The Visual Computer (2024) 40:2933–2948
https://doi.org/10.1007/s00371-023-02994-9

ORIGINAL ARTICLE

Arthur and Bella: multi-purpose empathetic AI assistants for daily conversations

Paulo Ricardo Knob[1] · Natalia Dal Pizzol[1] · Soraia Raupp Musse[1] · Catherine Pelachaud[2]

Accepted: 18 June 2023 / Published online: 12 July 2023

Abstract
The paper presents a novel approach to developing an embodied conversation agent (ECA) that is capable of displaying empathy toward its human partner during interactions. The virtual agents are equipped with both memory and empathy capabilities, with the main focus being on modeling an empathy model associated with the ECA's memory. The paper presents the proposed model of empathy, as well as its connection with memory, and evaluates how this relationship affects the user's experience (UX) through experiments with volunteers who participated in long and short-term interactions. The results of the experiments show that the association of memory with the empathy model makes interactions with embodied conversational agents more enjoyable and to the user. This suggests that the ECA's ability to display empathy can have a positive impact on the user's experience, which is an important factor to consider when designing conversational agents for various purposes. Overall, the paper presents an interesting and valuable contribution to the field of embodied conversational agents and human–computer interaction. The incorporation of empathy and memory capabilities into an ECA has the potential to improve the user's experience and make interactions with machines more human-like.

Keywords Embodied conversational agent · Virtual agents · Empathetic agent

1 Introduction

Emotions are a fundamental aspect of human behavior that have long been recognized as important for intelligent systems [7, 27]. Empathy, in particular, is a crucial socio-emotional behavior that allows individuals to share and understand the emotions of others [9]. For example, if someone is talking to someone who just lost a beloved relative, he/she can perceive that this person is truly sad and show sadness.

The term "Computational Empathy" has been gaining popularity in the research field with the need for more virtual agents to act as social partners [42]. Such research aims to discover how virtual agents can exhibit empathetic behavior toward users during interactions. Several studies in the literature have demonstrated the benefits of adopting an empathetic virtual agent, including reducing stress and frustration [2, 32] and providing more engagement [6, 32]. Additionally, many recent methodologies have proposed models of empathy for virtual agents in the last few years [4, 35, 36, 44], as later discussed in Sect. 2.

Although significant contributions have been made in the area, little effort is being applied in the investigation of the relationship between the agent's empathy and memory. It is established that we, as human beings, need both our empathetic and mnemonic skills to cope with our social routine [40]. When we see another person suffering, our empathy process can be influenced by the retrieval of memory details from ourselves [39]. This work proposes two ECAs endowed with an empathy model connected with their memory. To do this, we use an ECA platform we have developed previously [18]. It is a multi-purpose embodied conversational agent, endowed with a memory model. The main differences between the two agents proposed in this work (namely Arthur and Bella) are their embodiment and their respective

✉ Paulo Ricardo Knob
paulo.knob@edu.pucrs.br

Natalia Dal Pizzol
natalia.pizzol@edu.pucrs.br

Soraia Raupp Musse
soraia.musse@pucrs.br

Catherine Pelachaud
catherine.pelachaud@upmc.fr

[1] PUCRS, Porto Alegre, Brazil

[2] CNRS - ISIR, Paris, France

facial expression animations. Their behavior, i.e., memory, emotions, chatting, and empathy, work in the same way, independently of the embodiment chosen (Arthur or Bella). The main contribution of this work is an empathy model and its relationship with an artificial memory. Regarding the purpose of the ECAs, we intend to keep them without a specific application, intending to provide multi-purpose agents. In addition, another contribution of this work is the long-term and short-term user interactions we propose to study the impact of empathy on the user experience. It is worth noting that, to the best of our knowledge, there has been little investigation in the literature on long-term interactions.

2 Related work

One of the definitions of empathy states that it is "the ability to understand and react toward the emotions of others" [8], being an essential trait for smooth interpersonal interactions. Nevertheless, the complexity of an empathetic model is related to the wide range of behaviors it arises from, such as mirroring, affective matching, empathetic concern, altruistic helping, and perspective taking [6, 9]. The work of Yalcin et al. [43] aims to model empathetic behavior on ECAs. The ECA built by the authors has three stages: listening, where the agent captures input from the person it is talking to; thinking, where the agent processes the information; and speaking, where the agent gives a proper response, both with words and gestural behavior. Concerning empathetic behavior, it should allow the agent to respond to the user in a verbal and nonverbal way. Since the empathetic behavior relies on the user's emotion, an emotion recognition module is used.

Sajjadi et al. [35] conducted an experiment to investigate the effect of a person interacting with an ECA with personality. The authors hypothesized that an ECA with its non-verbal behavior governed by a personality-driven behavioral model would increase the level of social presence of the person and provide a better game experience. If the user says something to the virtual agent, Linda (the virtual agent) can update its emotional state toward a given state. On the other hand, if the user does not provide any stimuli while talking, Linda can become bored. An experiment was conducted with 41 participants 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, if compared with a less emotionally personified agent displaying no nonverbal 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.

Spitale et al. [36] present a framework for an empathetic conversational agent, which is grounded on the empathy theory. Their framework is divided into three modules: empathic perception module (what the artificial agent can perceive from the human interaction partner); empathic behavior module (decides how the virtual agent can act, both in the listener and speaker roles); and empathic regulation module. The authors comment that one possible limitation of their framework is that it might be too general and unable to capture some specific empathetic properties intrinsic to the agents and their application.

Pereira et al. [30] presented a robot aimed to act as a social companion, expressing 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 toward one of them and neutral with the other. The study results suggest that the players with whom the robot was being empathetic perceived it more as a friend than the other players.

Concerning memory models, Wang et al. [41] built a model to mimic the way human memory works based on the autobiographical memory model (called AM-ART). They proposed a three-layer neural network that encodes lifetime periods, general events, and event-specific knowledge. Their results show that AM-ART performed better than the keyword-based query method since the latter cannot deal with noise in many existing imagery or memory repositories. Edirisinghe et al. [10] also modeled an autobiographical memory, but for a robot that can store knowledge about users during friendly interactions, recalling them for future interactions. The autobiographical memory was developed in a three-layer architecture. Once the robot interacts with a new person, it creates a user profile. The results show the potential of such memory mechanism for robots, which can improve the long-term interactions between humans and robots.

Kasap et al. [17] focus on the problem that people often lose interest in virtual agents or robots after the novelty effect disappears. In order to build a long-term interaction model that can keep users' interest, they developed a robotic tutor, Eva, endowed with many aspects such as emotion and memory. The results achieved by their work provide evidence that the use of a memory system in a long-term interaction can effectively help keep the users' attention as time passes by.

Martinez et al. [21] consider 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").

Petit et al. [31] implement Autobiographical Memory in a robot, named ICub. 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 recognized 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.

3 Proposed model

In this work, we propose two ECAs equipped with empathy, where the empathy is associated with the agents' memory. Thus, we aim to investigate the relationship between memory and empathy. For this purpose, we extend an existing ECA named Arthur, as proposed by Knob et al. [18]. The overview of our method is illustrated in Fig. 1. In this paper, we focus on the new features (highlighted in yellow); the other modules are already described in the previous work [18].

The main difference between the previous work [18] and the current study lies in the inclusion of an empathy module between the two controls (memory control and behavior control, in blue). In the original model, theses controls are directly connected. The addition of the empathy module creates a relationship between empathy and memory, allowing the virtual agent to update its emotional state based on what it remembers. This, in turn, helps the agent provide empathetic behavior. The following sections describe in detail all aspects highlighted in yellow in Fig. 1.

