5. Then, she alternated between 1 and 2 in the remaining days. Man 2 scored a 3 twice (days 4 and 9), alternating between 1 and 2 in the remaining days, while Woman 2 scored a 3 only once (day 2), alternating between 1 and 2 in the remaining days. Again, although no interesting pattern could be perceived, it is interesting to note what happened on each day which caused a change of perception. For instance, we can see in Figure 4.6 that Man 2 scored 3 in his fourth day of interaction, but dropped his score to 2 in the fifth day and to 1 in the sixth day. In the free text filed, he commented that Bella was uttering several strange phrases and was mistaking his name. Also, he comments that Bella offered herself to be a calculator, "but didn't understand simple operations half the time".

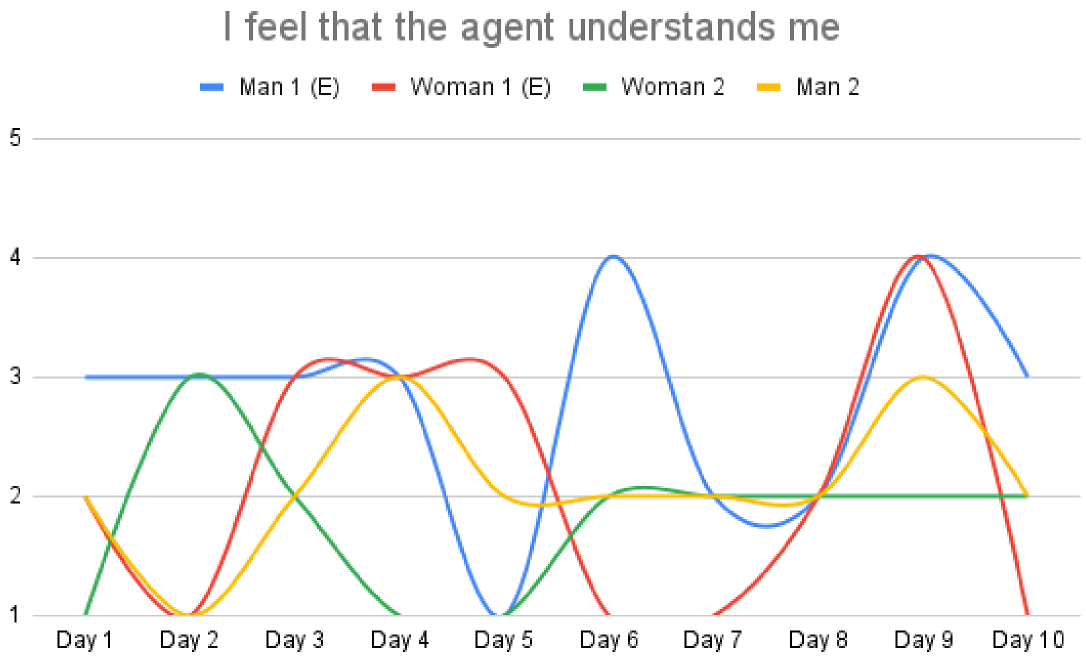


Figure 4.5: Temporal evolution of the results regarding question 2 in Table 4.4, for the ten days interaction and all four participants. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5.

Moving to the STIs, the eight volunteers conducted interactions, only once, with the ECAs with and without empathy. To do so, Man 1, Man 2, Woman 1 and Woman 2 interacted with the virtual agent with empathy, while the others interacted with the virtual agent without empathy. Figure 4.7 presents the average scores of the eight STI participants in the three evaluated questions. It is possible to note in Figure 4.7 that the results for question 1 (How do you evaluate your satisfaction with the agent's empathy) were very similar. Five of the participants scored 4, while the other three participants scored 5. Concerning question 2 (I feel that the agent understands me), half of the participants scored 4, while the worst score was achieved by Man 4 (1) and the best score was achieved by Woman 3 (5). For question 3 (Sometimes the agent seems to have real feelings), most of the participants scored 3 and 4, with Man 1 scoring 2.

In addition, we can calculate the average score of each group (with and without Empathy Module). Concerning the group with Empathy Module activated, the average scores were 4.5, 3.75 and 3.25 for questions 1, 2 and 3, respectively. Concerning the group with Empathy Module deactivated, the average scores were 4.25, 3 and 3.5 for questions 1, 2 and 3, respectively. Although the averages seem,

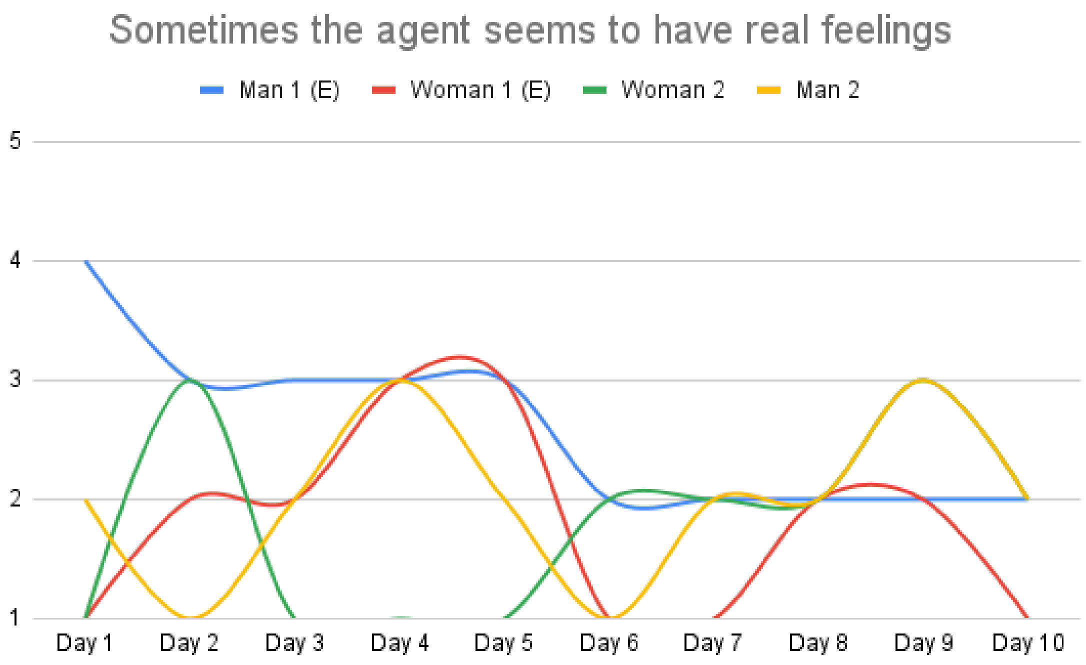


Figure 4.6: Temporal evolution of the results regarding question 3 in Table 4.4, for the ten days interaction and all four participants. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5.

