International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00855-w

Bidirectional Multi-modal Signs of Checking Human-Robot Engagement and Interaction

Umberto Maniscalco[1] · Pietro Storniolo[1] · Antonio Messina[1]

Accepted: 30 November 2021

Abstract
The anthropomorphization of human-robot interactions is a fundamental aspect of the design of social robotics applications. This article describes how an interaction model based on multimodal signs like visual, auditory, tactile, proxemic, and others can improve the communication between humans and robots. We have examined and appropriately filtered all the robot sensory data needed to realize our interaction model. We have also paid a lot of attention to communication on the backchannel, making it both bidirectional and evident through auditory and visual signals. Our model, based on a task-level architecture, was integrated into an application called W@ICAR, which proved efficient and intuitive with people not interacting with the robot. It has been validated both from a functional and user experience point of view, showing positive results. Both the pragmatic and the hedonic estimators have shown how many users particularly appreciated the application. The model component has been implemented through Python scripts in the robot operating system environment.

Keywords Social signs · Human-robot engagement · Interaction model · User experience

1 Introduction

Communication between human beings is a highly dynamic social activity in which at least two subjects must cooperate consciously to generate the meaning of their interaction. This assumption implies a fundamental concept: effective communication and information extraction are two distinct but equally essential phenomena. It is valid even in the presence of non-verbal communication, and it is linked exclusively to the signs [30]. In this regard, communication science has a long tradition of misunderstandings. The most striking is probably the one made by Watzlawick et al. [63], who claim that "one cannot not communicate". This statement implies that we communicate something anyway, whatever behaviour we adopt.

If this is true from a certain point of view, it does not consider the other subject's willingness or ability to extract and give meaning to the signifier. Thus, the willingness and ability to give meaning to the signifiers are essential elements that the subjects involved in communication must put into play.

If, on the one hand, the willingness belongs to the sphere of individual behaviours, the ability turns out to be a more objective and measurable element. Thus, willingness and ability are two fundamental elements that we must consider, even when one of the two interacting subjects is a humanoid robot.

In addition to communication of contents, there is another kind of communication, so to speak, of control. By control communication, we mean the continuous exchange of mainly non-verbal information to establish and maintain engagement between the subjects communicating and dictating the communication times and states.

Think about the walkie-talkie communication, where the channel is half-duplex[1] and there is no non-verbal communication. In this case, the dialogue is a bit unnatural because the subjects need to make explicit signals to control the communication. For example, when an interlocutor has finished transmitting, he says "k" or "kk" to signal that he has completed its transmission and has unlocked the channel.

One of the components of this control communication is the one that goes by the name of backchannel, intro-

✉ Umberto Maniscalco
name.surname@icar.cnr.it

[1] Istituto di Calcolo e Reti ad Alte Prestazioni (Human-Robot Interaction group) - Italian National Research Council, Via Ugo La Malfa, 153, Palermo, Italy

[1] Both parties can communicate with each other, but not simultaneously.

Published online: 07 April 2022

duced by Victor Yngve in 1970 [68] for the first time. The name backchannel (as opposed to the speaker's main-channel) indicates that two communication channels operate simultaneously during a conversation [64]. Through the backchannel, which can be considered a feedback channel, the interlocutor sends back to the speaker a set of verbal and non-verbal signals, thanks to which the speaker can evaluate the progress of the dialogue [40].

Very significant examples of vocal backchannel are short words like *yeah*, *mmm*, *uh-huh* used by the listener during dialogue to show attention to the speaker. The absence of these signals makes the dialogue unnatural, so much so that a well-known advertisement from a voice assistant has used them (although this feature is not implemented) to surprise the public[2]. Indeed, the current voice assistants are an example of unnaturalness in dialogue. In fact, not being equipped with sensors capable of perceiving non-verbal communication requires using a wake-up word every time we want to ask a question.

Considering what has been said, it is quite natural that, in human-robot interaction, scientists try to emulate the same mechanisms of interaction between humans to make the interaction as natural as possible. It implies that the robot should have both the willingness and the ability to decode non-verbal communication. We assume that the robot also knows how to interpret and reproduce natural language. Besides, the robot must also produce both verbal and non-verbal communication on the main-channel and the backchannel. The communication that goes from the robot to the listener is also crucial because it allows the interlocutor to understand the robot's states, manage the conditions of engagement, and communicate the appropriate feedback.

The term engagement typically refers to a relationship between individuals that has the character of stability and durability. The word engagement is also used widely in the robotics field, where it concerns the human-robot interaction as well as for the first time defined by C.L. Sidner et al. in [58]. They represent the engagement as: "*the process by which individuals in an interaction start, maintain and end their perceived connection to one another*".

The concept of engagement is defined in [20], where it is thought of as a binary concept. That is, two subjects are considered to be fully engaged or not engaged. In reality, this point of view can be limiting in some circumstances. The conditions in which to determine whether an engagement is determined between two or more participants can be various.

For example, the number of subjects considered in the engagement process can influence how it is defined. We leave out the classical situation in which only two subjects are considered, and we think a group involving some subjects. In this case, the behaviour of each of the subjects that make up the group can vary over time concerning the so-called "affiliation". The affiliation [9] represents the role acknowledged for each individual who constitutes the social group.

When one member of the group is the chairman and the others are spectators, the verification of the specific conditions of engagement is less severe for the discussion's conduct. The speaker does not need to check that all spectators are engaged while continuing his communication. Nevertheless, every onlooker must be somehow engaged in following the speech. Instead, if subjects of a group are on an equal footing, as friends chatting with each other, speakers and listeners' affiliations will vary with time and then will the engagement conditions. In both these circumstances, the engagement's continuity does not constitute such a determining element for communication. Any subjects of the groups could be distracted without thereby losing the fundamental requirements for communication.

Furthermore, another fundamental aspect of the interaction, which might seem trivial, but not at all, is to be sure that you are talking to the desired interlocutor. This issue takes on even more critical when one of the two interlocutors is a robot.

In this work, we will focus on the latter case and define a model based on bidirectional multi-modal signs of checking human-robot engagement and interaction.

The anthropomorphization of the interaction between human and robot cannot be based only on vocal interaction. Visual, auditory, tactile, proxemic, and other aspects must be considered and integrated to manage the interaction. This article considers some of these aspects (see Sects. 2.1, 2.2, 2.3 and 2.4), describing how they individually contribute to improving the interaction between humans and robots. We have also paid a lot of attention to communication on the backchannel, making it evident through auditory and visual signals.

The model has been implemented through Python scripts in the robot Operating System (ROS) environment[3] and has been successfully tested in the real world through W@ICAR (Welcome To Istituto di CAlcolo e Reti ad alte prestazioni). It is an application for an unedited and appealing experience that guides the visitors to discover our Institute and the research activities we conduct. The robot guide knows how to identify the visitant, accompanying him/her on tour, capturing their emotional signals, and showing additional multimedia content thanks to its display. The robot understands the user's natural language questions (in this case, Italian) and provides answers based on his previously created knowledge. The robot can profile the user and capture the visitor's emotional state, interests, and knowledge. This way, it builds personalized experiential itineraries.

[2] https://www.youtube.com/watch?v=yv_8dx7g-WA.

[3] http://wiki.ros.org/.

The next section shows what sensory data the robot uses to manage the phases of communication. In Sect. 3, we describe how sensory data can be merged in a suitable model and used to verify the conditions of the engagement and its persistence. Section 4 reports some details about the ROS implementation. Section 5 describes the results of the system validation. Section 6 reports conclusions and some notes on future developments.

2 The Multi-Modal Signs

Referring to relationships between humans, each individual has his model of reference for interpreting the signals that coordinate communication, through which he deduces if his interlocutors are attentive and follow his speech [35,48,57].

This model is not the same for all individuals. For example, it may be influenced by cultural or geographic aspects [25]. Moreover, this model may also slightly vary in the individual, depending on the social circumstances. Despite this variability, it is always based on a composition of some elements like facial recognition and expression, body gesture, voice, distance and more.

One of the most significant aspects that humans consider during a social engagement and interaction is a non-verbal behaviour based on face-to-face interaction, through which they communicate quite a lot about purpose [15,52].

Other fundamental concepts that humans use to manage social engagement are related to visibility (e.g., facial recognition and expression, body gesture), audibility (e.g., voice, tone, sound) and the social distance that separates the interlocutors [26,55].