3.1 Memory

In this work, we adopt the concept of autobiographical memory to model the memory of our virtual agent. According to Conway et al. [5], autobiographical memory can be categorized into three levels: lifetime periods, general events, and event-specific knowledge (ESK). We use General Events and ESK to represent our agent's memory. General Events refer to the events that occur during the interaction between the virtual agent and the user, while ESK contains detailed information about each General Event. To simplify the nomenclature, we refer to each detail in ESK as a resource. Therefore, our ESK consists of a pool of resources that provide various details about events.

Besides that, we store the memory of our virtual agent in two levels: Long-term memory (LTM) and Short-term memory (STM). According to Loftus et al. [20], STM is used to store important information for a short period of time, while LTM is an information storage with virtually unlimited capacity that each human being possesses. In our work, we model STM and LTM separately. The STM comprises a list of resources that can have, at most, seven items, as defined by Miller's Law [25]. If a new resource needs to enter the STM, the less important information is forgotten based on its weight (i.e., the resource with the lower weight is removed). In our work, each general event comprises six basic pieces of information: timestamp, id, type (e.g., meeting a new person), emotion (recognized by the virtual agent in the user), polarity (positive or negative information), and resources (details of the event). On the other hand, each resource present in the ESK is also comprised of six basic pieces of information: timestamp, id, type (e.g., text), information, activation (rehearsal process), and weight (importance).

When the virtual agent retrieves information from its memory, it follows the Generative Retrieval method. This means that when the user provides information during an interaction, it can be used as a cue to trigger the retrieval of relevant memories. For instance, if the user says "I went fishing with my dad", the words "fishing" and "dad" can be used as cues to search for relevant memories in the database. In addition, we developed a module for memory consolidation that prioritizes important memories over mundane ones. Emotional memories, for example, are deemed more important and have a stronger impact than neutral memories. Thus, less important information may be pushed to the background or even forgotten. This module considers two pieces of information from the resources in STM: Activation and Weight.

For further details on the behavior of the memory model, please refer to our previous work [18].

3.2 Bella

While Arthur is embodied as a 2D male cartoon face, Bella is embodied as a 3D female cartoon face. Both Arthur and Bella are illustrated in Fig. 1. We acquired a 3D rigged model of a cartoon female head.[1] Using this model, we created six basic emotions defined by Ekman [11], including happiness, fear, disgust, anger, surprise, and sadness. In addition, we developed a simple lip-sync tool for both Arthur and Bella. The tool analyzes the audio frequency of what they need to speak and generates motion values. These values are then used to rotate

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

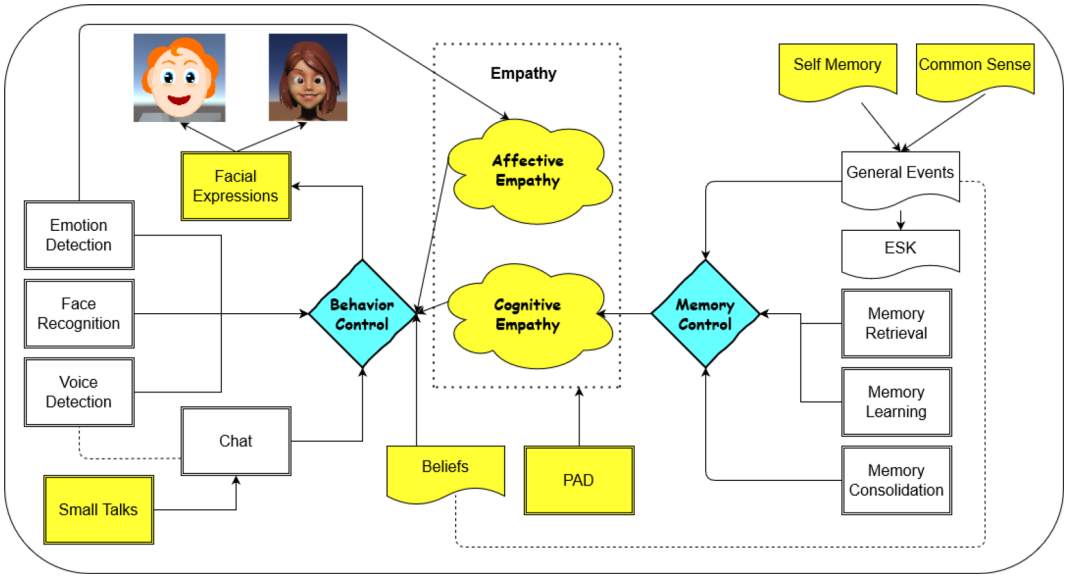

Fig. 1 Overview of the proposed model. The main aspects presented in this work are highlighted in yellow: a new avatar (in facial expressions), beliefs module, self-memory module, common sense module, PAD module, small talks module and empathy

the virtual agents' jaws (or blendshapes), allowing Arthur and Bella to open and close their mouths in sync with their speech. As suggested in Fig. 1, the only differences between the two virtual agents (i.e., Arthur and Bella) are their embodiment and facial expression animations. Independently of choosing Arthur or Bella, the virtual agent works in the same way and with the same modules presented in this work (i.e., memory, empathy, and so on).

3.3 Facial expressions evaluation

In our study, we manually modeled Bella's facial expressions to portray the six basic emotions. To assess users' ability to recognize these emotions, we conducted a perceptive study. Fifty-eight volunteers participated in the experiment, of whom 22 were men and 36 were women. Of these, 26 reported familiarity with graphical computing, while another 26 did not, and 6 preferred not to answer the question. Participants were shown video sequences in which Bella displayed facial expressions corresponding to the six basic emotions, and were asked to identify the emotion portrayed. The results showed that participants were able to correctly identify all six emotions, with Happiness being the easiest to recognize (98.8%), and Anger and Fear being the most difficult (82.7% for both).

Furthermore, we carried out a similar experiment with Arthur, recruiting 58 volunteers—32 women, 25 men, and 1 person identifying as Other. Of the participants, 38 reported some familiarity with graphical computing, while 16 did not, and 4 did not provide a response. The results indicate that the majority of participants were able to identify the different emotions, with Happiness being the easiest to identify (96.6%). However, many participants had difficulty identifying Anger (34.5%). In the next sections, we discuss the new modules presented in this work.

3.4 ECAs' beliefs

The beliefs model was developed to enable our virtual agent to be able to reason with respect to different pieces of information. For example, if a person states that they have two children named John and Mary, we automatically infer that John and Mary are siblings. To incorporate this kind of reasoning into our ECAs, we integrated a knowledge-based system of PROLOG statements. These statements can be created manually to meet specific requirements, or automatically generated based on the agent's memory. By using this method, we are able to efficiently encode such features into our ECAs.

As the virtual agent interacts with people, information about the conversation is stored in its artificial memory (as explained in Sect. 3.1 and detailed in the work of Knob et al. [18]). In Fig. 1, a dotted line connects the agent's memory to its Beliefs. Based on this, we developed a script that automatically generates beliefs for the virtual agent by retrieving the resources stored in its event-specific knowledge (ESK) and general events and using them to create PROLOG statements reflecting that knowledge. For example, if a user tells Arthur/Bella that they have two children named John and Mary, the ECA stores this information in its memory and creates two PROLOG statements: parent(user, john) and parent(user, mary). The ECA can use these statements to

understand that “user" is a “parent" of “john" and “user" is a “parent" of “mary", and even infer more complex relationships such as sibling relationships.

3.5 ECAs’ self-memory

When people converse, it is common to ask questions about each other, such as ’What’s your favorite food?’ or ’Do you like music?’. If a user asks ’personal’ questions to an ECA (Embodied Conversational Agent), it is important for the ECA to have knowledge about itself to respond appropriately. The self-memory Model was created to provide Arthur/Bella with information about themselves. We manually included some information on various casual topics (such as name, age, and music) into the ECA’s initial memories. Whenever the user inquires about these topics, Arthur/Bella can search its memory and respond accordingly.