in general, a bit higher for the Empathy group, it seems to have little impact for the STIs. For instance, for question 2, the worst score was achieved by Man 4 (1) and the best score was achieved by Woman 3 (5), and both interacted with the virtual agent without Empathy. We performed a Mann-Whitney test using the values presented in Figure 4.7 and grouped by each group (with and without Empathy Module), resulting in a p-value of 0.70, which indicates a similarity between the answers of the participants of both groups. We hypothesized that the short-term interactions of 10-15 minutes did not offer a conversation where the ECA could apply the empathy model or, at least, be fully perceived. Therefore, these results seem to indicate that *H1* is valid when the interaction contemplates scenarios where empathy can be perceived, i.e., in the LTIs. We argue that in LTIs scenarios, the participants have more time interacting with the virtual agent and, thus, can better perceive the empathy conveyed by Arthur or Bella than participants which interact only once (i.e., STIs).

Concerning **H2**, while the obtained average scores in LTIs (2.55 and 1.99, respectively for ECA with and without empathy) are lower than STIs (3.83 and 3.58), confirming this hypothesis, another aspect can be noted: the percentage difference between the two LTI groups (with or without empathy) is 21.96%, i.e., greater than

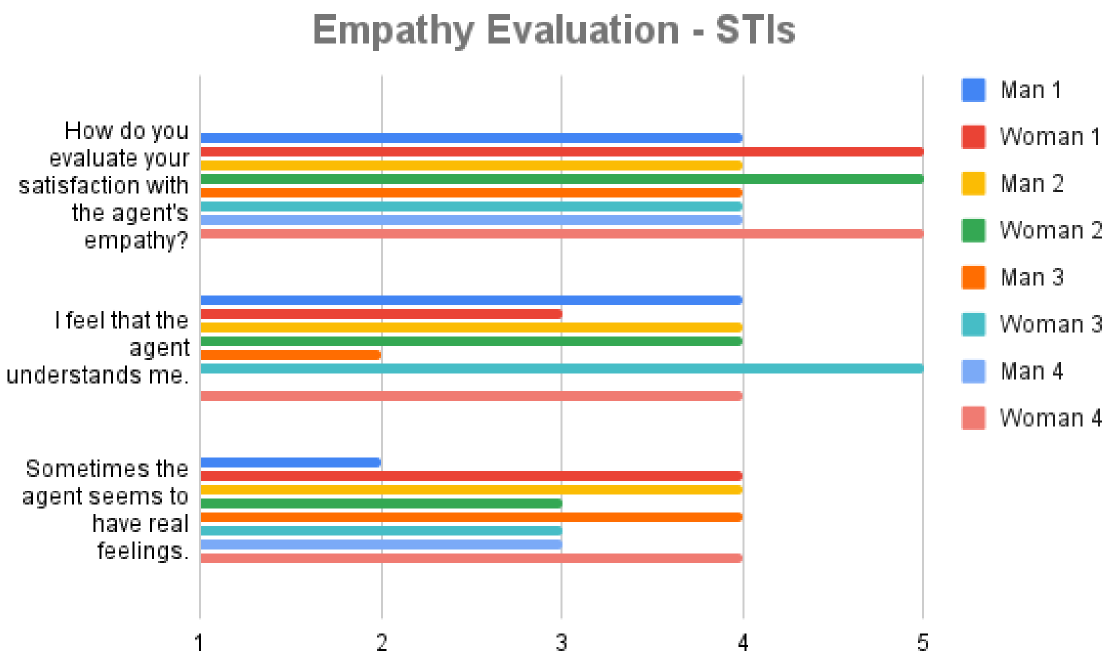


Figure 4.7: Average scores of the empathy assessment for the STIs, referring to questions 1-3 in Table 4.4. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5.

the difference between the two STI groups (6.52%). Here, we hypothesize that the reason for this difference is that in STIs the content of the conversation between agent and participants were more neutral, and performed in a single interaction. On the other hand, the LTI participants had probably more emotional conversations, accessing the Arthur/Bella emotional module. There is another possibility: the lower scores observed in the LTIs could have been caused by lower engagement, when compared with STIs. Since the LTIs occurred for 10 days, the engagement of the users can have been reduced after each day of interaction, which would cause a drop in the evaluations. This phenomenon would not occur in the STIs, because users only interacted once with the virtual agent.

4.2.2 Empathetic Memory experiment

This experiment was conducted with only Short-term Interactions (STIs). Participants were recruited to interact with Arthur or Bella and answer an online questionnaire, summing up 30 people (22 Men and 8 Women). Of these 30 volunteers, 13 are Undergraduates, another 13 are Graduated, 3 completed High School and 1 person is a high school student. Concerning their past experience interacting with virtual agents, 6 participants answered as Very Low, 9 as Low, 9 as being Regular, 4 as being High and 2 as being Very High. The average age of the participants was 27.43, with a standard deviation of 11.84. Each participant was asked to accomplish a set of tasks to complete, as presented in Table 4.6. Before starting these tasks, all participants read and agreed with the ethics term presented at the beginning of the questionnaire. After that, they were encouraged to download the ECA's executable file and freely interact with it for a short while to get used to it. They were also presented with a brief explanation about emotion and empathy and answered the Toronto empathy questionnaire [SMML09] (TEQ) to measure their empathy level. The mean score computed for men was 43.63, while the mean score computed by for women was 47.12. All tasks presented in Table 4.6 are related to some data that is present in the agent's memory, with a respective emotion associated. In T1, T2 and T3 participants should find some information saved in ECA's memory and recognize the expressed ECA's facial emotion. For T4, participants were asked to freely ask about the subject they want. Tasks T1 and T3 are about the ECA's self-memory, while T2 and T4 are related to what the agent knows about the participant. Following one of the definitions of empathy cited by de Wall [De 08] (the ability to understand and react towards the emotion of others), we believe that such emotional memories can be seen as an empathetic behavior, being able even to trigger such behavior in the participants. Finally, all tasks ask the participants to evaluate the agent's empathy on a Likert scale from 1 (no Empathy) to 5 (Extremely Empathetic). In order to conduct the evaluation, we raise one main hypothesis: *H3:* We expect that participants can trigger ECA's memories and identify the associated emotion.

Table 4.6: Tasks of the empathetic memory experiment.