Thus, humans decide, from time to time, both if engagement exists and persists and the different states of interaction based on multi-modal information composition.

In human-robot interaction, we should try to reproduce the human model in the robot to make the communication as similar and natural as possible. Then, we have to arrange the sensory data in a suitable model manageable by the robot. Furthermore, we should try to make visible in the robot the non-verbal signals that we usually perceive in our interlocutor.

We wish to underline here we consider anthropomorphic robots or a robot with anthropomorphic capabilities. Thus, the robot has auditory and visual abilities and it can also measure, in some way, the distance that separates itself from objects.

Let us consider Pepper and Nao humanoid robots by SoftBank Robotics[4], used in our experiments. These kinds of robots have almost all the needed capabilities to design an anthropomorphic interaction model. More specifically, the ability to measure the distances between oneself and objects is entrusted to sonars and precisely the one in the front position. The vision skills are made possible by the RGB camera and the audio-related abilities are made possible by the presence of microphones and speakers.

Besides, the Pepper robot has a tablet used to transpose visual information both of content and control. Also, Pepper can modify the individual LED segments of one's own eyes to create animations. This feature is used together with others to enrich the robot's non-verbal communication.

2.1 Visual Information

The visual information, which, for example, can be acquired utilizing an RGB camera, is of fundamental importance for the management of the interaction by a robot. Much of the non-verbal information flows on the visual channel.

Concerning the information produced by the user and perceived by the robot, we can highlight, for example, the presence or absence of one or more human beings in front of the robot. It is required to try to recognize the face [16] and assign a name or an Id and determine the gaze direction (In this work, Boolean information is used to indicate if the user is looking at the robot in the eye or not). The gaze's focus has high value both as a social signal and an element of synchronization of the conversation [2].

The robot can also use non-verbal communication, producing signals that humans can use to understand what the robot is doing. We have used the robot's tablet to communicate visual information about the robot's status and activities. Animation created using the LED segments of the robot's eyes also help to give visual information about the robot's activities. Moreover, the colour of the eyes was used to communicate different information. The eyes colour and animation make the human-robot interaction more natural regardless of the information they transmit [10,50].

Therefore, we can conclude that the visual channel is used in a bidirectional way: both the robot and the human acquire and produce information that flows on this channel.

Let's consider the visual information that goes from the human to the robot. From a theoretical point of view, it would be quite simple to merge this data to say whether a specific user is in front of the robot and is looking or not at it. However, if we consider the variability over time of this basic information and their noisy nature, their composition produces an even more variable and noisy result.

To overcome the variability and noise of the data, we can consider and evaluate appropriately, for each kind of information a time series of values organized into a First

[4] https://www.softbankRobotics.com/us/Robots.

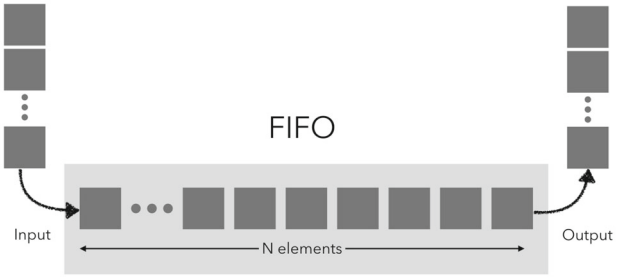

Fig. 1 The generic structure of a FIFO queue. The first element to enter the queue is also the first to exit once the queue is filled

Input First Output (FIFO) queue (see Fig. 1) instead of the instantaneous values. In doing so, we replace the instantaneous values with suitably smoothed values. We will call the values obtained from the analysis of the time series V_{gFIFO}, V_{pFIFO}, V_{idFIFO} respectively for the direction of the gaze, the presence of a person and the recognition of a face.

If sensory data are sampled at a specific frequency $f(Hz)$ and, for example, with a queue of n samples, we get a time window of $t = n/f$. Having established the information's sampling frequency, we can use a larger or smaller number of samples to stabilize our measurements.

In the case of the assessment of the direction of the gaze, for example, by calculating the average of the samples $g_m = \frac{\sum_{i=1}^{n} g_i}{n}$, where g_i is the $i\text{-}th$ instantaneous gaze value, and given a threshold t_r, it is possible to attribute an overall value to the samples contained in the queue according to the Formula (1). The value of t_r in the range [0; 1] establishes how stable the value of the queue's content must be to give the queue value of 1 or 0.

$$V_{gFIFO} = \begin{cases} 1 \text{ if } g_n \geq t_r \\ \\ 0 \text{ if } g_n < t_r \end{cases} \text{ where}$$

$$n = \text{ number of elements} \tag{1}$$

$$t_r = \text{ threshold}$$

$$g_i = \begin{cases} 1 & \text{if } \textit{gaze towards the Robot} \\ 0 & \text{if } \textit{not gaze towards the Robot} \end{cases}$$

Similar considerations can be made in the case of the evaluation of the presence/absence of a person in front of the robot. In this case, by calculating the average of the samples $p_m = \frac{\sum_{i=1}^{n} p_i}{n}$, where p_i is the $i\text{-}th$ instantaneous presence/absence value and given a threshold t_r, it is possible to attribute an overall value to the samples contained in the queue according to the Formula (2).

$$V_{pFIFO} = \begin{cases} 1 \text{ if } p_n \geq t_r \\ \\ 0 \text{ if } p_n < t_r \end{cases} \text{ where}$$

$$n = \text{ number of elements} \tag{2}$$

$$t_r = \text{ threshold}$$

$$p_i = \begin{cases} 1 & \text{if } \textit{presence} \\ 0 & \text{if } \textit{absence} \end{cases}$$

In the case of face recognition, instead, the samples are given by the ID (or by the name) of the recognized face. Therefore, the formula must be reinterpreted, assigning an ID to the face in front of the robot if, within the queue, $MaxEqID(ID_i)$ (maximum of amount of instances of equal ID) divided by n overcomes a t_r value. Otherwise, the formula returns "Unknown" as shown in Formula (3).

$$V_{idFIFO} = \begin{cases} \text{ID} & \text{if } \frac{MaxEqID(ID_i)}{n} \geq t_r \\ \text{Unknown} & \text{if } \frac{MaxEqID(ID_i)}{n} < t_r \end{cases} \tag{3}$$

In other words, we evaluate if the same face has been recognized a sufficient number of times to affirm that, during the time window covered by the FIFO queue, the identified person is always the same.

Let's now consider the visual information that goes from the robot to the human. As previously mentioned, the Pepper robot has a tablet that is used to transpose visual information. We have used this visual device to communicate to the interlocutor the different robot's states. All the images displayed on the tablet are animated GIFs and here is shown just a significant frame for each one. In picking the animations, we have chosen widely consolidated visual metaphors [14,21] in the field of human-computer interaction.

This non-verbal communication made by the robot is essential because it dictates the timing of interaction with the human. As described in the next section, the robot is governed by a finite-state automaton in our model. The robot's ability to express its state makes the interlocutor conscious.

Figure 2 shows two different "waiting for" states of the robot. On the left side, where the typical waiting circle is grey, it communicates that the robot is waiting to meet an interlocutor. Instead, when the tablet shows the red waiting circle, it communicates that it has identified a possible interlocutor. Still, the engagement conditions have not yet been verified (see the next section).