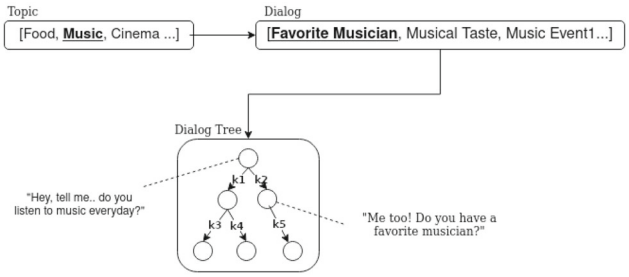

Fig. 2 Small talks structure. Topics define the broad subject, dialogs refer to a given topic, and each dialog has a dialog tree with different utterances. In the example, the selected topic was “Music" and the selected dialog was “Favorite Musician". The first thing Arthur or Bella says inside this dialog is the utterance present in the root of the dialog tree (i.e., “Hey, tell me... Do you listen to music every day?"). Each “k" ($k1$, $k2$, $k3$, etc.) represents a set of keywords. If the user’s answer has many words present in $k1$, the path is traveled downleft. Otherwise, if the answer has many words present in $k2$, it is traveled downright (i.e., “Me too! Do you have a favorite musician?")

3.6 ECAs’ common sense

The common sense module aims to provide Arthur/Bella with a basic understanding of a wide range of topics. To accomplish this, we chose to use WordNet,[2] an extensive lexical database of English terms that includes concepts for many nouns and verbs. The database contains about 150,000 words, each with a corresponding description. Due to performance constraints, we were unable to include the entire database in our model. Therefore, we sorted the words by their sense number, a way to represent word relevance provided by the database, and included the first 10,000 word/description pairs in our agent’s memory. Additional pairs can be added in future versions.

When asked about any of these terms, Arthur/Bella can respond with the appropriate description. As the word/description pairs are stored in the agent’s memory, the common sense module is directly related to the agent’s memory. These pairs are translated into the agent’s autobiographical memory, becoming part of its knowledge. For example, consider the word/description pair ’dog/best friend of men’. The term ’dog’ is added to the event-specific knowledge as a new resource, and a new general event is created with the word’s description (i.e., ’best friend of men’) and connected to the new resource ’dog’.

3.7 Small talks

According to the Cambridge Dictionary, small talk is defined as ’conversation about unimportant things, often between people who don’t know each other well’.[3] Bringing this concept to Arthur and Bella offers the advantage of building a more comfortable relationship between a virtual agent and a human, especially for long-term interactions [29]. The main idea is to allow both Arthur and Bella to engage in conversation about topics that people often discuss when they do not have a specific task or problem to solve.

For example, asking someone about their favorite band may not be directly relevant to achieving a particular goal, but it is important for establishing a friendly and positive relationship between two individuals. Small talk is commonly used during interactions, often to create closer relationships with others.

The small talks module can be seen in Fig. 1, at the bottom left, connected with the chat module, which allows for the user to communicate with the virtual agent, be it by voice or text (for more information, please consult the previous work [18]). To build our small talks, we need to define their structure. While there are many research papers on dialogue systems [26], we decided not to create a new system from scratch. Instead, we chose to use a simple conversational structure based on a decision tree [3]. We opted for a decision tree because it is a straightforward and sturdy way to model a dialog flow. Our small talk feature is structured into three main components: topics, dialogues, and dialog trees, as shown in Fig. 2. Topics define broad subject areas that Arthur or Bella can talk about, such as Music, Food, and Sports. Each topic includes a set of dialogues that can be used to initiate a conversation.

For each dialogue within a topic, a dialog tree is manually created. The dialog tree consists of nodes and branches that determine what Arthur/Bella will say to the user during the conversation. For example, in the “Music" topic, a dialogue could be “Favorite Musician", and the corresponding dialog

[2] https://wordnet.princeton.edu/.

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

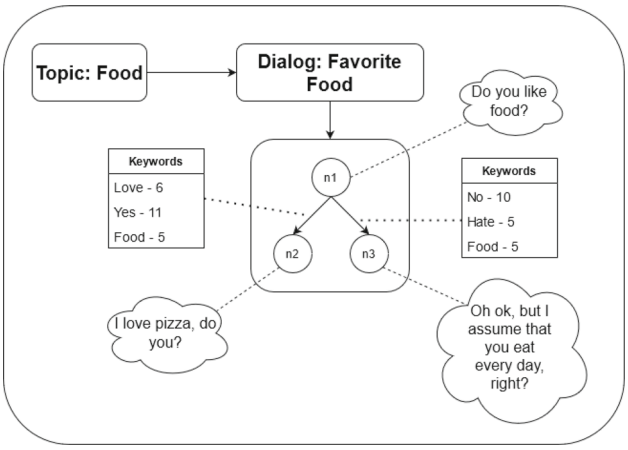

Fig. 3 Small talk example. Inside the topic: Food and the dialog: Favorite Food there is a dialog tree with three nodes (which means, three possible utterances for the agent). The dialog starts with the root node n1 (Do you like food?). If the answer of the user contains words like "Love", "Yes" or "Food", 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?)

tree would contain various nodes with different utterances related to the topic.

Figure 3 presents an example of this process, where the chosen topic is "Food" and the dialogue is "Favorite Food". Inside this dialog, there is a dialog Tree with three nodes, each representing a possible utterance for the agent. Arthur or Bella begins the conversation with the utterance associated with the root node n1: "Do you like food?". If the user's response contains words like "Love", "Yes", or "Food", which are the keywords associated with the branch connecting n1 to n2, the agent's next utterance will be n2: "I love pizza, do you?". On the other hand, if the user's response contains words like "No", "Hate", or "Food", which are the keywords associated with the branch connecting n1 to n3, the agent's next utterance will be n3: "Oh, okay, but I assume that you eat every day, right?". The association between a keyword and a branch is determined based on frequency. For instance, if a user responds to n1 with something like "I hate food" and travels to n3, the words "hate" and "food" are added to the keyword set of the respective branch. This allows the same keyword to be present in different branches, such as "food" in both "like food" and "hate food".

It is important to clarify a few things here. First, let's discuss how the keywords are associated with each branch. In Fig. 3, we can see that the keyword "Food" appears in both branches (leading to n2 and n3). These keywords are selected based on their frequency of occurrence. For example, in the case of the example shown in Fig. 3, depending on how people respond to the first question n1 ("Do you like food?") and where they go next (n2 or n3), keywords are assigned to each set. If someone responds to n1 with something like "I hate food" and goes to n3, the words "hate" and "food" are added to the keyword set of the respective branch.

Secondly, in Fig. 3, there are numbers associated with each keyword (for example, "Love" has 6 and "Food" has 5). These numbers represent the frequency of usage of the respective keyword during interactions and are used to reach their respective nodes. For instance, 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 of each keyword's usage in interactions, both alone and in relation to other keywords of the set. As a result, we can determine which keywords are more significant in progressing from one node to another. For example, if many people respond to n1 with the keyword "Love," its frequency is high, and therefore, we can give it more weight when deciding which node the agent should travel to. This approach has two advantages. Firstly, it does not matter if the same keyword appears in different sets (like "Food" in the example of Fig. 3) because it can have different weights, and other keywords can have much higher weights. Secondly, based on the occurrence of words, Arthur and Bella can learn from previous interactions and improve their decision-making when traveling down the dialog decision tree.