Task	**Description**	**Emotion**
T1	Discover if the virtual agent likes video games and if it has a favorite game.	Happiness
T2	Discover if the virtual agent remembers about the participant's study and work.	Happiness
T3	Discover if the virtual agent has any pets, as well as more information about it.	Sadness
T4	Discover if the virtual agent remembers about any other subject that the participant already spoke with it.	Varied

Figure 4.8 presents the scores of the thirty participants from the experiment. "Correct Answer", in blue, refers to the amount of people who answered as

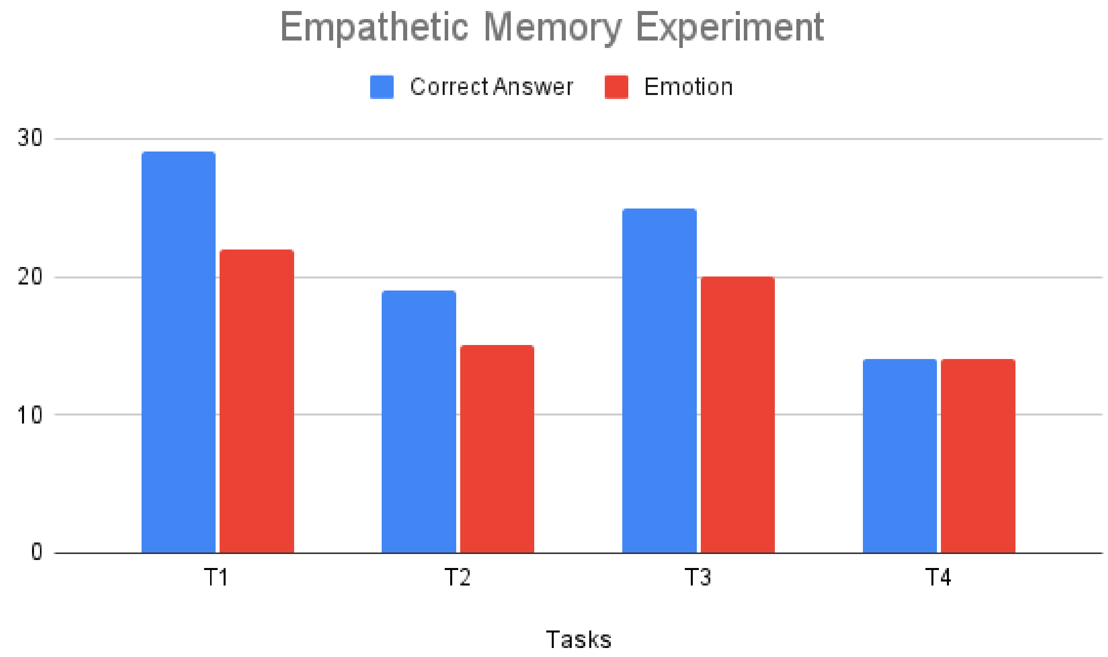


Figure 4.8: Scores of the thirty participants from the experiment. "Correct Answer", in blue, refers to the amount of people who answered as expected. "Emotion", in red, refers to the amount of people who correctly identified the agent's conveyed emotion.

expected. For instance, in **T1** (Table 4.6), it was expected that the participants were able to discover that the virtual agent enjoys playing video games. "Emotion", in red, refers to the amount of people who correctly identified the agent's conveyed emotion. Concerning **T1** (Find out if the ECA likes video games and if he/she has a favorite game), from 30, 29 participants were able to find out that the virtual agent likes video games. Also, 23 participants were able to identify the agent's favorite game, while 22 participants correctly identified the emotion conveyed by Arthur or Bella (i.e., Happiness). Concerning **T2** (Find out if the virtual agent remembers about the participant's study and work), from 30, 19 participants reported that the ECA was able to remember information about their study/work, and 15 of them correctly identify the emotion conveyed (i.e., Happiness). Regarding **T3** (Discover if the ECA has any pets, as well as more information about it), 25 of 30 participants were able to answer that the virtual agent had a pet, and 24 were also able to identify the pet's name. Moreover, 20 participants could correctly identify the emotion conveyed by Arthur or Bella (i.e., Sadness). Concerning **T4** (Find out if the ECA remembers about any other subject that the participant already spoke with it), 14 participants reported that Arthur or Bella was able to remember about some other subject that they chose to speak about and conveyed an appropriate emotion.

The results presented suggest that the participants were, in general, able to trigger the expected memories from Arthur or Bella and correctly identify the emotion associated with it, thus validating *H3*. It is also possible to notice that the worst results were found when the 30 participants had to retrieve a memory about him/herself (19 participants answered correctly in **T2** and 14 in **T4**), when compared with memories about the agent itself, i.e., 29 participants correctly answer about video games in **T1**, and 25 concerning pets in **T3**. In this case, we hypothesize that **T1** and **T3** are more straight-forward tasks than **T2** and **T4**. Figure 4.9 presents the average scores and standard deviations of the evaluated empathy for all the four tasks. As commented before, the evaluated empathy is a Likert scale which goes from 1 (no Empathy) to 5 (Extremely Empathetic). The average score values were 3.71, 3.23, 3.71 and 3.38, with standard deviation of 0.90, 1.18, 0.9 and 1.02, for tasks **T1-T4**, respectively. It is possible to notice that the best scores (3.71) were achieved in **T1** and **T3**, which are the tasks where the participants should find out something about Arthur or Bella, and more direct tasks, as discussed before.