The microphone with the green bullet shown on the left side of Fig. 3 (in the animated version, the bullet blinks green) communicates to the interlocutor that he can start speaking. Therefore the conditions of engagement have been verified. The red microphone is shown on the right side of Fig. 3 (in the animated version, it blinks red), replaces the green

Fig. 2 Two different robot waiting states. The grey waiting circle (left) communicates that the robot is waiting to meet someone. The red waiting circle (right) indicates that the robot has identified a possible interlocutor, but he is not yet engaged

Fig. 3 The microphone with the green bullet communicates to the interlocutor that he can start speaking. The red microphone indicates that the robot is listening to him

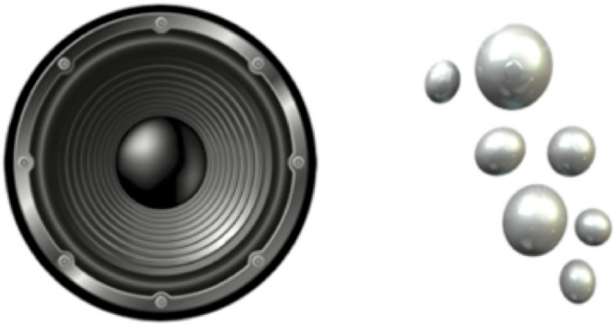

Fig. 4 The speaker used when the robot is speaking on the left side. A symbolic set of spheres represents the moments of elaboration on the right side of the figure

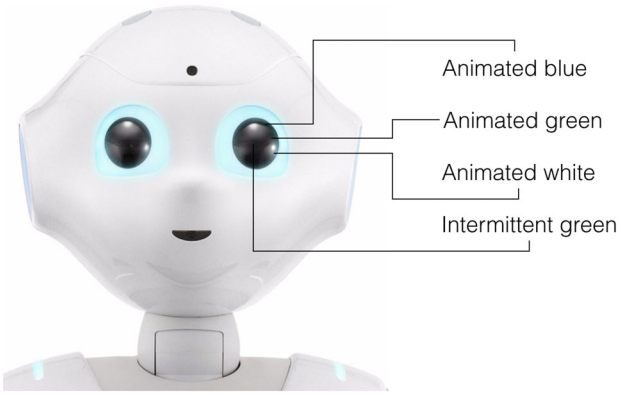

Fig. 5 The four different ways in which the robot's eyes are animated to enrich non-verbal communication

microphone as soon as the interlocutor starts talking; thus, it communicates that the robot is listening to the talker.

Figure 4 show two other animated GIFs, completing the robot's non-verbal communication image set. The left side shows a speaker the robot displays on his tablet in an animated version when he starts talking. The right side shows the image used to communicate to the user that the robot is meditative. That is a state in which it is processing information to find answers to user requests.

The Pepper robot can also modify the individual LED segments of its own eyes to create animations. The ability to animate the robot's eyes has been exploited in two ways. The first one is used to improve the robot's facial expressiveness, which is an essential feature in the field of human-robot interaction [1,12,27]. The eyes of the robot are animated to give the idea that it is blinking. This animation is always used, except when the eyes are in intermittent green mode. It does not communicate changes in the robot's status, but it aims to make the robot's face more natural and increase the interlocutor's trust. The second way is to enrich the non-verbal communication that the robot can produce on the main-channel and the backchannel. In addition to what the robot communicates via the tablet, as shown in Fig. 5, the colour of the eyes is used to indicate to the interlocutor four different states of the robot:

- Animated white, when the robot communicates that it is waiting for an engagement with an interlocutor;
- Animated green, when the robot announces that the conditions of engagement have been verified with an interlocutor and that the engagement has started;
- Animated blue, when the robot indicates that it is responding to the interlocutor;
- Intermittent green, when the robot indicates that all engagement conditions have been verified but the social distance is still too high.

2.2 Proxemics Information

As previously mentioned, the social distance between two interlocutors is also an essential element in determining whether an engagement exists. We assume that most people keep the same distances when interacting with each other and when interacting with a humanoid robot [62]. The robot

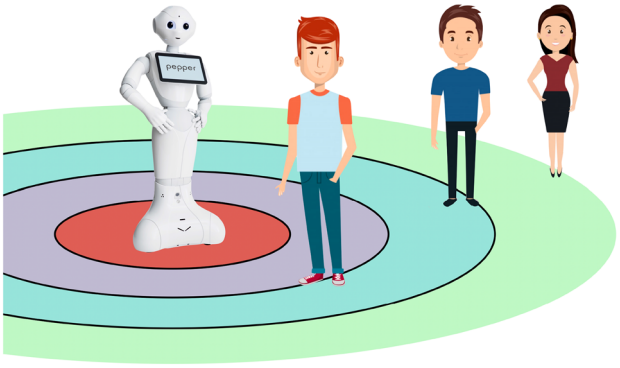

Fig. 6 The figure shows the different social distances identified by Edward T. Hall: intimate space, personal space, social space, and public space

measures distance either via lasers or sonars. Having the latter a wider cone of irradiation, they are generally employed to measure distances even from moving objects.

Sonar measurements are often noisy and not very precise, so, even in this case, it is necessary to proceed with a filtering operation before using them to determine the social distance of an interlocutor. Also, in this case, the distances' instantaneous values are not considered because they are replaced by the median of the content of a FIFO queue of distance values.

Formula (4) shows that, as in the case of the previous measurements, a FIFO queue can be used to stabilize the distance measurement, evaluating if a sufficient number of measures in the FIFO is less than the established social distance.

Here, formula (4) returns 1 if the interlocutor is at a distance (d_i is the $i-th$ instantaneous distance value) less than or equal to t_r, returns 0 if the interlocutor is more distant than t_r.

$$V_{dFIFO} = \begin{cases} 1 \text{ if } \left(\sum_{i=1}^{n} d_i \leq S_d\right) \geq t_r \\ 0 \text{ if } \left(\sum_{i=1}^{n} d_i \leq S_d\right) \leq t_r \end{cases} \quad \text{where} \tag{4}$$

$n =$ number of elements
$S_d =$ social distance
$t_r =$ threshold
$d_i =$ measured distance

2.3 Auditory Information

Auditory information is essential and it alone is often enough to establish whether there is an engagement between two (or more) individuals. Think, for example, of a telephone conversation; as long as the audio channel carries information between one subject and another, we can say that there is

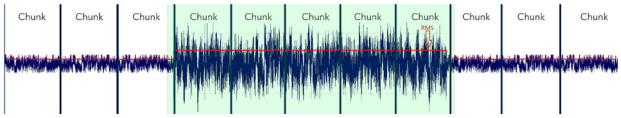

Fig. 7 The audio signal divided into chunks. The red line indicates the RMS of the power signal for each chunk. The area highlighted in light green indicates the active audio channel

an engagement. Conversely, prolonged silence will arouse suspicion in one of the two interlocutors that the engagement is, for some reason, concluded.

In our interaction model, we use a dual audio channel to establish the conditions of the engagement. The first audio channel uses a matrix of 4 microphones which allows locating the direction of the sound's origin to the robot frame [60], even in noisy environments [59]. This channel is used to attract the robot's attention through auditory signals, and therefore, it can be considered, as we will see, a proper tool to achieve engagement. The second audio channel allows real communication between the human and the robot. In this case, we take into account the analysis of the power in an audio signal.

Many of the audio formats such as AVI, ANI and WAV are based on the [5]. The basic building block of a RIFF file is called a chunk. For each chunk into which the audio stream is divided, the power's root-mean-square (RMS) is calculated. If it exceeds a certain threshold t_r, the robot considers its interlocutor speaking to it. Similarly, if after the activation of the audio channel, the RMS of the power of the chunks turns under the established threshold for a specific time t, then the robot considers that its interlocutor has stopped talking.

We build the wave file to be sent to google's speech to text service by collecting the consecutive chunks with an RMS of the power greater than a given threshold. We calculate this threshold experimentally considering an average noisy environment. Moreover, among the engagement's conditions, we involve the user proximity to the robot. All this allows us to be quite certain that the chunks with RMS of the power greater than the threshold contain the user's voice and not just the noise. The file we build also contains noise beyond the user's voice. However, this does not affect the recognition of google's speech to text service.

Figure 7 shows the audio signal divided into chunks and the red line indicates the RMS for each fragment. The area highlighted in light green represents the area in which the robot considers the audio channel active. This area is larger than the one in which the RMS remains over a certain threshold. An earlier part is added to this portion so as not to miss the beginning of the conversation (otherwise, the first chunks that allow you to check the condition $RMS > t_r$ would be

[5] https://www.aelius.com/njh/wavemetatools/doc/riffmci.pdf.

lost). Furthermore, the active portion does not end immediately when the RMS drops below the threshold, but only when the condition lasts for a suitable time. This way, the user can take natural pauses during his speech without the robot interpreting them as the end of the speech [34,67].

The audio channel is used by the robot once again for enriching communication. The robot emits a sound like a beep every time it considers the user's speech finished. This way, the robot signals to the user it finished the listening phase and that the reasoning phase has started.