We chose to work with two different frequencies: Simple Frequency (FSimple) and Sibling Frequency (FSibling). The simple frequency is simply the number of times a given keyword k appears in node n (nt), divided by the total number of keywords in this node (nt^n_{all}). The formula is as follows:

$$\text{FSimple}^n_{\text{k}} = \frac{\text{nt}^n_{\text{k}}}{\text{nt}^n_{\text{all}}}, \tag{1}$$

where $\text{FSimple}^n_{\text{k}}$ is 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 all nodes of the same tree level ($\text{nt}^{\text{level}}_{\text{k}}$). The formulation goes as follows:

$$\text{FSibling}^n_{\text{k}} = \frac{\text{nt}^n_{\text{k}}}{\text{nt}^{\text{level}}_{\text{k}}}, \tag{2}$$

where $\text{FSibling}^n_{\text{k}}$ is also a value between 0 and 1. Finally, the final frequency is simply the mean value between the simple and the sibling frequencies, as follows:

$$F^n_{\text{k}} = \frac{(\text{FSimple}^n_{\text{k}} + \text{FSibling}^n_{\text{k}})}{2}, \tag{3}$$

where F^n_{k} will be the frequency that keyword k appears in node n.

Finally, since Arthur or Bella only engage in small talk when the interaction seems to have "cooled down", a timer was implemented. Empirically, it was defined that if the user

does not respond to the virtual agent for 30 s, but remains in front of the webcam, Arthur or Bella randomly select a pair of topic/dialog and initiate a small talk conversation.

3.8 ECAs' empathy model

Empathy can be defined as the sharing of emotions between individuals and the ability to adopt another person's point of view [9]. Verbal and non-verbal communications are known to be helpful when emulating empathy [38]. In our empathy model, Arthur/Bella exhibit three human-like characteristics: personality, emotion, and mood, which are described in detail in the following sections. According to Kshirsagar [19], personality is defined as "the characteristics of a virtual human that distinguish it from others", while emotion is a "momentary state of mind," and mood is "a prolonged state of mind resulting from the cumulative effect of emotions." Our agent's personality is defined based on the OCEAN model [12], while the emotional states of Arthur and Bella are modeled using the PAD dimensions [34]. We chose to work with both OCEAN and PAD models as they are widely accepted for personality and emotion representation. By dynamically changing its emotional state, our virtual agent can behave empathetically toward the user. The following sections provide more details on these characteristics.

3.8.1 Pleasure–arousal–dominance

To endow Arthur/Bella with emotional states, we chose to work with the pleasure–arousal–dominance (PAD) model. The PAD model is a three-dimensional space used to represent emotional states, which was introduced by Russell and Mehrabian [34], and it is considered a good alternative to define and represent many emotional states. The authors suggest 151 different emotional states represented inside the PAD space. However, in this work, we decided to use 13 emotional states, as defined by Russell and Mehrabian [34] (except the Neutral emotional state $PAD = (0, 0, 0)$, which we defined as a starting point at the intersection of the three PAD dimensions), shown in Table 1. 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 during the interaction with the user, and then, it is used to change the agent's emotion. More details are provided in Sect. 3.8.4.

Table 1 Emotional states of our ECA, adapted from Russell and Mehrabian [34].

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

P stands for Pleasure, A stands for Arousal and D stands for Dominance

3.8.2 Personality

To define the personality of our agent, we opted to use the OCEAN model, also known as the Big Five, proposed by Goldberg [12]. We assigned a personality profile to our virtual agent based on the extrovert/introvert trait, limited by the Extraversion (E) trait, as suggested by Sajjadi et al. [35]. Specifically, we considered the agent introverted if E falls within the range $[0, 0.5)$ and extroverted if E falls within the range $[0.5, 1]$. We then translated the personality profile into our PAD three-dimensional space to generate a default emotional state for Arthur/Bella, as follows: for the extroverted personality profile, the default PAD value was set to $PAD_{\mathrm{E}} = (0.8, 0.5, 1)$, corresponding to high pleasure, moderate arousal, and high dominance, respectively. This value is closest to the Happy emotional state in Table 1. For the introverted personality profile, the default PAD value was set to $PAD_{\mathrm{I}} = (-0.8, 0.3, -1)$, which corresponds to low pleasure, low arousal, and low dominance. This value is closest to the depressed emotional state in Table 1. We set these default values based on the work of McCrae et al. [22].

Moreover, besides the Extraversion dimension used by Sajjadi et al. [35], we also include the Neuroticism dimension to define the default emotional state. We chose to use Neuroticism to change Dominance based on its own definitions. As defined by Mehrabian [23], 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). In addition, as commented by Kagan et al. [16], people who reach high scores in the Neuroticism trait tend to be emotionally reactive and vulnerable to stress, also tending to be shallow in the way they express emotions. If Arthur/Bella has a Neuroticism value above 0.5 (values lie between 0 and 1), we assume that it is a bit paranoid and may not feel in control of its own emotions. Therefore, if it is extrovert (i.e., $PAD_{\mathrm{E}} = (0.8, 0.5, 1)$), we can reduce its Dominance, resulting in $PAD_{\mathrm{E}} = (0.8, 0.5, 0.5)$. Otherwise, if the agent is an introvert (i.e., $PAD_{\mathrm{I}} = (-0.8, 0.3, -1)$) with

a Neuroticism value lower or equal 0.5, we can increase his Dominance, resulting in $PAD_{\mathrm{I}} = (-0.8, 0.3, -0.5)$. In the other two cases (i.e., Neuroticism above 0.5 and introvert; Neuroticism lower or equal 0.5 and extrovert), no changes are made in the Dominance value. It is important to note that Dominance has its initial value set and does not change during the interactions, but its value is important to define the emotional state of the agent inside the 3D spatial dimension of PAD. For instance, if we have P = 1 and A = 1, a value of D= −1 will result in a different emotional state when compared with a value of D = 1. PAD_E and PAD_I are used to define the initial emotional state of the agent, depending on the personality given, which in our case is based on values of E and N, from OCEAN. It is important to emphasize that each instantiated ECA has a fixed value of personality, which does not change during the interaction with the user.

3.8.3 Emotional state

We first discuss our modeling of Bored emotional state in our ECAs. When two or more individuals interact, the conversation can become awkward if they spend too much time without saying anything. Awkward silences can make people uncomfortable and even bored. Moreno et al. [28] state that interactions that experience an "under-loading" of information can lead to boredom and disengagement. In order to mimic such behavior, we included a Boredom value (based on [35]) in our virtual ECA, 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 decrease if the user stays 15 s (empirically defined) without interacting with Arthur/Bella. When an interaction occurs, the Boredom value is reset to 0. The boredom varies linearly, i.e., Bor− = 0.005, if the time without interaction is greater than 15. Finally, the Boredom value Bor is also used for the emotional state update, as it will be explained in Sect. 3.8.4.

3.8.4 Emotional states update and empathetic behavior

Empathy can involve cognitive and affective attributes, which also can be combined [13]. Cognitive attributes of empathy involve cognitive reasoning used to understand another person's experience [15]. Emotional or affective attributes involve physiological enthusiasm and spontaneous affective responses to someone else's display of emotions [33]. Our emotions can change depending on how the interaction flows when we talk with someone else. Similarly, an Embodied Conversational Agent endowed with emotion should be able to change its emotional state as interactions occur. In order to do so, we compute the emotional state of Arthur/Bella in three specific updating 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; and
3. When something is remembered by the virtual agent.