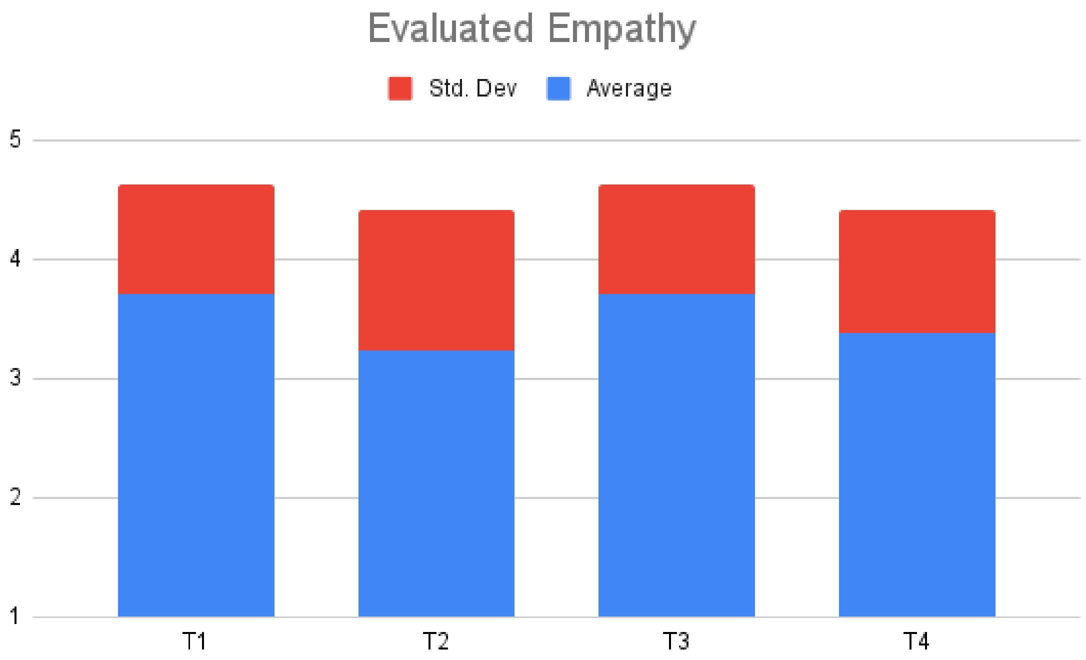


Figure 4.9: Evaluated empathy for all four tasks. The average score values were 3.71, 3.23, 3.71 and 3.38 for tasks **T1-T4**, respectively. The standard deviation values were computed as being 0.90, 1.18, 0.9 and 1.02 for tasks **T1-T4**, respectively.
]

4.3 Chapter Considerations

This chapter presented and discussed the results achieved by this work. Such results include: the behavior of the agent's memory 4.1 and experiments concerning the empathy of the agent and its relationship with memory 4.2. Concerning the memory model, we were able to confirm it is working as it was intended. Additionally, the results achieved suggest that the memory model helped the participants to understand the context of the interaction better and were more satisfied with the interactions. Concerning the empathy model, the results achieved suggest that the participants were more satisfied interacting with the empathetic agent, when compared with the virtual agent without empathy. Also, they were able to perform the tasks of the Empathetic Memory experiment 4.2.2, suggesting that the empathy of Arthur/Bella could be perceived in the memories retrieved. The next Chapter presents the final considerations.

5. FINAL REMARKS

This work presented a model of an empathetic Embodied Conversational Agent (ECA) endowed with many abilities, like face recognition, emotion detection, expressiveness, empathy and memory modeling. The main contribution of this work lies on the memory model, the empathy model and on the interplay between them. Some experiments were conducted in order to test the proposed model and collect both quantitative and qualitative information. The results achieved seem to confirm that Arthur/Bella presented the expected behavior. Also, concerning the ethic of our Embodied Conversational Agent, it is important to note two things: our model and the online chatbot model. While the conversation is inside our model (e.g., ice-breakers, small talks, etc.), everything that Arthur/Bella says is controlled, since it is manually defined. In fact, if needed, we can deactivate the chatbot responses and work with only our model. However, since we also use the chatbot model for when Arthur/Bella is unsure about how to answer, answers delivered to the user may be "unethical" or weird. It is important to emphasize that it may not be the case: the online API used in the chosen chatbot may be ethical and friendly: we just do not have control over it. If needed, indeed, we could choose another chatbot model to use.

Concerning the memory of the agent, two experiments were conducted: one based on scripted tasks and one experiment with volunteers. In the scripted tasks experiment, we modeled two different scenarios: one for introduction (i.e., Arthur/Bella meets someone new) and other for remembering (i.e., Arthur/Bella learns information with the user and tries to remember about it later). In the introduction scenario, Arthur was able to meet someone new, store this information in his memory and remember about this new person in further interactions. In the remembering scenario, Bella is able to learn a few things about the user, such as his/her age and pizza preference, being able to retrieve such information from her memory and answers questions about the subject. In the experiment with volunteers (i.e., Perceptive Study), 51 participants had to watch three video sequences and answer a questionnaire about these videos. Among other things, we expected that the participants would better evaluate Arthur/Bella's comprehension abilities when the Memory Module was activated, which seems to have been validated by the results. In summary, the results achieved in both experiments suggest that the memory worked as intended and its influence on the interaction could be perceived.

Concerning the empathy of the agent, another two experiments were conducted: one focused on the empathy itself and other focused on its relationship with memory. In the Empathy experiment, the results achieved suggest that the Empathy module was perceived by people and impacted in their interactions, but it could be better perceived in Long-Term Interactions (LTIs) when compared with Short-Term Interactions (STIs). We hypothesize that LTIs give more time for people to interact with the virtual agent and so, the user's perception is affected by much more exposition to problems and other characteristics, if compared with the restricted time of STIs. In the Empathetic Memory experiment, 30 participants were endowed with four tasks to be done while interacting with Arthur or Bella, which were related with some piece of information present in the agent's memory. It was expected that the participants would be able to make Arthur or Bella retrieve those memories and identify the associated emotion. The results achieved suggest that, indeed, the participants were able to trigger the correct memories and correctly identify the emotion conveyed. In addition, the participants evaluated the empathy of Arthur or Bella, as perceived by them, for each task, for which average scores between 3 and 4 (in a scale from 1 to 5) were computed.

This work has some limitations. Firstly, the number of users is certainly an issue that we want to work in a future, specially when we look at the first empathy experiment in Section 4.2.1. However, in this case, we argue that LTIs are much more interesting to be used to evaluate ECAs than STIs. The reason is, as mentioned before, with a little time of interacting with an ECA, the user can not explore all possibilities in the possible dialogues. Nevertheless, having more participants is going to allow us to explore other hypotheses, such as the perception of people concerning Arthur and Bella. Another limitation is the absence of a chatbot model. Everything that the user says is analyzed by NLTK, which is able to conduct natural language operations (see Section 3.8.2). What is said by Arthur/Bella depends on what the user is talking, but when Arthur/Bella is unsure about how to answer, he/she still needs to have a "escape path" to use, in order to keep the conversation alive (after all, Arthur/Bella should act as a friend). In order to do so, we send the sentence said by the user to an online chatbot API, which is able to deliver an answer back to us. But, this way, we become dependent of the availability of such API: if it becomes unavailable for some reason, Arthur/Bella loses their "escape path". In addition, the chatbot API works in a different system: it means that all the information stored in Arthut/Bella's memory is not used in the construction of the answer, which can lead to weird sentences. Indeed, many volunteers who participated in the experiments commented that Arthur/Bella was calling them by different names, like "Aco" or "Junior".

For future work, there are many avenues to follow. For instance, we want to invest more time in the visual behavior and facial animation of Arthur and Bella. Besides the modeling of different emotions, we would like to make this experience more personal, as it would be with a friend. In this topic, Melgare et al. [MMSQ19] suggested the existence of emotion styles, where each person would have their own way to demonstrate an emotion. In this sense, one interesting future work would be to endow Arthur/Bella with the ability to identify such style on the face of the user and mimic it. This way, we believe that the user would feel more comfortable with the facial expressions of Arthur/Bella.