2.4 Body Movement and Posture

One way that we humans use to show attention to their interlocutor is to maintain face-to-face contact. It implies that when our interlocutor moves into space, we automatically follow his movements [19,36]. Exclusively from a postural and movement point of view, this type of behaviour, in the field of human-robot interaction, is known as face tracking [31,38]. The robot always tries to keep the face at the centre of its field of vision by suitably moving the head or the whole body. In the Pepper robot case, we use its features to create effective behaviour to ensure the most natural face-to-face interaction possible. Pepper can maintain eye contact with the following movements[6]:

- Just the head;
- The head and the rotation of the body;
- The whole body, without rotation;
- The head and autonomously performs small moves such as approaching the tracked person, stepping backwards, rotating, etc.

This last mode is the most appreciated by many users who interacted with the robot through W@ICARR.

The small movements of the robot's advancement, when the user moves away, and those of the robot's backward movement, when the user approaches, contribute to enriching non-verbal communication and transmit to the user the robot's awareness about social distance.

Furthermore, the robot can be configured to have different behaviours about the type of engagement:

- When the robot is engaged with a user, it can be distracted by any stimulus and engages with another person;
- As soon as the robot is engaged with a person, it stops listening to stimuli and stays engaged with the same person. If it loses the engaged person, it will listen to stimuli again and may engage with a different person;
- When the robot is engaged with a person, it keeps listening to the stimuli, and if it gets a stimulus, it will look in its direction, but it will always go back to the person it is engaged with. If it loses the person, it will listen to stimuli again and may engage with a different person.

Again, we have established that this latest behaviour seems to be the one that users most appreciate, making interaction more natural.

However, our interaction model can be configured differently, as will be described in Sect. 4, for each interaction session through appropriate parameters.

We use another interesting basic feature of the Pepper robot: the micro-movements of breathing. These movements mean that the humanoid is perceived as alive (or, in any case, active) even when it is not performing any evident task (i.e., when it is listening or thinking).

3 The Robot Model of Interaction

In Sect. 2, we introduced the sensory data employed in the interaction model between humans and robots. Here, we show how these data are merged to manage the engagement and, more generally, the interaction. Besides, we explained how these information sources are processed and treated to make them easily usable for managing the interaction and communication between humans and robots. We said the sensors' raw data are inherently noisy and unstable. For this reason, we have introduced the V_{*FIFOs} (the * replaces subscripts used in Eqs. (1–4). They allow stabilizing the measurements of the sensors by operating the appropriate averages for each type of data. This information filtered by the V_{*FIFOs} is used to determine the model's state transitions represented in Fig. 8. From an implementation point of view, the measurement of each type of sensory data occurs asynchronously by exploiting the ROS topic mechanism, as explained in the next session.

The proposed model is based on the finite state automaton represented in Fig. 8. It is general and, therefore, can be customized for different applications. Furthermore, it can be easily scaled, adding other sensory aspects if necessary. However, the model presented here forms a perfectly functional core.

In the next section, you can find some implemented details of both the automaton and the ROS topics that compute and communicate the sensory information used to evolve the automaton.

The robot is initially in its resting state ***wait***. In this state, the robot, while active, has his eyes off, and his tablet shows the classic animated waiting icon (see the left part of Fig. 2). From this state, the robot tends to get to the ***engaged with known*** state to start an iteration with someone it knows.

[6] http://doc.aldebaran.com/2-5/naoqi/interaction/autonomousabilities/albasicawareness.html?highlight=fullyengaged.

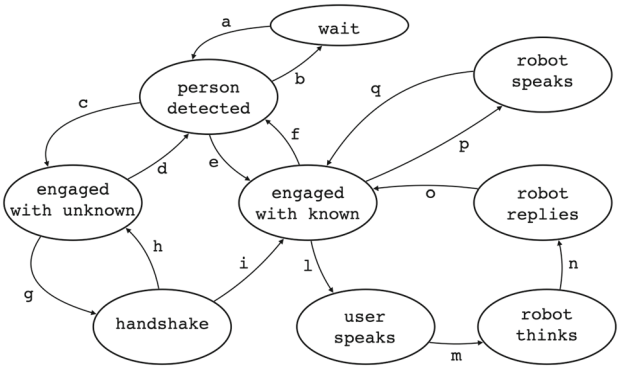

Fig. 8 The model based on the finite-state automaton. The model constitutes a perfectly functional core, and it can be scaled and customized for different applications

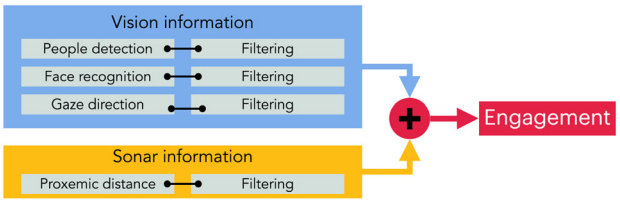

Fig. 9 The figure shows the sensory data taken into account for the initiation and verification of the continuation of the engagement

When the robot detects a person's presence, its status changes, by the transaction a becoming ***person detected***. The transition a is determined using only the values of the V_{pFIFO} described in Eq. (2). Associated with the ***person detected*** state is non-verbal communication. The robot's eyes light up and begin to blink, simulating the eyelids' movement in the mode of animated white (see Fig. 5). The image on the tablet changes and becomes that of the right side of Fig. 2, the face tracking begins, and the robot makes all the appropriate movements to follow the person's face.

Now, having the robot detected a person if he is close enough, the Eq. (4) returns a value 1 and if he is looking at the robot in the eyes the Eq. (1) returns a value 1. So, according to the Eq. (3) returns an "user ID" or "Unknown", we will get the transition e reaching the state ***engaged with known*** or c reaching the state ***engaged with unknown***. Figure 9 schematically summarizes what has been said. Distance, facial recognition, person detecting and gaze direction are the variables involved in determining engagement. The face recognition result determines if the engagement takes place with a known or unknown person.

In the case of ***engaged with unknown***, the robot begins a handshake phase with the user. In this phase, the robot takes the initiative by telling the user that he does not think he knows him and invites him to say his name. After the user says his name, the robot repeats this name and asks for confirmation that the name he understood is correct. During this phase, the robot acquires the user's facial features and stores them in a user database for future recognition. At the same time, the robot gets information about the user's gender and age. Both of these two pieces of information are used by the conversational agent dealing with the dialogue. Knowing the user's sex allows differentiating between male and female, some sentences addressed to the user. The estimate of the user's age instead, in our application, allows formulating simple answers for children/teenagers and more complex explanations for adults. The robot concludes this phase with a pantomime simulating the gesture of taking a photograph to remember the user. During the experiments, we noticed that users appreciate and are amused by this simple gesture.

The handshaking phase can be more or less complex than the one just described for our application to obtain user profiling based on the specific application. This step is skipped entirely if the application does not require profiling. The handshake state is represented here in atomic form as a single state.

When the robot is in the ***engaged with known*** state, there are two ways of interacting. The robot takes the initiative autonomously and says something to the user or the user asks the robot a question. In the first case, the p transition occurs and it brings the robot to state ***robot speaks***. At the end of the robot's speech, the transition q takes place and the robot goes back to the state ***engaged with known***. In addition to verbal communication, non-verbal communication is also used. The ***engaged with known*** state is characterized by the green blinking eyes (see Fig. 5) and an animated microphone with a blinking bullet (see the left side of Fig. 3), indicating that the user can speak if he wishes. When the robot starts talking and its state changes, its appearance also changes. The eyes become blinking blue (see Fig. 5) and the animation of a speaker appears on the tablet (see the left side of Fig. 4).

In the second case, when the user asks the robot a question, the succession of changes of state $l \rightarrow m \rightarrow n \rightarrow o$ occurs. The robot, depending on the events, crosses the states ***user speaks***, ***robot thinks***, ***robot replies***, to finally return to the state ***engaged with known***.

When the user starts talking to ask the robot a question, the RMS of the audio signal's power exceeds the stability threshold. As described in the Sect. 2.3, the robot starts to record the user message. The image shown on the tablet changes, becoming the flashing red microphone (previously, it was the microphone with the flashing green bullet), indicating that a listening phase has begun.

When the user finishes his question, once again, following what is described in the Sect. 2.3, the status of the robot changes, becoming ***robot thinks***. The image shown in the tablet changes again, becoming that of the right side of Fig. 4 and the eyes go back to becoming animated white. The robot also emits a beep to emphasize the change of state.