As commented in Sect. 3.8.2, PAD_{E} and PAD_{I} are defined for our ECA based on its personality, specifically the E and N factors from OCEAN [12]. Indeed, the initial value of PAD, when the ECA is initialized, is considered its comfort zone (C_{z}), which means an emotional state in which the ECA feels comfortable. ECA's PAD can be updated toward close or far from the C_{z} during the interaction. Thus, we implemented a bonus or penalty of 0.05 (empirically defined), 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). The emergent effect is that the virtual agent wants to be in its comfort zone C_{z}: it approaches faster than C_{z} and distances slower from C_{z}. The update of the PAD values is done based on the work of Sajjadi et al. [35], adapted to our model to include the comfort zone (C_{z}), as following defined:

$$P = ((P + \mathrm{Pol})/2) + C_{\mathrm{z}}, \tag{4}$$

where P is the Pleasure dimension of PAD, Pol is a value that lies between −1 and 1 (i.e., Pol = [−1, 1]) and C_{z} represents the bonus or penalty of the comfort zone. In updating situation (1), the Pol value stands for the polarity of the sentence, meaning how positive, neutral or negative the information is. For example, if someone says "I woke up feeling great today", it is interpreted as a positive sentence (e.g., Pol = 1). On the other hand, if someone says "I woke up feeling so bad today.", it is interpreted as a negative sentence (e.g., Pol = −1). In updating situations (2) and (3), Pol stands for the valence of the emotion, meaning how positive or negative the emotion is, being recognized in the face of the person (situation 2) or associated with a given memory (situation 3). For situation (2), the Affectiva plugin[4] is used to capture this information and for recognition of user's emotion. For situation (3), the emotional information is stored in the memory of the virtual agent. In addition to Pleasure, the Arousal dimension is defined as follows:

$$A = |\mathrm{Pol}| + \mathrm{Bor} + C_{\mathrm{z}}, \tag{5}$$

where A is the Arousal dimension of PAD and Bor is the boredom, as previously explained. The |Pol| indicates that only the modulus of the Pol value is used. Finally, the Dominance's initial value is defined depending on the agent's personality (as explained in Sect. 3.8.2), and remains fixed during interactions. At any given moment, when the PAD

[4] https://affectiva.com/.

value is updated, the virtual agent's facial expressed emotion is updated to reflect this change, by selecting the closest emotion value, according to a previous work [34]. A simple distance function between two three-dimensional points is used. No changes are made if the actual facial emotion is still the closest; otherwise, if a different emotion is found to be closer than the current one, the new emotion is set and Arthur or Bella play the respective animation.

4 Experimental results

To evaluate our model, we conducted two different experiments. For the first experiment, we decided to conduct both long-term interactions (LTIs) and short-term interactions (STIs) with users. For the second experiment, we decided to conduct only short-term interactions (STIs). For the experiments discussed in this section, the personality of the agent is set as the following OCEAN values (corresponding to the Extrovert personality): $O = 0.9$; $C = 0.5$; $E = 0.9$; $A = 0.7$; $N = 0.5$. The initial PAD value (calculated from the Extrovert personality, i.e., PAD_{E}) is, thus, set as follows: $P = 0.8$; $A = 0.5$; $D = 1$. A video available online[5] showing examples of interactions with the virtual agent. It is important to note that, given the number of participants in the experiments, as presented next, we are not going to analyze differences between people (e.g., men X women), but focus on the contributions provided by this work, such as the empathy module and its relationship with the memory model, as well as the comparison between LTIs and STIs. Details are provided in the next sections.

4.1 Empathy experiment

This experiment was conducted with both LTIs and STIs. The LTI participants filled in the questionnaire in September/2021, interacting with Arthur or Bella for 10 days, only once daily. Each interaction lasted for about 10–15 min. Four users, two men and two women, interacted daily with our ECAs for 10 days, resulting in 40 answers for our survey, which has three questions. All 4 participants were Brazilian. Details about each participant can be seen in Table 2.

All participants (from STI and LTI) read and agreed with the ethics term presented at the beginning of the questionnaire. Since empathy is essential to our work, all participants were presented with a brief explanation about emotion and empathy before starting the interactions. Additionally, before the start of the experiment, participants filled in the Toronto empathy questionnaire [37] (TEQ) to measure their empathy level of the participants, also shown in Table 2. It is important to remember that the average score for men lies on the interval [43.46; 44.45], while the average score for women is in [44.62; 48.93], according to [37]. Man 1 scored below the average, while the other three participants scored above the average.

After each interaction with one of the ECAs, the participants answered a questionnaire compounded of the following parts (illustrated in Table 3): (1) One question concerning the user satisfaction (UX) (Question 1 from Table 3); (2) one question from Bartneck "Godspeed" questionnaire [1] and one question from Heerink questionnaire [14], in a total of 2 questions (Questions 2 and 3 from Table 3); (3) Free text field, where the participant could freely write his/her impressions about the ECA and the interactions. Although both questionnaires (i.e., Bartneck and Heerink) were mainly used in evaluating robots, some questions can be adapted to a virtual agent.

Additionally, eight volunteers, consisting of four men and four women, participated in our STIs, engaging with our ECAs for a single session and completing a standardized questionnaire. Each interaction lasted between 10 and 15 min. Further details regarding each participant are presented in Table 4. Prior to the STIs, all participants received a brief explanation about emotions and empathy and completed the Toronto Empathy Questionnaire [37] (TEQ) to measure their empathy levels. Notably, Man 4 scored below average, Woman 1 scored within average, while the remaining 6 participants scored above average in empathy levels.

In order to conduct this evaluation, we define two conditions: empathy (activated/deactivated) and type of interaction (LTI/STI). Based on these conditions, we formulated two hypotheses:

- *H1* We expect that the interactions with the agent in which the empathy module is activated are going to be more satisfying to the user than the interactions with the same agent without empathy; and
- *H2* We expect that the results achieved in the STIs will be represented by higher empathy values than the results achieved by the LTIs. Our hypothesis here is justified by the fact that users may not notice some problems (e.g., vocabulary, agent not being able to answer something, software errors) with STIs, while LTI users deal with prolonged interaction time.

It is important to note that the memory was activated in both LTI and STI cases. In order to test our hypothesis H1 (We expect that the interactions with the empathetic agent will 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. Regarding the LTI, Man 1 and Woman 1 interacted with the ECA with empathy, while Man 2 and Woman 2 interacted with the ECA without empathy. It is important to note

[5] https://youtu.be/kI8eHW30W8U.

Table 2 Participants of the LTI experiment

Participant	Age	Education level	Experience 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

Table 3 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 Participants of the STI experiment

Participant	Age	Education level	Experience 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

that users did not know previously if the Empathy skill was available on the ECA or not. For this hypothesis, we evaluate questions 1–3 from Table 3.

Figure 4 presents the average scores of the four LTI participants in the three evaluated questions for the 10 days of interaction. It is possible to note in Fig. 4 that the best scores were reached by Man 1 (3.3, 2.8, and 2.6), who used 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). Additionally, it is noteworthy that in question 3 ("Sometimes the agent seems to have real feelings"), Man 2, who interacted with the agent without empathy, achieved a higher score (2) compared to Woman 1 (1.8), who interacted with empathy. One possible explanation could be linked to the nature of the third question itself. On closer inspection, we can observe that the results for question 3 are lower when compared to the other two questions. It is possible that because both of our embodiments have a cartoon-like appearance (Arthur and Bella), users evaluated the visualization of emotions combined with graphical realism, which might have caused this distortion. Another possibility is that while some traces of empathy can be seen in the other two questions, the only question that directly measures empathy is question 1. Hence, the results for question 1 align with our expectations, as the evaluation values for Man 1 and Woman 1 were higher compared to Man 2 and Woman 2. However, further experiments would be necessary to confirm or refute this hypothesis.