Another interesting avenue to explore is the dialog system. Besides building and integrating a generic chatbot model, we would like to take a deeper look into small talks. As it was commented in Section 3.3, our Embodied Conversational Agent is endowed with a small talk module, allowing it to start a conversation with the user about a pre-defined topic. The main problem is that such topics and dialogues need to be manually defined: each dialog tree needs to be manually written and added to the virtual agent. One possible future work would be to find a way to automatize this process. For instance, it would be interesting to investigate if is possible to use previous interactions as material to create new dialog trees. If so, it would be possible to build a script which can automatically gather previous interactions between Arthur/Bella and the users and build new topics and dialogues for the small talk module.

Since Arthur/Bella aims to be a friend to the user, yet another avenue for future work is the befriending process and the process of relationship formation. According Levinger [Lev80], such process occurs in stages and follows an ABCDE sequence of relationship development, where A stands for Attraction, B for Build-up, C for Continuation, D for Decline and E for Ending. Such process could be incorporated into Arthur/Bella, especially for Long-Term Interactions. Finally, other avenues that could be explored are: sound information, where sound files could be stored in the autobiographical memory of Arthur/Bella (the main problem would be the size of these files); body expressions, in which we could model bodies (or parts of the body) and use them to express emotions and non-verbal behavior; object recognition, where a trained neural network could be applied to recognize objects in the webcam (e.g., a glass of water, a dog, a cat, etc.) and pass it to Arthur/Bella, who would be able to search his/her memory for information about these objects and talk about this subject.

References

[AC76] Argyle, M.; Cook, M. “Gaze and mutual gaze.”, 1976. (Citado na página 25.)

[ACP+18] Almeida, A. M.; Cerri, R.; Paraiso, E. C.; Mantovani, R. G.; Junior, S. B. “Applying multi-label techniques in emotion identification of short texts”, *Neurocomputing*, vol. 320, 2018, pp. 35–46. (Citado na página 36.)

[AHS19] Ayedoun, E.; Hayashi, Y.; Seta, K. “Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners’ willingness to communicate”, *International Journal of Artificial Intelligence in Education*, vol. 29–1, 2019, pp. 29–57. (Citado nas páginas 26, 27, and 48.)

[Amb14] Ambridge, B. “Psy-Q: You know your IQ-now test your psychological intelligence”. Profile Books, 2014. (Citado na página 76.)

[And05] Anderson-Cook, C. M. “Practical genetic algorithms”, 2005. (Citado na página 45.)

[Bad92] Baddeley, A. “Working memory”, *Science*, vol. 255–5044, 1992, pp. 556–559. (Citado na página 39.)

[BCK08] Bartneck, C.; Croft, E.; Kulic, D. “Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots”, 2008. (Citado na página 90.)

[BD19] Bamford, J. M. S.; Davidson, J. W. “Trait empathy associated with agreeableness and rhythmic entrainment in a spontaneous movement to music task: Preliminary exploratory investigations”, *Musicae Scientiae*, vol. 23–1, 2019, pp. 5–24. (Citado na página 76.)

[BDR+21] Beinema, T.; Davison, D.; Reidsma, D.; Banos, O.; Bruijnes, M.; Donval, B.; Valero, Á. F.; Heylen, D.; Hofs, D.; Huizing, G.; et al.. “Agents united: An open platform for multi-agent conversational systems”. In: 21st ACM International Conference on Intelligent Virtual Agents, 2021. (Citado na página 47.)

[Ber96] Berntsen, D. “Involuntary autobiographical memories”, *Applied Cognitive Psychology*, vol. 10–5, 1996, pp. 435–454. (Citado na página 67.)

[BKW04] Becker, C.; Kopp, S.; Wachsmuth, I. “Simulating the emotion dynamics of a multimodal conversational agent”. In: Tutorial and Research Workshop on Affective Dialogue Systems, 2004, pp. 154–165. (Citado na página 77.)

[BL98] Bluck, S.; Levine, L. J. “Reminiscence as autobiographical memory: A catalyst for reminiscence theory development”, *Ageing & Society*, vol. 18–2, 1998, pp. 185–208. (Citado nas páginas 38 and 62.)

[BVDHMAM07] Bartneck, C.; Van Der Hoek, M.; Mubin, O.; Al Mahmud, A. ““daisy, daisy, give me your answer do!” switching off a robot”. In: 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2007, pp. 217–222. (Citado na página 76.)

[BWM^{+}19] Biancardi, B.; Wang, C.; Mancini, M.; Cafaro, A.; Chanel, G.; Pelachaud, C. “A computational model for managing impressions of an embodied conversational agent in real-time”. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 2019, pp. 1–7. (Citado nas páginas 15, 26, 27, 42, 43, and 48.)

[CA21] Croes, E. A.; Antheunis, M. L. “Can we be friends with mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot”, *Journal of Social and Personal Relationships*, vol. 38–1, 2021, pp. 279–300. (Citado na página 35.)

[CC14] Collier, G.; Collier, G. J. “Emotional expression”. Psychology Press, 2014. (Citado na página 25.)

[CG11] Coplan, A.; Goldie, P. “Empathy: Philosophical and psychological perspectives”. Oxford University Press, 2011. (Citado na página 42.)

[CGRFK20] Castle-Green, T.; Reeves, S.; Fischer, J. E.; Koleva, B. “Decision trees as sociotechnical objects in chatbot design”. In: Proceedings of the 2nd Conference on Conversational User Interfaces, 2020, pp. 1–3. (Citado na página 55.)

[CPB$^+$94] Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Becket, T.; Douville, B.; Prevost, S.; Stone, M. "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents". In: Proceedings of the 21st annual conference on Computer graphics and interactive techniques, 1994, pp. 413–420. (Citado na página 25.)

[CPP00] Conway, M. A.; Pleydell-Pearce, C. W. "The construction of autobiographical memories in the self-memory system.", *Psychological review*, vol. 107–2, 2000, pp. 261. (Citado nas páginas 17, 38, 39, 62, 63, and 67.)

[CRP19] Cafaro, A.; Ravenet, B.; Pelachaud, C. "Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions", *IEEE Transactions on Affective Computing*, 2019. (Citado nas páginas 27 and 45.)

[Csá] Csáky, R. "Deep learning based chatbot models (2017)", *DOI*, vol. 10, pp. 13140. (Citado na página 55.)

[CSCP00] Cassell, J.; Sullivan, J.; Churchill, E.; Prevost, S. "Embodied conversational agents". MIT press, 2000. (Citado na página 25.)