In this state, the system performs some actions:

- It sends the recorded wave file to google's speech to text service and obtains a string;
- It encapsulates this string in a JSON structure which also contains the user's name, age, sex and other information related to the user profile got by the handshaking phase[7];
- It sends the JSON structure to a conversational agent and receives another JSON structure that contains the answer to the user's question;
- It decodes this last JSON structure and extracts the phrase (a string) to say to the user.

At the end of the described actions, the robot is able to respond adequately to the user and its status becomes ***robot replies***. Now, the robot's eyes become animated blue, and the tablet's image becomes as shown in the left part of Fig. 4. The robot pronounces the appropriate answer [13,49]. At the end of the response, the *o* event is determined and the robot returns to the state ***engaged with known***.

In this section, some of the previous state's return events have not been described for reasons of brevity. Moreover, all the error conditions that the implemented model manages have not been highlighted.

The model presented is very scalable and allows the easy integration of other fundamental aspects in the human-robot interaction. For example, we are currently integrating the understanding of the deictic gesture [54] in the introduced model. We imagine a scenario where the human being interacts with the robotic agent in natural language, and he can also indicate the objects he intends to refer to.

By recognizing the "stroke hold" of the deictic gesture, the robot can understand some descriptive phrases in which a gesture describes something [18,37]. Humans sometimes substitute descriptive words with gestures because they presume listeners will understand the meaning by integrating visual information.

In this case, the robot recognizes the object or subject indicated by the user. It replaces the pronoun used in the sentence with the entity's name referred to before requesting the sentence's understanding from the conversational agent and the appropriate response.

4 The ROS Implementation

The interaction model described in Sect. 3 has been validated through a complex application called W@ICAR. The software is available at GitHub https://github.com/hri-cnr-lab/w_icar and Zenodo https://doi.org/10.5281/zenodo.5144893

W@ICARR is a ROS-based software. ROS (Robot Operating System) is a robotics middleware for robot software development. It is a language and platform-independent framework that allows low-level device control, message-passing between processes, and package management.

W@ICAR has the typical structure of a ROS-based software project. It is a package of Python scripts that implements the nodes of the software architecture. W@ICAR consists of 2 modes. The first is called engagement and manages all sensory information. The second has the application's name and manages all aspects of the actual interaction by implementing the finite-state automaton. The ROS nodes are processes that perform the computation. Nodes are combined into a graph and communicate using streaming topics, RPC services, and the parameter server. In the case of W@ICAR, only the topics are used to generate and exchange information. A launch file is associated with each node. It allows you to start the node and parameterize it appropriately to obtain the desired behaviour.

The finite-state automaton that manages the interaction described in Fig. 8 has been implemented through the SMACH package[8]. It is a task-level architecture for rapidly creating complex robot behaviour based on a Python library.

Here is the code fragment that implements the task-level architecture where the match between the code and the Fig. 8 is evident.

```
with sm_top: smach.StateMachine.add('Wait',
    Wait(),
transitions={'human_presence':'PersonDetected
    ','end':'stop'})
smach.StateMachine.add('PersonDetected',
    PersonDetected(),
transitions={'recognized':'EngagedWithKnown','
    unrecognized':'EngagedWithUnknown','
    human_absence':'Wait'})
smach.StateMachine.add('EngagedWithKnown',
    EngagedWithKnown(),
transitions={'listening':'UserSpeaks','speaking
    ':'RobotSpeaks','error':'PersonDetected'})
smach.StateMachine.add('EngagedWithUnknown',
    EngagedWithUnknown(),
transitions={'handshake':'Handshake','error':'
    PersonDetected'})
smach.StateMachine.add('UserSpeaks', UserSpeaks
    (),
transitions={'done':'RobotThinks','error':'
    EngagedWithKnown'})
smach.StateMachine.add('RobotSpeaks',
    RobotSpeaks(),
transitions={'done':'EngagedWithKnown'})
smach.StateMachine.add('Handshake', Handshake()
    ,
transitions={'eng2know':'EngagedWithKnown','
    eng2unknow':'EngagedWithUnknown'})
smach.StateMachine.add('RobotThinks',
    RobotThinks(),
transitions={'done':'RobotReplies'})
smach.StateMachine.add('RobotReplies',
    RobotReplies(),
transitions={'done':'EngagedWithKnown'})
```

The individual elements of the interaction, or part of them grouped by functionality, are implemented through ROS

[7] The conversational agent uses this information to organize an appropriate response to both the question and the profile information, for example, by building a simplified answer in the case of children or one more articulated in the case of adults.

[8] http://wiki.ros.org/smach.

nodes. Each node can publish or receive messages from other nodes through the topic mechanism. For example, considering Fig. 9, we developed one topic for each sensory information involved in the engagement.

The three topics publish processed and filtered (see Eqs. 1, 2 and 3) information relating to face detection, face recognition and the direction of the user's gaze. A topic publishes processed and filtered (see Eq. 4) proxemics information about the user's distance from the robot.

Indeed, the software also contains four other topics similar to the previous ones. These other four topics publish the sensory information without the filtering operated by the respective V_{*FIFOs}. They are not used in the final application but allowed us, as explained in the next session, to estimate the improvement in performance due to the introduction of V_{*FIFOs}.

Through the mechanism of topics, the node that manages the application can continuously read the information needed to verify the beginning and the maintenance of the engagement between robot and user.

5 Model Testing and Validation

Current research work rarely addresses AI software testing problems. Various articles discuss data quality and assurance in the literature [11,22,66], but rarely researches focus on validation for AI software from a function and feature view. In [24] is widely discussed what AI software testing should be and why.

We use various approaches to test and validate our model. It has been examined as a white box for all aspects of the software code. Obviously, this type of verification is not reported in this article as it does not have any noteworthy research content. Instead, every single functional aspect has been verified, considering the model a black box. Considering that the system was implemented in a ROS environment, it was natural to analyze the individual functions by testing the respective ROS topics that implemented them.

Furthermore, being a model implemented in an application with which thousands of users interacted, it was also evaluated from the User Experience point of view, referred to below UX [29,41,53,61]. It should be emphasized that UX focus on the interaction between human and robot and not on the robot's behaviour and functionality. Certainly, in this last case, the evaluation concerns the user experience with the application of which the interaction model is only a part, even if it is dominant.

5.1 Functional Aspect Evaluation

In this section, only a part of the tests that have been carried out is reported. For reasons of brevity, we will refer to the sanity test, the integration test and the system test used to validate the functional aspects of the model [8,32].

In the sanity test, we focused on the engagement's functionalities (see Fig. 9). We have analyzed the components individually, including the filtering operations described in the Sects. 2.1, 2.2, 2.3 and 2.4.

People detection, face recognition, gaze direction and proxemic distance estimation, in controlled conditions (e.g., a laboratory), reach performances very close to 100% correct operation. However, in an uncontrolled environment, we report a slight degradation of performance concerning the person's recognition functions and estimation of the gaze's direction. These degradation of performance are often due to poor lighting conditions. In particular, back-light conditions are those that cause the worst degradation. Proxemic distance estimation and people detection continue to perform well, even in an uncontrolled operating environment.

We have also validated the use of V_{*FIFOs} by comparing the results obtained with and without their use. The results are significantly different. The use of V_{*FIFOs} makes the engagement condition much more stable than what happens without their use.

Table 1 shows the results of experiments conducted to evaluate the improvements introduced by the V_{*FIFOs}. We considered the features of people detection, face recognition, gaze direction and proxemic distance estimation. For each of them, we have performed measurements to verify the percentage of correct functioning with (columns marked with a two asterisks) and without (columns marked with a single asterisk) V_{*FIFOs}.

The quantities involved to determine the engagement are four (see Fig. 9 and they are sampled at $3Hz$. If we consider the probability that one of them produces an incorrect value, it is easy to understand how engagement verification becomes unstable. The instability does not depend only on a measurement error. A slight distraction of the user may cause it. If the user distracts his gaze for a moment (remember that the gaze direction is sampled at $3Hz$), in the absence of the V_{gFIFO} there would be an immediate loss of engagement. The same happens if the user has a borderline position and the proxemic values become unstable.

Correct functioning performances have been found in close to 90% of cases concerning the sound detection and audio segmentation function. The most frequent causes of malfunctions were due to incorrect segmentation due to the user's excessive pauses in the sentence's pronunciation. Results about the speech to text functionality are not reported because it is provided by Google Cloud Speech API[9] and therefore external to the system..