Moreover, we explored the temporal evolution of the answers of the four participants during the 10 days of interaction. Figures 5, 6 and 7 present such temporal evolution for questions 1–3 in Table 3, respectively. In Fig. 5, concerning question 1, it is possible to notice that Man 1 scored a 2 on his first day and alternated between 3 and 4 on the remaining days. Woman 1 presented a great variety of values, from 1 to 5. Man 2 and Woman 2 presented a similar behavior: from day three onward, they scored 2 until the end of the interaction. It seems to indicate that the group with the empathy module activated (Man 1 and Woman 1) was more satisfied with the agent's empathy than the group with the empathy module deactivated (Man 2 and Woman 2).

In Fig. 6, concerning question 2, it is possible to notice that Man 1 scored a 3 in the first 4 days, alternating between 2 and 4 on the remaining days. Woman 1 presented a greater variation, starting with a 2 on her first day, 1 on her second day, and passing through 3, 2, and 4 in the remaining days. Man 2 and Woman 2 varied from 1 to 3 in the 10 days, never scoring 4 or 5. Although no interesting pattern could be perceived, it is interesting to note what happened in each day that caused a change of perception. For instance, we can see in Fig. 6 that Woman 2 scored 1 on her first interaction day,

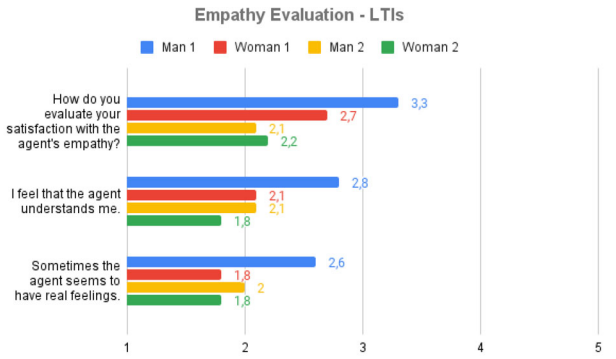

Fig. 4 Average scores of the empathy assessment for the LTIs, referring to questions 1–3 in Table 3 for 10 days of interaction. The Likert scale was converted to numbers, so Very Unsatisfied is 1 and Very Satisfied is 5

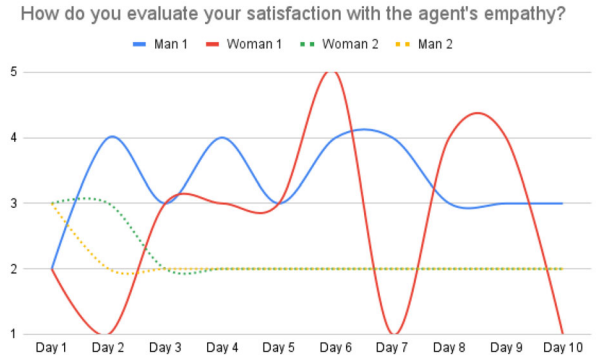

Fig. 5 Temporal evolution of the results regarding question 1 in Table 3, for the 10 days of interaction and all 4 participants. The Likert scale was converted to numbers, so very unsatisfied is 1 and very satisfied is 5

but raised her score to 3 on the second day. In the free text field, she commented that "Arthur was funny today, he even made me laugh. He was also kinder and friendlier than yesterday.". It is also possible to note that she dropped her score to 1 again on day 4, to which she commented that the agent was presenting some unexpected behavior, like mistaking her name. Moreover, we can note that Man 1 went from score 1 to score 4 between days 5 and 6, while Woman 1 went from 3 to 1 in the same period. However, the participants did not provide any insight into the free text field, so nothing could be inferred.

In Fig. 7, concerning question 3, it is possible to notice that Man 1 scored a 4 on the first day, 3 between days 2 and 5, and 2 on the remaining days. Woman 1 started with a 1, then scored 2 in the next 2 days and 3 on 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 pattern could be perceived, it is interesting to note what happened in each day that caused a change of perception. Once again we look at change of perceptions. For instance, we can see in Fig. 7 that Man 2 scored 3 on his fourth day of interaction, but dropped his score to 2 on the fifth day and to 1 on the sixth day. In the free text field, he commented that Bella was uttering several strange phrases and was mistaking his name. Also, he commented that Bella offered herself to be a calculator, "but did not understand simple operations half the time".

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 8 presents the average scores of the eight STI participants in the three evaluated questions. It is possible to note in Fig. 8

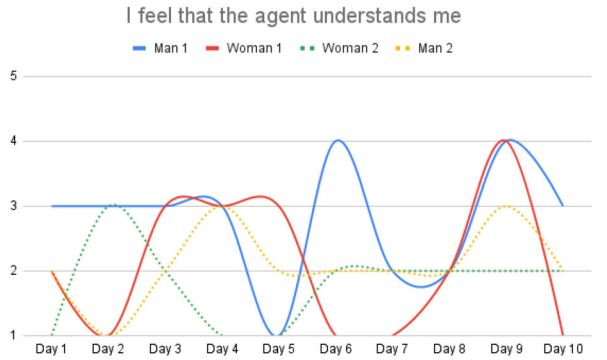

Fig. 6 Temporal evolution of the results regarding question 2 in Table 3 (10 days of interaction and 4 participants). The Likert scale was converted to numbers, so very unsatisfied is 1 and very satisfied is 5

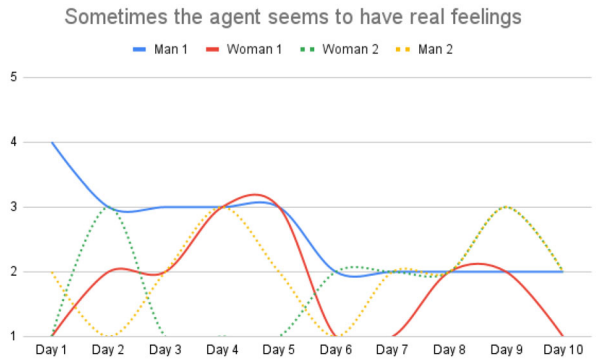

Fig. 7 Temporal evolution of the results regarding question 3 in Table 3 (10 days of interaction and 4 participants). The Likert scale was converted to numbers, so very unsatisfied is 1 and very satisfied is 5

that the results for question 1 (How do you evaluate your satisfaction with the agent's empathy) were very similar. Five participants scored 4, while the other three 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

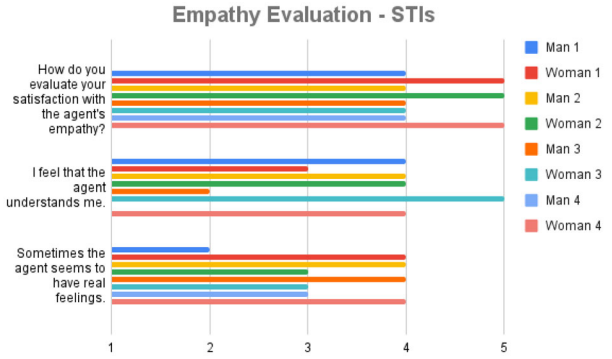

Fig. 8 Average scores of the empathy assessment for the STIs, referring to questions 1–3 in Table 3. The Likert scale was converted to numbers, so very unsatisfied is 1 and very satisfied is 5

Woman 3 (5). For question 3 (Sometimes the agent seems to have real feelings), most participants scored 3 and 4, with Man 1 scoring 2.