[CW19] Chetty, G.; White, M. "Embodied conversational agents and interactive virtual humans for training simulators". In: Proc. The 15th International Conference on Auditory-Visual Speech Processing, 2019, pp. 73–77. (Citado nas páginas 26, 27, and 48.)

[DBODAH19] Das, K. S. J.; Beinema, T.; Op Den Akker, H.; Hermens, H. "Generation of multi-party dialogues among embodied conversational agents to promote active living and healthy diet for subjects suffering from type 2 diabetes". In: 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AWE 2019, 2019, pp. 297–304. (Citado nas páginas 26, 27, and 48.)

[De 08] De Waal, F. B. “Putting the altruism back into altruism: the evolution of empathy”, *Annu. Rev. Psychol.*, vol. 59, 2008, pp. 279–300. (Citado na página 97.)

[DFH⁺12] Dill, V.; Flach, L. M.; Hocevar, R.; Lykawka, C.; Musse, S. R.; Pinho, M. S. “Evaluation of the uncanny valley in cg characters”. In: International Conference on Intelligent Virtual Agents, 2012, pp. 511–513. (Citado nas páginas 41 and 48.)

[DP19] Dermouche, S.; Pelachaud, C. “Generative model of agent’s behaviors in human-agent interaction”. In: 2019 International Conference on Multimodal Interaction, 2019, pp. 375–384. (Citado na página 45.)

[dSKSM21] dos Santos, J. B. S.; Knob, P. R.; Scherer, V. P.; Musse, S. R. “Is my agent good enough? evaluating embodied conversational agents with long and short-term interactions”. 2110.00114, 2021. (Citado na página 90.)

[dWP17] de Waal, F. B.; Preston, S. D. “Mammalian empathy: behavioural manifestations and neural basis”, *Nature Reviews Neuroscience*, vol. 18–8, 2017, pp. 498–509. (Citado nas páginas 25 and 42.)

[Ekm92] Ekman, P. “An argument for basic emotions”, *Cognition & emotion*, vol. 6–3-4, 1992, pp. 169–200. (Citado na página 70.)

[EMJ18] Edirisinghe, M.; Muthugala, M.; Jayasekara, A. “Application of robot autobiographical memory in long-term human-robot social interactions”. In: 2018 2nd International Conference On Electrical Engineering (EECon), 2018, pp. 138–143. (Citado nas páginas 27 and 39.)

[FDNP15] Ferreira, L.; Dosciatti, M.; Nievola, J.; Paraiso, E. C. “Using a genetic algorithm approach to study the impact of imbalanced corpora in sentiment analysis”. In: The Twenty-Eighth International Flairs Conference, 2015. (Citado na página 36.)

[GM85] Goldstein, A. P.; Michaels, G. Y. “Empathy: Development, training, and consequences”. Lawrence Erlbaum, 1985. (Citado nas páginas 72 and 77.)

[Gol90] Goldberg, L. R. “An alternative" description of personality": the big-five factor structure.”, *Journal of personality and social psychology*, vol. 59–6, 1990, pp. 1216. (Citado nas páginas 73 and 74.)

[HBMW19] Hancock, B.; Bordes, A.; Mazare, P.-E.; Weston, J. “Learning from dialogue after deployment: Feed yourself, chatbot!”, *arXiv preprint arXiv:1901.05415*, 2019. (Citado nas páginas 15, 34, and 35.)

[HKEW09] Heerink, M.; Krose, B.; Evers, V.; Wielinga, B. “Measuring acceptance of an assistive social robot: a suggested toolkit”. In: RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication, 2009, pp. 528–533. (Citado na página 90.)

[Hoj07] Hojat, M. “Empathy in patient care: antecedents, development, measurement, and outcomes”. Springer Science & Business Media, 2007. (Citado nas páginas 73, 77, and 79.)

[JBD$^+$18] Jonell, P.; Bystedt, M.; Dogan, F. I.; Fallgren, P.; Ivarsson, J.; Slukova, M.; Wennberg, U.; Lopes, J.; Boye, J.; Skantze, G. “Fantom: A crowdsourced social chatbot using an evolving dialog graph”, *Proc. Alexa Prize*, 2018. (Citado na página 55.)

[KMT12] Kasap, Z.; Magnenat-Thalmann, N. “Building long-term relationships with virtual and robotic characters: the role of remembering”, *The Visual Computer*, vol. 28–1, 2012, pp. 87–97. (Citado na página 39.)

[KNB19] Klinzing, J. G.; Niethard, N.; Born, J. “Mechanisms of systems memory consolidation during sleep”, *Nature neuroscience*, vol. 22–10, 2019, pp. 1598–1610. (Citado na página 69.)

[KRK13] Kope, A.; Rose, C.; Katchabaw, M. “Modeling autobiographical memory for believable agents”. In: Ninth Artificial Intelligence and Interactive Digital Entertainment Conference, 2013. (Citado nas páginas 27 and 41.)

[Ksh02] Kshirsagar, S. “A multilayer personality model”. In: Proceedings of the 2nd international symposium on Smart graphics, 2002, pp. 107–115. (Citado na página 73.)

[KtSM+19] Kramer, L.; ter Stal, S.; Mulder, B.; de Vet, E.; van Velsen, L. “Developing embodied conversational agents for healthy lifestyles: a scoping review”. In: ARPH, 2019. (Citado nas páginas 26, 27, and 48.)

[LBB02] Lee, S. P.; Badler, J. B.; Badler, N. I. “Eyes alive”. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, 2002, pp. 637–644. (Citado nas páginas 25, 41, 48, and 70.)

[Lev80] Levinger, G. “Toward the analysis of close relationships”, *Journal of experimental social psychology*, vol. 16–6, 1980, pp. 510–544. (Citado nas páginas 36 and 103.)

[LGB+16] Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; Dolan, B. “A persona-based neural conversation model”, *arXiv preprint arXiv:1603.06155*, 2016. (Citado na página 34.)

[LK16] Lison, P.; Kennington, C. “Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules”. In: Proceedings of ACL-2016 system demonstrations, 2016, pp. 67–72. (Citado na página 47.)

[LL19] Loftus, G. R.; Loftus, E. F. “Human memory: The processing of information”. Psychology Press, 2019. (Citado nas páginas 15, 25, 26, 37, 38, 49, 65, and 66.)

[LPSP15] Lowe, R.; Pow, N.; Serban, I.; Pineau, J. “The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems”, *arXiv preprint arXiv:1506.08909*, 2015. (Citado na página 33.)

[LZ15] Leigh, R. J.; Zee, D. S. “The neurology of eye movements”. OUP USA, 2015. (Citado nas páginas 25, 41, and 70.)