[9] https://cloud.google.com/speech-to-text.

Table 1 Comparison of positive results with (columns **) and without (columns *) the use of V_{*FIFOs}

Seconds of measure	Number of experiment	People detection		Face recognition		Gaze direction		Proxemic distance	
		*	**	*	**	*	**	*	**
10 seconds	50	96%	100%	92%	99%	74%	100%	100%	100%
20 seconds	50	94%	100%	90%	98%	68%	98%	98%	100%
30 seconds	25	90%	100%	86%	98%	58%	98%	94%	98%

The integration tests and the system test produced good results, not showing any deterioration in performance due to integrating the individual functions.

5.2 UX Evaluation

Since the UX design cycle is intrinsically iterative, often described as UX wheel [28], the results reported here are to be considered cumulative, including the changes that have been gradually made to the application based on previous experiences.

At the end of their experience, the users filled out a short questionnaire consisting of a few but precise questions to evaluate the model's essential elements. In addition to the positive aspect of the experience, the questionnaire assessed both pragmatic and hedonic aspects.

The subjects involved were students, undergraduates, doctoral students, researchers from other institutes and people who participated in conferences or events held at our office. In about two years, 501 people interacted with the robot and filled in the questionnaire [39], but only 467 filled the questionnaire in a useful way for the evaluation. The evaluation group consists of 203 women and 264 men of predominantly young age. Most of them were familiar with new communication technologies, especially conversational.

The subjects received just a basic tutorial, including essential information, to start interacting with the robot. No other assistance was provided to them during the interaction.

We use a seven stage Likert scale to allow the person to express how much they agree or disagree with a particular statement [33]. The UX questionnaire often adopts the Likert scale to reduce the well-known central tendency bias for such items. The following is an example of a topic of the questionnaire:

Negative ① ② ③ ④ ⑤ ⑥ ⑦ Positive

Table 2 reports the questionnaire topics and the mean and standard deviation values obtained for each of them and it is made up of three parts. The first part consists of an overall evaluator to understand if the user has had a positive or negative experience. Four items constitute the section concerning the pragmatic aspects (Ease-of-use, Effectiveness, Learnability, Reliability). The other five items compose the last part regarding the hedonic aspects (Attractiveness, Trust, Fun, Acceptance).

Figure 10 shows the graph of the obtained results in which the standard deviation was also reported for each item. The results show a very good evaluation of the application in general (All test results administered are available as supplementary material in CSV format). Even looking at the detailed result, they are all very positive in terms of pragmatic and hedonic aspects.

We found the minimum score for acceptance. The lowest score in the acceptance category is consistent with other robotics applications [51,56,65]. Acceptance and adoption (A&A) are one of the most critical aspects of the development of robotic applications. Furthermore, the A&A is a process that often requires a much longer time than the few minutes in which the interaction with our application took place. These technologies deeply affect users' life and are often viewed with distrust. People are still quite reluctant to interact with a robot rather than a person. This aspect is also confirmed by the result obtained from the Trust item.

Particularly interesting and encouraging is the result obtained for the "Ease-of-use" item. Users have found it easy to interact with the robot and this is an excellent result for us. The many bidirectional multi-modal signs used and the interaction model aimed to make the interaction as natural as possible. The score obtained from the item "Ease-of-use" seems to confirm the goodness of the approach.

The result obtained from the topic "Learnability" confirms that the objectives of the approach have been achieved. The excellent score obtained by this item confirms that the anthropomorphization process presented in this article has made learning the interaction model very simple.

6 Conclusion and Future Works

We have looked at many of the aspects of the interaction between human beings and have found robotic counterparts. At times, we have used metaphors to replicate some elements of communication. Furthermore, we explained how each of these elements could individually contribute to enriching and improving the interaction's anthropomorphisation. Each of these elements has been involved in an interaction model

Table 2 User Experience Questionnaire (UEQ)

	Evaluator	Question	Mean; STD
Overall	Positiveness	How do you evaluate the experience you just had with the robot? Negative ① ② ③ ④ ⑤ ⑥ ⑦ Positive	5.96; 1.02
Pragmatic	Ease-of-use	Was it easy to interact with the robot? Complicate ① ② ③ ④ ⑤ ⑥ ⑦ Simple	5.21; 1.37
	Effectiveness	Was the robot effective considering your needs? Ineffective ① ② ③ ④ ⑤ ⑥ ⑦ Effective	4.98; 1.29
	Learnability	Did you easily and intuitively understand how to interact with the robot? Hard ① ② ③ ④ ⑤ ⑥ ⑦ Easy	5.11; 1.21
	Reliability	How many errors or malfunctions did you observed during the interaction? Many ① ② ③ ④ ⑤ ⑥ ⑦ Few	4.71; 1.37
Hedonic	Attractiveness	How attractive did you find the experience? Ugly ① ② ③ ④ ⑤ ⑥ ⑦ Attractive	5.51; 1.22
	Trust	Do you think you can trust the interaction with the robot? Distrust ① ① ② ③ ④ ⑤ ⑥ ⑦ Trust	4.35; 1.93
	Fun	Was it fun interacting with the robot? Boring ① ② ③ ④ ⑤ ⑥ ⑦ Fun	5.84; 1.35
	Acceptance	Do you think humans can be replaced with a robot in applications like this? Irreplaceable ① ② ③ ④ ⑤ ⑥ ⑦ Replaceable	4.33; 1.83
	Usefulness	Do you think this application was useful? Unuseful ① ② ③ ④ ⑤ ⑥ ⑦ Useful	4.90; 1.62

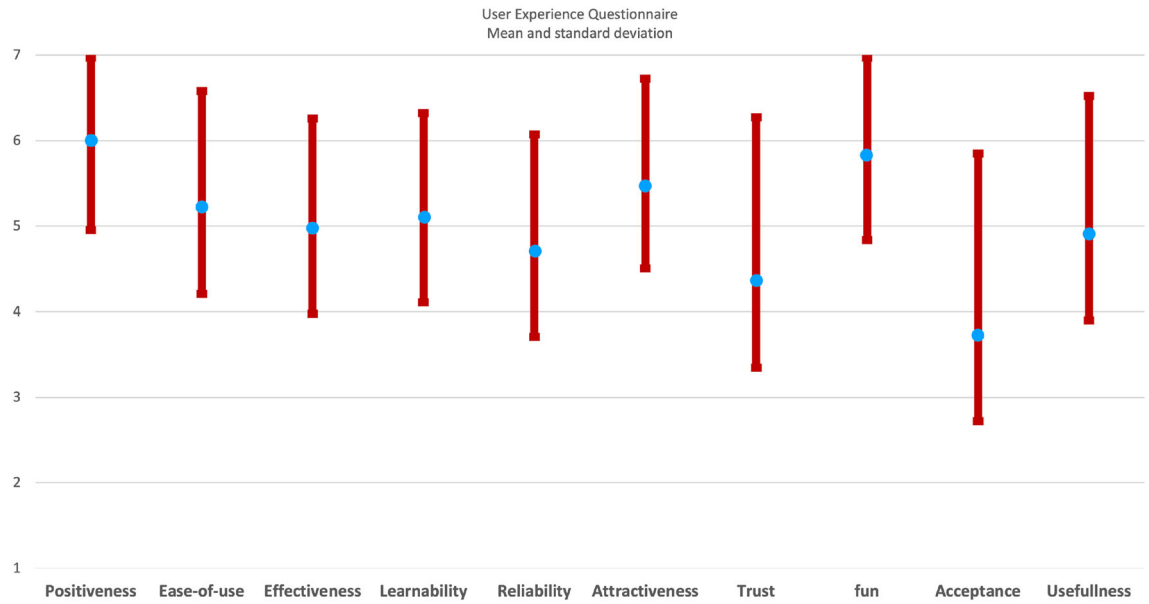

Fig. 10 User Experience Questionnaire results: Mean and Standard deviation are reported for each item

based on a finite state automaton that evolves based on events arising from the interaction between human and robot.

The presented model is theoretical and has been implemented in a ROS environment to ensure flexibility and portability. Furthermore, the model has been widely used in the W@ICAR application, proving its effectiveness with non-expert users interacting with robots. They interacted naturally with the robot and immediately understood the interaction paradigm.