In addition, we can calculate the average score of each STI group (with and without the 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 generally seem a bit higher for the Empathy group, they seem to have little impact on the STIs compared to LTI. 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 Fig. 8 and grouped by each group (with and without empathy module), resulting in a p-value of 0.70, which indicates a similarity in the answers of the participants of both groups. We hypothesized that the short-term interactions of 10–15 min 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 who 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 the difference between the two STI groups (6.52%). Here, we hypothesize that this difference is because the conversation content between agent and participants, in STIs, was more neutral and performed in a single interaction. On the other hand, the LTI participants probably had more emotional conversations, further 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 users' engagement could have been reduced after each day of interaction, which would cause a drop in the evaluations. This phenomenon would not occur in STIs because users only interacted once with the virtual agent.

Finally, when we group the participants' results based on their interaction with either Arthur or Bella, we observe that the Bella group obtained an average score of 3.91, whereas the Arthur group scored an average of 3.74. Furthermore, the ANOVA test confirms the consistency in the participants' responses, with the results showing ($F(5.98) = 0.08, p = .77$)). Therefore, the gender and visual attributes of the tested ECAs do not significantly affect the results. Nevertheless, the population who interacted with Bella scored marginally higher values.

4.2 Empathetic memory experiment

This experiment was conducted to measure the impact of the relationship between our ECA's memory and empathy modules on user perception. In our previous work [18] we assessed the impact of memory on the ECA architecture. In this current study, we aim to explore the influence of empathy, which is inherently linked with the presence of memory in virtual agents. Therefore, all interactions analyzed in this paper take into account the existence of artificial memory. We performed it only with 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, all Brazilians). Of these 30 volunteers, 13 are Undergraduates, another 13 are Graduated, 3 completed High School, and one is a high school student. Concerning their experience interacting with virtual agents, 6 participants answered as very low, 9 as low, 9 as regular, 4 as high and 2 as 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, as presented in Table 5.

First, the participants were presented with the consent form approved by the Ethics Committee of University of Pontifical Catholic University of Rio Grande do Sul, referring to the research project entitled "Estudos e Avaliações da Percepção Humana em Personagens e Multidões Virtuais", number 46571721.6.0000.5336. After that, they were encouraged to download the ECA's executable file and freely interact with it for a short time to get used to it. They were also presented with a brief explanation about emotion and empathy and answered the Toronto empathy questionnaire

Table 5 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 the participant's study and work	Happiness
T3	Discover if the virtual agent has any pets and more information about it	Sadness
T4	Discover if the virtual agent remembers any other subject that the participant already spoke with it	Varied

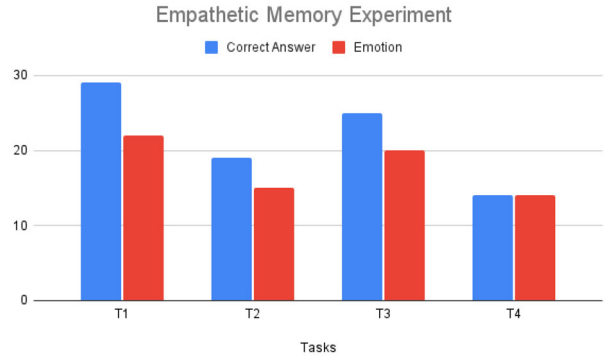

Fig. 9 Scores of the thirty participants from the experiment. "Correct Answer", in blue, refers to the number of people who answered as expected. "Emotion", in red, refers to the number of people correctly identifying the agent's conveyed emotion

[37] (TEQ) to measure their empathy level. The mean score for men is 43.63, while the mean score for women is 47.12, which puts them both inside the average TEQ values.

All tasks presented in Table 5 are related to some data inside agent's memory, with a respective emotion associated. In T1, T2 and T3, participants should find some information saved in ECA's memory related to the fact, e.g., the ECA is happy talking about games (T1). 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 [8] (the ability to understand and react toward the emotion of others), we modeled such emotional memories as an empathetic behavior. Finally, after each task, the participants were asked 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.

Figure 9 presents the scores of the thirty participants from the experiment. "Correct Answer", in blue, refers to the number of people who answered as expected. For instance, in T1, it was expected that the participants were able to discover that the virtual agent enjoys playing video games. "Emotion", in red, refers to the number of people who correctly identified the agent's conveyed emotion. Concerning T1, (Find out if the ECA likes video games and if it has a favorite game), from 30, 29 participants were able to find out that the virtual agent likes video games. Also, 23 participants could identify the agent's favorite game, while 22 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 answered 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), 14 participants reported that Arthur or Bella could remember about some other subject that they chose to speak about and conveyed an appropriate emotion.

The presented results suggest that the participants were generally 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 straightforward tasks than T2 and T4.

As commented before, the evaluated empathy is a Likert scale from 1 (no empathy) to 5 (extremely empathetic). Figure 10 presents the average scores and standard deviations of the evaluated empathy for all four tasks. 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. A possible explanation is that when users interact with the ECA and talk about themselves, there is a more significant variation in answers and questions, so the information saved by the ECA can vary

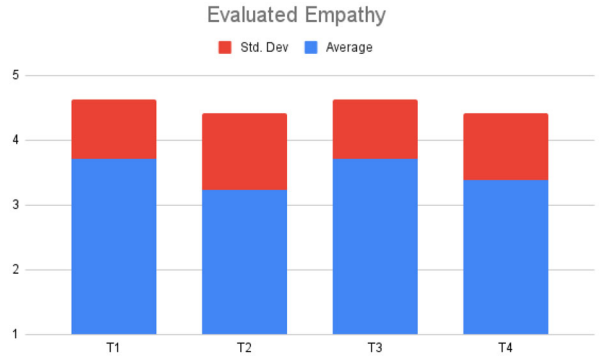

Fig. 10 Evaluated empathy for all four tasks. The average score values were 3.71, 3.23, 3.71, and 3.38 for tasks T1–T4, respectively. The standard deviation values are 0.90, 1.18, 0.9, and 1.02 for tasks T1–T4, respectively

too. On the other hand, when users ask controlled questions about the ECA, the saved information is always the same, so the answers seem more coherent. Free dialogues are expected to be much more difficult to control than controlled ones.

5 Final considerations

In this study, we present a novel embodied conversational agent (ECA) named Bella. Unlike previous ECAs, Bella is equipped with an empathy module that is connected to her memory. Along with Arthur, a multi-purpose agent introduced in our earlier work, Bella is designed without a specific application in mind, allowing for versatile use across various domains. To assess the effectiveness of our proposed approach, we conducted two experiments, involving a total of 42 volunteers. The first experiment included 12 participants, while the second involved a larger sample of 30 participants.

The primary objective of our first experiment was to evaluate the efficacy of our empathy module using both short-term interactions (STIs) and long-term interactions (LTIs) between human participants and our ECAs, Bella and Arthur. The results of this study indicate that while the empathy module was perceived by participants in both types of interactions, it was more effective in LTIs. This finding aligns with our hypothesis that STIs may not provide enough information in the dialogue to activate empathy feelings, as is often the case in real-life interactions. We also compared the user ratings for the two types of interactions: STIs and LTIs. Our hypothesis, which was confirmed by the results, was that STIs have a more positive impact on users, as they limit exposure to potential problems or inconsistencies that might be observable in a long-term interaction. Additionally, our study revealed that the empathy module had a more notable impact in LTIs than in STIs, further highlighting the importance of longer interactions when assessing empathy in ECAs.

For our second experiment, we focused solely on STIs and examined how ECAs' empathetic behavior, based on their memories, impacted the user. Our hypothesis was that when ECAs exhibited empathy associated with memories, participants would be better able to identify such emotions and the agent would behave more naturally. As we discuss in our paper, our results confirm this hypothesis. Specifically, our findings showed that participants were able to elicit the expected memories from Arthur or Bella and accurately identify the conveyed emotion. Additionally, we observed a difference in the results between LTIs and STIs, with the evaluation of LTIs consistently lower than that of STIs. However, we believe that the evaluation of LTIs was more accurate because participants in STIs were not exposed to communication issues for a sufficient amount of time.