[MB03] Milward, D.; Beveridge, M. “Ontology-based dialogue systems”. In: Proc. 3rd Workshop on Knowledge and reasoning in practical dialogue systems (IJCAI03), 2003, pp. 9–18. (Citado na página 55.)

[MCM05] McCrae, R. R.; Costa, Jr, P. T.; Martin, T. A. “The neo–pi–3: A more readable revised neo personality inventory”, *Journal of personality assessment*, vol. 84–3, 2005, pp. 261–270. (Citado na página 75.)

[ME72] Mehrabian, A.; Epstein, N. “A measure of emotional empathy.”, *Journal of personality*, 1972. (Citado na página 78.)

[Meh80] Mehrabian, A. “Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies”, 1980. (Citado na página 75.)

[Mil56] Miller, G. A. “The magical number seven, plus or minus two: Some limits on our capacity for processing information.”, *Psychological review*, vol. 63–2, 1956, pp. 81. (Citado na página 66.)

[Mil95] Miller, G. A. “Wordnet: a lexical database for english”, *Communications of the ACM*, vol. 38–11, 1995, pp. 39–41. (Citado na página 61.)

[Min91] Minsky, M. “Society of mind: a response to four reviews”, *Artificial Intelligence*, vol. 48–3, 1991, pp. 371–396. (Citado na página 25.)

[MK20] Martinez, V. R.; Kennedy, J. “A multiparty chat-based dialogue system with concurrent conversation tracking and memory”. In: Proceedings of the 2nd Conference on Conversational User Interfaces, 2020, pp. 1–9. (Citado na página 40.)

[ML19] Mathur, S.; Lopez, D. “A scaled-down neural conversational model for chatbots”, *Concurrency and Computation: Practice and Experience*, vol. 31–10, 2019, pp. e4761. (Citado na página 31.)

[MM07] Moreno, R.; Mayer, R. “Interactive multimodal learning environments”, *Educational psychology review*, vol. 19–3, 2007, pp. 309–326. (Citado na página 76.)

[MMMR$^+$19] Martínez-Miranda, J.; Martínez, A.; Ramos, R.; Aguilar, H.; Jiménez, L.; Arias, H.; Rosales, G.; Valencia, E. “Assessment of users’ acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour”, *Journal of medical systems*, vol. 43–8, 2019, pp. 246. (Citado nas páginas 26 and 48.)

[MMSQ19] Melgare, J. K.; Musse, S. R.; Schneider, N. R.; Queiroz, R. B. “Investigating emotion style in human faces and avatars”. In: 2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames), 2019, pp. 115–124. (Citado nas páginas 70 and 103.)

[Mor05] Morville, P. “Experience design unplugged”. In: ACM SIGGRAPH 2005 Web Program, 2005, pp. 10–es. (Citado na página 54.)

[ODMP$^+$17] Ochs, M.; De Montcheuil, G.; Pergandi, J.-m.; Saubesty, J.; Pelachaud, C.; Mestre, D.; Blache, P. “An architecture of virtual patient simulation platform to train doctors to break bad news”. In: Conference on Computer Animation and Social Agents (CASA), 2017. (Citado nas páginas 15, 46, and 47.)

[Omd14] Omdahl, B. L. “Cognitive appraisal, emotion, and empathy”. Psychology Press, 2014. (Citado na página 25.)

[PDA$^+$20] Philip, P.; Dupuy, L.; Auriacombe, M.; Serre, F.; de Sevin, E.; Sauteraud, A.; Micoulaud-Franchi, J.-A. “Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients”, *npj Digital Medicine*, vol. 3–1, 2020, pp. 1–7. (Citado nas páginas 26 and 48.)

[PFD15] Petit, M.; Fischer, T.; Demiris, Y. “Lifelong augmentation of multimodal streaming autobiographical memories”, *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8–3, 2015, pp. 201–213. (Citado nas páginas 15, 40, and 41.)

[PLM$^+$10] Pereira, A.; Leite, I.; Mascarenhas, S.; Martinho, C.; Paiva, A. “Using empathy to improve human-robot relationships”. In: International Conference on Human-Robot Personal Relationship, 2010, pp. 130–138. (Citado na página 42.)

[PMI05] Prendinger, H.; Mori, J.; Ishizuka, M. “Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game”, *International journal of human-computer studies*, vol. 62–2, 2005, pp. 231–245. (Citado nas páginas 73 and 77.)

[RM77] Russell, J. A.; Mehrabian, A. “Evidence for a three-factor theory of emotions”, *Journal of research in Personality*, vol. 11–3, 1977, pp. 273–294. (Citado nas páginas 19, 73, and 75.)

[SCCG20] Spitale, M.; Catania, F.; Crovari, P.; Garzotto, F. “Multicriteria decision analysis and conversational agents for children with autism”. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020. (Citado nas páginas 26, 27, and 48.)

[SHCK19] Sajjadi, P.; Hoffmann, L.; Cimiano, P.; Kopp, S. “A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users”, *Entertainment Computing*, vol. 32, 2019, pp. 100313. (Citado nas páginas 15, 26, 27, 43, 44, 48, 74, 75, and 76.)

[SMML09] Spreng*, R. N.; McKinnon*, M. C.; Mar, R. A.; Levine, B. “The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures”, *Journal of personality assessment*, vol. 91–1, 2009, pp. 62–71. (Citado nas páginas 89, 90, and 97.)

[SVL14] Sutskever, I.; Vinyals, O.; Le, Q. V. “Sequence to sequence learning with neural networks”. In: Advances in neural information processing systems, 2014, pp. 3104–3112. (Citado na página 34.)

[SYF$^+$19] Swanson, K.; Yu, L.; Fox, C.; Wohlwend, J.; Lei, T. “Building a production model for retrieval-based chatbots”, *arXiv preprint arXiv:1906.03209*, 2019. (Citado na página 32.)

[Tar10] Tarasenko, S. “Emotionally colorful reflexive games”, *arXiv preprint arXiv:1101.0820*, 2010. (Citado nas páginas 17 and 74.)

[TB12] Toegel, G.; Barsoux, J.-L. “How to become a better leader”, *MIT Sloan Management Review*, vol. 53–3, 2012, pp. 51–60. (Citado na página 76.)

[Tec20] Technologies, U. “Unity - game engine”. Source: https://unity3d.com/, 2020. (Citado na página 49.)

[Tie09] Tiedemann, J. “News from opus-a collection of multilingual parallel corpora with tools and interfaces”. In: Recent advances in natural language processing, 2009, pp. 237–248. (Citado na página 31.)

[TM07] Tapus, A.; Mataric, M. J. “Emulating empathy in socially assistive robotics.” In: AAAI spring symposium: multidisciplinary collaboration for socially assistive robotics, 2007, pp. 93–96. (Citado na página 73.)