As reported in Sect. 5, the results from a functional point of view are very encouraging. The user experience results also showed that the application was highly rated and the model largely met the user's expectations.

Now, we are starting to use the same model, obviously with some specific changes, for a medical assistance project (AMICO - Assistenza Medicale In COntextual Awareness) at the patient's home. In this project, the robot interacts with

the patient to check that the therapeutic process is followed scrupulously.

We are firmly convinced that a fundamental element in the anthropomorphization process of human-robot interaction is the manifestation, management and exchange of sensations and emotions. We have already dealt with the aspects of robots' emotions and sensations that we call "roboceptions". In particular, we have designed and implemented an artificial somatosensory system for a humanoid robot [4,47] able to make the robot perceive some "roboceptions" [23,42]. Thanks to the soft sensor paradigm [17,44–46], the robot processes its sensory data and transforms them into information with greater semantic content [43]. Therefore, considering that the robots influence the behaviour of the robot [3,5,6], they must necessarily also influence the interaction of the robot with other subjects. Social distance, for example, can be linked to the concept of anxiety. We perceive too much closeness with a stranger as a disturbing element (getting into the elevator with a stranger). Instead, as an element of pleasure if it is a friend or a partner. In future works, we will integrate these aspects into the human-robot interaction model [7].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12369-021-00855-w.

Acknowledgements Special thanks to Massimo Esposito and colleagues from the CNR-ICAR Language and Knowledge Engineering Group, who developed the conversational agent on the dialogue with the robot.

Funding This research was partially supported by the project AMICO - Assistenza Medicale In COntextual Awareness, with funding from the National Programs of the Italian Ministry of Education, Universities and Research (code: ARS01_00900). The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript or in the decision to publish the results.

Data availability The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

Declarations

Conflict of interest the authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethical approval the content of this article and the methods used are compliant with the ethical guidelines of the CNR.https://www.cnr.it/sites/default/files/public/media/doc_istituzionale/ethics/guidelines-for-research-integrity-2019.pdf.

References

1. Admoni H, Scassellati B (2017) Social eye gaze in human-robot interaction: a review. J Hum Robot Interact 6(1):25–63
2. Argyle M, Cook M, Cramer D (1994) Gaze and mutual gaze. Br J Psychiatry 165(6):848–850. https://doi.org/10.1017/S0007125000073980
3. Augello A, Infantino I, Maniscalco U, Pilato G, Rizzo R, Vella F (2018) Robotic intelligence and computational creativity. Encycl Semant Comput Robot Intell 2(1):1850011
4. Augello A, Infantino I, Maniscalco U, Pilato G, Vella F (2018) Robot inner perception capability through a soft somatosensory system. International Journal of Semantic Computing 12(01):59–87
5. Augello A, Città G, Gentile M, Infantino I, La Guardia D, Manfré A, Maniscalco U, Ottaviano S, Pilato G, Vella F, et al (2017) Improving spatial reasoning by interacting with a humanoid robot. In: International conference on intelligent interactive multimedia systems and services. Springer, Cham, Berlin, Heidelberg. pp 151–160
6. Augello A, Infantino I, Maniscalco U, Pilato G, Vella F (2017) The effects of soft somatosensory system on the execution of robotic tasks. Robotic computing (IRC). In: IEEE international conference on. IEEE, New York, pp 14–21
7. Augello A, Infantino I, Maniscalco U, Pilato G, Vella F (2018) A cognitive architecture for social robots. In: 2018 IEEE 4th International forum on research and technology for society and industry (RTSI). IEEE, New York, pp 1–5. https://doi.org/10.1109/RTSI.2018.8548520
8. Baresi L, Pezzè M (2006) An introduction to software testing. Electron Notes Theor Comput Sci 148(1):89–111. http://www.sciencedirect.com/science/article/pii/S1571066106000442. Proceedings of the School of SegraVis research training network on foundations of visual modelling techniques (FoVMT 2004)
9. Bartl C, Dorner D (1998) Psi: a theory of the integration of cognition, emotion andmotivation. In: Proceedings of the 2nd European conference on cognitive modelling. DTIC Document, Nottingham, England, pp 66–73
10. Bethel CL, Murphy RR (2008) Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. IEEE Trans Syst Man Cybern Part C Appl Rev 38(1):83–92
11. Bowring JF, Rehg JM, Harrold MJ (2004) Active learning for automatic classification of software behavior. In: Proceedings of the 2004 ACM SIGSOFT international symposium on software testing and analysis, ISSTA '04, pp 195–205. Association for Computing Machinery, New York. https://doi.org/10.1145/1007512.1007539
12. Bruce A, Nourbakhsh I, Simmons R (2002) The role of expressiveness and attention in human-robot interaction. In: Proceedings 2002 IEEE international conference on robotics and automation (Cat. No.02CH37292), vol 4. IEEE, New York, USA, pp 4138–4142
13. Caggianese G, Pietro GD, Esposito M, Gallo L, Minutolo A, Neroni P (2020) Discovering leonardo with artificial intelligence and holo-