While our study provides valuable insights into the use of ECAs with empathy in human–computer interaction, we acknowledge several limitations that need to be addressed in future research. One such limitation is the small number of users who participated in our experiments. However, we argue that LTIs provide more informative results for evaluating ECAs than STIs since users cannot explore all possible dialogues with limited interaction time. Furthermore, while our ECA can start conversations on pre-defined topics, the manual definition of such topics and dialogues poses a significant challenge. Thus, we plan to explore ways to automate this process, which will enable our ECA to initiate more engaging and varied conversations with users.

Our future work will focus on improving the visual behavior and facial animation of ECAs, as well as their dialogues. Specifically, we aim to invest more time in automating the process of defining topics and dialogues for our Embodied Conversational Agent's small talk module. Melgare et al. [24] proposed the concept of "emotion styles," unique ways individuals express emotions. It would be intriguing to investigate whether ECAs (Embodied Conversational Agents) can detect and utilize these styles to display their own emotions and offer empathy. Incorporating this feature may enhance the user experience by making the facial expressions of Arthur/Bella more relatable, potentially leading to increased user comfort.

Funding Soraia R. Musse is funded by CNPq (Grant No. 305084/2016-0).

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bartneck, C., Croft, E., Kulic, D.: Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots (2008)
2. Burleson, W., Picard, R.W.: Affective agents: Sustaining motivation to learn through failure and a state of stuck. In: Workshop on Social and Emotional Intelligence in Learning Environments (2004)
3. 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, pp. 1–3 (2020)
4. Chen, J., Zhang, D., Qu, Z., Wang, C.: Artificial empathy: a new perspective for analyzing and designing multi-agent systems. IEEE Access **8**, 183649–183664 (2020)
5. Conway, M.A., Pleydell-Pearce, C.W.: The construction of autobiographical memories in the self-memory system. Psychol. Rev. **107**(2), 261 (2000)
6. Coplan, A., Goldie, P.: Empathy: Philosophical and Psychological Perspectives. Oxford University Press (2011)
7. Damasio, A.R.: Descartes' error. Random House (2006)
8. De Waal, F.B.: Putting the altruism back into altruism: the evolution of empathy. Annu. Rev. Psychol. **59**, 279–300 (2008)
9. de Waal, F.B., Preston, S.D.: Mammalian empathy: behavioural manifestations and neural basis. Nat. Rev. Neurosci. **18**(8), 498–509 (2017)
10. 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), pp. 138–143. IEEE (2018)
11. Ekman, P.: An argument for basic emotions. Cognit. Emot. **6**(3–4), 169–200 (1992)
12. Goldberg, L.R.: An alternative" description of personality": the big-five factor structure. J. Personal. Soc. Psychol. **59**(6), 1216 (1990)
13. Goldstein, A.P., Michaels, G.Y.: Empathy: Development, Training, and Consequences. Lawrence Erlbaum (1985)
14. 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, pp. 528–533. IEEE (2009)
15. Hojat, M.: Empathy in Patient Care: Antecedents, Development, Measurement, and Outcomes. Springer Science & Business Media (2007)
16. Kagan, J., Snidman, N.: The Long Shadow of Temperament. Harvard University Press (2009)
17. Kasap, Z., Magnenat-Thalmann, N.: Building long-term relationships with virtual and robotic characters: the role of remembering. Vis. Comput. **28**(1), 87–97 (2012)
18. Knob, P., Dias, W.S., Kuniechick, N., Moraes, J., Musse, S.R.: Arthur: a new eca that uses memory to improve communication. In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), pp. 163–170. IEEE (2021)
19. Kshirsagar, S.: A multilayer personality model. In: Proceedings of the 2nd International Symposium on Smart Graphics, pp. 107–115 (2002)
20. Loftus, G.R., Loftus, E.F.: Human Memory: The Processing of Information. Psychology Press (2019)
21. 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, pp. 1–9 (2020)
22. McCrae, R.R., Costa, P.T., Jr., Martin, T.A.: The neo-pi-3: a more readable revised neo personality inventory. J. Personl. Assess. **84**(3), 261–270 (2005)
23. Mehrabian, A.: Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies (1980)
24. 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), pp. 115–124. IEEE (2019)
25. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81 (1956)
26. Milward, D., Beveridge, M.: Ontology-based dialogue systems. In: Proceedings of 3rd Workshop on Knowledge and Reasoning in Practical Dialogue Systems (IJCAI03), pp. 9–18 (2003)
27. Minsky, M.: Society of mind: a response to four reviews. Artif. Intell. **48**(3), 371–396 (1991)
28. Moreno, R., Mayer, R.: Interactive multimodal learning environments. Educ. Psychol. Rev. **19**(3), 309–326 (2007)
29. Morville, P.: Experience design unplugged. In: ACM SIGGRAPH 2005 Web Program, SIGGRAPH '05, p. 10-es. Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1187335.1187347
30. 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, pp. 130–138. Springer (2010)
31. Petit, M., Fischer, T., Demiris, Y.: Lifelong augmentation of multimodal streaming autobiographical memories. IEEE Trans. Cogn. Dev. Syst. **8**(3), 201–213 (2015)
32. Prendinger, H., Ishizuka, M.: The empathic companion: a character-based interface that addresses users'affective states. Appl. Artif. Intell. **19**(3–4), 267–285 (2005)
33. Prendinger, H., Mori, J., Ishizuka, M.: Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. Int. J. Hum. Comput. Stud. **62**(2), 231–245 (2005)
34. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. J. Res. Personal. **11**(3), 273–294 (1977)
35. 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. Entertain. Comput. **32**, 100313 (2019)
36. Spitale, M., Garzotto, F.: Towards empathic conversational interaction. In: Proceedings of the 2nd Conference on Conversational User Interfaces, pp. 1–4 (2020)
37. 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. J Personal. Assess. **91**(1), 62–71 (2009)
38. Tapus, A., Mataric, M.J.: Emulating empathy in socially assistive robotics. In: AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics, pp. 93–96 (2007)
39. Vollberg, M.C., Gaesser, B., Cikara, M.: Activating episodic simulation increases affective empathy. Cognition **209**, 104558 (2021)
40. Wagner, U., Handke, L., Walter, H.: The relationship between trait empathy and memory formation for social versus non-social information. BMC Psychol. **3**(1), 1–8 (2015)
41. 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, pp. 845–853. International Foundation for Autonomous Agents and Multiagent Systems (2016)
42. Yalçın, Ö.N.: Evaluating empathy in artificial agents. arXiv preprint arXiv:1908.05341 (2019)
43. Yalçın, Ö.N.: Empathy framework for embodied conversational agents. Cogn. Syst. Res. **59**, 123–132 (2020)
44. Yalcin, O.N., DiPaola, S.: A computational model of empathy for interactive agents. Biol. Inspired Cogn. Archit. **26**, 20–25 (2018)

Paulo Ricardo Knob is doctor in Computer Science. His topics of interest are crowd simulation, virtual agents and conversational agents.

Natalia Dal Pizzol is computer science undergraduate. Her research interests include human–computer interaction, natural language processing, and data analytics.

Soraia Raupp Musse is associate professor at Pontifical Catholic University of Rio Grande do Sul, where she created the VHLab. Her research interests include crowd simulation and analysis, facial animation and visual perception.

Catherine Pelachaud is Director of Research in laboratory ISIR, Sorbonne University. Her research interests include socially interactive agent, nonverbal communication (face, gaze, gesture and touch), and multiparty interaction.