[TWX+19] Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; Yan, R. “Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots”. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 267–275. (Citado na página 33.)

[vHO87] van Holthoon, F. L.; Olson, D. R. “Common sense: the foundations for social science”. University Press of America, 1987, vol. 6. (Citado na página 61.)

[VSP+17] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. “Attention is all you need”. In: Advances in neural information processing systems, 2017, pp. 5998–6008. (Citado na página 34.)

[WTM16] Wang, D.; Tan, A.-H.; Miao, C. “Modeling autobiographical memory in human-like autonomous agents”. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, 2016, pp. 845–853. (Citado nas páginas 27 and 39.)

[WWX+16] Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; Li, Z. “Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots”, *arXiv preprint arXiv:1612.01627*, 2016. (Citado na página 33.)

[Yal20] Yalçın, Ö. N. “Empathy framework for embodied conversational agents”, *Cognitive Systems Research*, vol. 59, 2020, pp. 123–132. (Citado nas páginas 26, 27, 42, and 48.)

[ZGL+19] Zhang, Y.; Gao, X.; Lee, S.; Brockett, C.; Galley, M.; Gao, J.; Dolan, B. “Consistent dialogue generation with self-supervised feature learning", *arXiv preprint arXiv:1903.05759*, 2019. (Citado na página 33.)

[ZHZ+18] Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. "Emotional chatting machine: Emotional conversation generation with internal and external memory". In: Thirty-Second AAAI Conference on Artificial Intelligence, 2018. (Citado na página 32.)

[ZMP+18] Zhao, Z.; Madaio, M.; Pecune, F.; Matsuyama, Y.; Cassell, J. "Socially-conditioned task reasoning for a virtual tutoring agent". In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 2018, pp. 2265–2267. (Citado na página 27.)

Appendix A – PUBLICATIONS

This appendix presents the relation of publications obtained during the development of this research. Section A.1 shows a list of already published re searches, including conference and journal papers. Section A.2 shows a list of submitted papers to conferences and journals without a decision by the moment of delivery of this manuscript.

A.1 Published Research

Generating background NPCs motion and grouping behavior based on real video sequences.
Paulo Knob, *Marlon Alcântara, Estêvão Testa, Rodolfo Favaretto, Gabriel Lima, Leandro Dihl and Soraia Raupp Musse.*
Journal Entertainment Computing (Elsevier). Volume 27, pages 179–187. 2018a.
DOI: https://doi.org/10.1016/j.entcom.2018.06.003.

Visualization of interactions in crowd simulation and video sequences.
Paulo Knob, *Victor Flavio de Andrade Araujo, Rodolfo Migon Favaretto, Soraia Raupp Musse.*
2018 17th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames).
DOI: https://doi.org/10.1109/SBGAMES.2018.00037.

Simulating crowds with ocean personality traits.
Paulo Knob, *Marcio Balotin, Soraia Raupp Musse.*
IVA '18: Proceedings of the 18th International Conference on Intelligent Virtual Agents, November 2018, Pages 233–238.
DOI: https://doi.org/10.1145/3267851.3267871.

Detecting personality and emotion traits in crowds from video sequences.
*Rodolfo Migon Favaretto, **Paulo Knob**, Soraia Raupp Musse, Felipe Vilanova, Ângelo Brandelli Costa.*
Machine Vision and Applications, volume 30, pages 999–1012 (2019).

DOI: https://doi.org/10.1007/s00138-018-0979-y.

Urban walkability design using virtual population simulation.
*CD Tharindu Mathew, **Paulo R Knob**, Soraia Raupp Musse, Daniel G Aliaga.*
Computer Graphics Forum, volume 38, pages 455-469 (2019).
DOI: https://doi.org/10.1111/cgf.13585.

Bioclouds: A multi-level model to simulate and visualize large crowds.
*Andre Da Silva Antonitsch, Diogo Hartmann Muller Schaffer, Gabriel Wetzel Rockenbach, **Paulo Knob**, Soraia Raupp Musse.*
Computer Graphics International Conference, CGI 2019: Advances in Computer Graphics pp 15-27.
DOI: https://doi.org/10.1007/978-3-030-22514-8_2.

How much do you perceive this? an analysis on perceptions of geometric features, personalities and emotions in virtual humans.
*Victor Araujo, Rodolfo Migon Favaretto, **Paulo Knob**, Soraia Raupp Musse, Felipe Vilanova, Angelo Brandelli Costa.*
IVA '19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, July 2019, Pages 179–181.
DOI: https://doi.org/10.1145/3308532.3329454.

Optimal Group Distribution based on Thermal and Psycho-Social Aspects.
***Paulo Knob**, Gabriel Wetzel Rockenbach, Claudio Rosito Jung, Soraia Raupp Musse.*
CASA '19: Proceedings of the 32nd International Conference on Computer Animation and Social Agents, July 2019, Pages 59–64.
DOI: https://doi.org/10.1145/3328756.3328765.

Moving virtual agents forward in space and time.
*Gabriel F Silva, **Paulo Knob**, Douglas A Schlatter, Carlos G Johansson, Soraia R Musse.*
2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames).
DOI: 10.1109/SBGames51465.2020.00026.

Arthur: a new ECA that uses Memory to improve Communication.
Paulo Knob*, Willian S Dias, Natanael Kuniechick, Joao Moraes, Soraia Raupp Musse.*
2021 IEEE 15th International Conference on Semantic Computing (ICSC).
DOI: 10.1109/ICSC50631.2021.00036.

Is my agent good enough? Evaluating Embodied Conversational Agents with Long and Short-term interactions.
Juliane Santos, ***Paulo Ricardo Knob****, Victor Putrich Scherer, Soraia Raupp Musse.*
Proceedings of Simpósio Brasileiro de Jogos e Entretenimento Digital, 2021, Brasil.
Link: https://repositorio.pucrs.br/dspace/bitstream/10923/20535/2/Is_my_agent_good_enough_Evaluating_Embodied_Conversational_Agents_with_Long_and_Shortterm_interactions.pdf.

A.2 Ongoing Publications

Bella: An Empathetic Agent to Serve as a Friend.
Paulo Ricardo Knob*, Natalia Pizzol, Soraia Raupp Musse, Catherine Pelachaud.*
IEEE Computer Graphics and Applications, 2022.
Submitted.

WebCrowds: An Authoring Tool for Crowd Simulation.
Gabriel Silva, ***Paulo Ricardo Knob****, Rubens Montanha, Soraia Raupp Musse.*
2022 Brazilian Symposium on Computer Games and Digital Entertainment (SBGames).
Submitted.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br