grams: a user study. Pattern Recognit Lett 131:361–367. https://doi.org/10.1016/j.patrec.2020.01.006
14. Carroll JM, Mack RL, Kellogg WA (1988) Chapter 3 - interface metaphors and user interface design. In: Helander M (ed) Handbook of human-computer interaction. North-Holland, Amsterdam, pp 67–85. https://doi.org/10.1016/B978-0-444-70536-5.50008-7
15. Cassell J (2001) Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. MIT Press, Cambridge, MA, USA, pp 1–27
16. Chellappa R, Wilson CL, Sirohey S (1995) Human and machine recognition of faces: a survey. Proc IEEE 83(5):705–741
17. Ciarlini P, Maniscalco U (2006) Mixture of soft sensors for monitoring air ambient parameters. In: Proceedings of the XVIII IMEKO world congress. IMEKO, Hungary
18. Duncan S (1996) Sgrammatical form and thinking-for-speaking in mandarin chinese and english: an analysis based on speech-accompanying gesture. In: Unpublished Ph.D. dissertation, University of Chicago. Chicago
19. Duncan S, Fiske D (2015) Face-to-face interaction: research, methods, and theory. Routledge Library Editions: Communication Studies, Taylor & Francis, Milton Park, Abingdon, Oxfordshire
20. Ehrlich S, Wykowska A, Ramirez-Amaro K, Cheng G (2014) When to engage in interaction - and how? eeg-based enhancement of robot's ability to sense social signals in hri. In: 2014 IEEE-RAS international conference on humanoid robots. IEEE, Madrid, Spain, pp 1104–1109
21. Erickson TD (1995) Working with interface metaphors. In: Baecker RM, Grudin J, Buxton WA, Greenberg S (eds) Readings in human-computer interaction, interactive technologies. Morgan Kaufmann, San Francisco, CA, pp 147–151. https://doi.org/10.1016/B978-0-08-051574-8.50018-2
22. Francis P, Leon D, Minch M, Podgurski A (2004) Tree-based methods for classifying software failures. In: 15th international symposium on software reliability engineering. IEEE, New York, pp 451–462. https://doi.org/10.1109/ISSRE.2004.43
23. Galipó A, Infantino I, Maniscalco U, Gaglio S (2017) Artificial pleasure and pain antagonism mechanism in a social robot. International conference on intelligent interactive multimedia systems and services. Springer, Cham, Berlin, Heidelberg, pp 181–189
24. Gao J, Tao C, Jie D, Lu S (2019) Invited paper: what is ai software testing? and why. In: 2019 IEEE international conference on service-oriented system engineering (SOSE). IEEE, New York, pp. 27–2709. https://doi.org/10.1109/SOSE.2019.00015
25. Haarmann H (2011) Language in its cultural embedding. De Gruyter Mouton, Berlin, Boston. https://www.degruyter.com/view/title/11678
26. Hall ET, Birdwhistell RL, Bock B, Bohannan P, Diebold AR, Durbin M, Edmonson MS, Fischer JL, Hymes D, Kimball ST, La Barre W, McClellan JE, Marshall DS, Milner GB, Sarles HB, Trager GL, Vayda AP (1968) Proxemics [and comments and replies]. Curr Anthropol 9(2/3):83–108. https://doi.org/10.1086/200975
27. Hamacher A, Bianchi-Berthouze N, Pipe AG, Eder K (2016) Believing in bert: using expressive communication to enhance trust and counteract operational error in physical human-robot interaction. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, New York, pp 493–500
28. Hartson R, Pyla P (2012) The UX book: process and guidelines for ensuring a quality user experience. Elsevier, USA. https://books.google.it/books?id=w4I3Y64SWLoC
29. Hassenzahl M (2013) User experience and experience design, chap. 3, pp 80, 94. The Interaction Design Foundation, Aarhus, Denmark . http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.html
30. Hauser MD (1996) The evolution of communication. MIT Press, Cambridge, MA
31. Hjelmås E, Low BK (2001) Face detection: a survey. Comput Vis Image Underst 83(3):236–274
32. Hooda I, Chhillar RS (2015) Software test process, testing types and techniques. Int J Comput Appl 111(13):10–14
33. Joshi A, Kale S, Chandel S, Pal D (2015) Likert scale: explored and explained. Br J Appl Sci Technol 7(4):396–403. https://doi.org/10.9734/bjast/2015/14975
34. Kane J, Yanushevskaya I, Looze CD, Vaughan B, Chasaide AN (2014) Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions. In: Li H, Meng HM, Ma B, Chng E, Xie L(eds) INTERSPEECH 2014, 15th annual conference of the international speech communication association, Singapore, September 14-18, 2014. ISCA
35. Kendon A (1970) Movement coordination in social interaction: some examples described. Acta Physiol (Oxf) 32:101–125
36. Kendon A, Harris R, Key M (2011) Organization of behavior in face-to-face interaction. World anthropology. De Gruyter, Berlin, Germany . https://books.google.it/books?id=Lq4UvjsROToC
37. Kita S (1990) The temporal relationship between gesture and speech: a study of japanese-english bilinguals. Masters Thesis, Department of Psychology, University of Chicago. https://ci.nii.ac.jp/naid/10009705492/en/
38. Lalonde M, Byrns D, Gagnon L, Teasdale N, Laurendeau D (2007) Real-time eye blink detection with gpu-based sift tracking. In: Fourth Canadian conference on computer and robot vision (CRV '07), IEEE, New York City, pp 481–487
39. Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: Holzinger A (ed) HCI and usability for education and work. Springer, Berlin Heidelberg, pp 63–76
40. Li HZ (2006) Backchannel responses as misleading feedback in intercultural discourse. J Intercult Commun Res 35(2):99–116. https://doi.org/10.1080/17475750600909253
41. Lindblom J, Andreasson R (2016) Current challenges for ux evaluation of human-robot interaction. In: Schlick C, Trzcieliński S (eds) Advances in ergonomics of manufacturing: managing the enterprise of the future. Springer, Cham, pp 267–277
42. Maniscalco U, Infantino I (2017) An artificial pain model for a humanoid robot. International conference on intelligent interactive multimedia systems and services. Springer, Cham, Berlin Heidelberg, pp 161–170
43. Maniscalco U, Infantino I (2018) Soft sensors to measure somatic sensations and emotions of a humanoid robot. Ser Adv Math Appl Sci. https://doi.org/10.1142/11100
44. Maniscalco U, Pilato G (2012) Multi soft-sensors data fusion in spatial forecasting of environmental parameters. Adv Math Comput Tools Metrol Test X 84:252–259
45. Maniscalco U, Rizzo R (2015) Adding a virtual layer in a sensor network to improve measurement reliability. Adv Math Comput Tools Metrol Test X 86:260–264
46. Maniscalco U, Rizzo R (2016) A virtual layer of measure based on soft sensors. J Ambient Intell Humaniz Comput 8:1–10
47. Maniscalco U, Messina A, Storniolo P (2020) Ass4hr - an artificial somatosensory system for a humanoid robot. the ros package. SoftwareX 11. https://doi.org/10.1016/j.softx.2020.100501
48. Mey J (2001) Pragmatics?: an introduction, 2nd edn. Blackwell, Malden, MA
49. Minutolo A, Esposito M, Pietro GD (2017) A conversational chatbot based on kowledge-graphs for factoid medical questions. In: Fujita H, Selamat A, Omatu S (eds) New trends in intelligent software methodologies, tools and techniques - proceedings of the 16th international conference, SoMeT_17, Kitakyushu City, Japan, September 26-28, 2017, Frontiers in Artificial Intelligence

and Applications, vol 297, pp 139–152. IOS Press, Amsterdam, The Netherlands. https://doi.org/10.3233/978-1-61499-800-6-139
50. Mizoguchi H, Sato T, Takagi K, Nakao M, Hatamura Y (1997) Realization of expressive mobile robot. In: Proceedings of international conference on robotics and automation, vol 1, pp 581–586. IEEE, Albuquerque, NM
51. Moradi M, Moradi M, Bayat F (2018) On robot acceptance and adoption a case study. In: 2018 8th conference of AI robotics and 10th robocup iranopen international symposium (IRANOPEN). pp 21–25 . https://doi.org/10.1109/RIOS.2018.8406626
52. Patterson M (1983) Nonverbal behavior: a functional perspective. Social psychology series. Springer, New York
53. Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: 2007 2nd ACM/IEEE international conference on human-robot interaction (HRI). IEEE, New York City, pp 145–152.https://doi.org/10.1145/1228716.1228736
54. Rauh G (1983) Aspects of deixis* gisa rauh. Essays on Deixis 188:9
55. Rauterberg GM, Dätwyler M, Sperisen M (1995) From competition to collaboration through a shared social space. In: Proceedings of the east-west international conference on human-computer interaction (EWHCI95). Springer-Verlag, New York, pp 49–57
56. Salem M, Ziadee M, Sakr M (2014) Marhaba, how may i help you? effects of politeness and culture on robot acceptance and anthropomorphization. In: 2014 9th ACM/IEEE international conference on human-robot interaction (HRI). pp 74–81
57. Shea N (2005) Varieties of meaning: The 2002 jean nicod lectures. The jean nicod lectures. Based on lectures held in paris, 31 May 2002, and 4, 6, and 11 June 2002. by ruth garrett millikan. Q Rev Biol 80(3), 344–344
58. Sidner C, Lee C, Kidds C, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 66(1–2):140–164
59. Trifa VM, Koene A, Moren J, Cheng G (2007) Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In: RO-MAN 2007 - The 16th IEEE international symposium on robot and human interactive communication. IEEE, New York, pp 393–398
60. Valin J, Michaud F, Rouat J, Letourneau D (2003) Robust sound source localization using a microphone array on a mobile robot. In: Proceedings 2003 IEEE/RSJ international conference on intelligent robots and systems (IROS 2003) (Cat. No.03CH37453), vol 2. IEEE, New York, pp 1228–1233
61. van Greunen D (2019) User experience for social human-robot interactions. In: 2019 amity international conference on artificial intelligence (AICAI). IEEE, New York, USA, pp. 32–36. https://doi.org/10.1109/AICAI.2019.8701332
62. Walters ML, Dautenhahn K, te Boekhorst RKLK, Kaouri C, Woods S, Nehaniv C, Lee D, Werry I (2005) The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. ROMAN 2005. ieee international workshop on robot and human interactive communication, 2005. IEEE, New York City, pp 347–352
63. Watzlawick P, Beavin J, Jackson D (1967) Pragmatics of human communication. W.W. Norton, New York
64. White S (1989) Backchannels across cultures: a study of americans and japanese. Lang Soc 18(1):59–76. https://doi.org/10.1017/S0047404500013270
65. Wu YH, Wrobel J, Cornuet M, Rigaud AS (2014) Acceptance of an assistive robot in older adults: a mixed-method study of human-robot interaction over a 1-month period in the living lab setting. Clin Interv Aging 9:801
66. Xie X, Ho JWK, Murphy C, Kaiser G, Xu B, Chen TY (2011) Testing and validating machine learning classifiers by metamorphic testing. J Syst Softw 84(4):544–558
67. Yanushevskaya I, Kane J, de looze C, Chasaide A The distribution of pitch patterns and communicative types in speech chunks preceding pauses and gaps. In: Proceedings of the international conference on speech prosody. https://doi.org/10.21437/SpeechProsody.2014-180
68. Yngve VH (1970) On getting a word in edgewise. CLS-70. University of Chicago, Chicago, pp 567–577