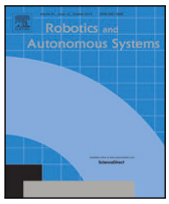
Robotics and Autonomous Systems 198 (2026) 105363

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot



Towards robot affective appraisal linking inner speech and emotion

Arianna Pipitone [a,c,*], Sophia Corvaia [b], Antonio Chella [b,c]

[a] Department of Humanities, University of Palermo, Viale delle Scienze, building 12, 90100, Palermo, Italy
[b] Department of Engineering, University of Palermo, Viale delle Scienze, building 6, 90100, Palermo, Italy
[c] ICAR - CNR, National Research Council, Palermo site, via Ugo la Malfa, 90100, Palermo, Italy

ARTICLE INFO

Keywords:
Robot's inner Speech
Computational models of emotions
Cognitive architecture
Robot appraisal evaluation

ABSTRACT

Recent studies in Robotics and AI suggest that robots "thinking out loud" can foster positive human feedback and support collaborative goal achievement. By externalizing their internal reasoning, robots enhance transparency and explainability, which are crucial for trust and robustness in human–robot interaction.

This work investigates the role of robot inner speech in supporting affective appraisal, focusing on the emergence, coherence, and interpretability of emotionally grounded evaluations. While the relationship between inner speech and emotion has been extensively explored within human psychology and cognitive science, this work is the first to operationalize this connection within a robotic architecture for affective appraisal.

Grounded in appraisal theories, the proposed model employs inner speech to simulate internal reflection, enabling the identification and evaluation of contextual variables relevant to affective assessment. Through this internal dialogue, the robot structures its appraisal process and externalizes it, allowing human partners to access and interpret the underlying affective reasoning.

The model is evaluated by comparing its appraisal dynamics with normative emotional patterns observed in adults under stress, and by assessing the interpretability of the robot's affective behavior through human observation. Results indicate that the model produces coherent and context-sensitive evaluations, improving upon a widely adopted computational model of emotion in terms of plausibility and transparency.

1. Introduction

Philosophers, psychologists, and neuroscientists have long studied the roles of the inner voice in cognitive and psychological human functions, including affectivity and passions [1–3]. The inner voice is the linguistic surface form of thoughts. A person is engaged in a self-dialogue when he/she thinks with a running verbal commentary [4]. This experience helps people to focus on the context, to plan actions, to make decisions, and to become conscious of facts and events.

A link between inner speech and emotions was first outlined by Vygotsky [5], which argued about the continued and dynamic interactions between the intellect, meant as thoughts, and the affective sphere. Such a link evolves and fluctuates in the course of life, and is bi-directional: *from the affective sphere of consciousness to thought and from thought to the affective sphere of consciousness* [6]. The same perspective is supported by Lazarus [7,8] and, in general, by cognitive appraisal theorists: thoughts are fundamental in affectivity because thinking must occur first in the sequence of human cognitive processes that lead to an emotional experience. More specifically, the sequence starts when a stimulus motivates the person, followed by an emergent thought in linguistic form associated to that stimulus, and ends with the experience of a physiological response or emotion. Thus, the role of thinking in feeling emotions is fundamental.

When modeling emotional behaviors in artifacts, such as robots and artificial agents, appraisal theories are frequently referred to, basing the emotion evaluation and its intensity on appraisal variables. Appraisal variables are links between the context and emotional components, and they are well suited for computation.

Despite the fact that existing contributions based on appraisal theories have produced significant results in the affective computing field, the role of inner dialogue has not been investigated as a fundamental process for computing appraisal variables and, consequently, for shaping artificial affective evaluations.

The main motivation of this work is that, although appraisal-based models have produced significant results within affective computing (e.g., [9]), existing robotic implementations still lack mechanisms that support transparent and self-explanatory affective evaluation. At the same time, inner speech has been recognized as a powerful cognitive

* Corresponding author at: Department of Humanities, University of Palermo, Viale delle Scienze, building 12, 90100, Palermo, Italy.
E-mail addresses: arianna.pipitone@unipa.it (A. Pipitone), sophia.corvaia@unipa.it (S. Corvaia), antonio.chella@unipa.it (A. Chella).

https://doi.org/10.1016/j.robot.2026.105363
Received 1 August 2024; Received in revised form 4 January 2026; Accepted 20 January 2026
Available online 21 January 2026

tool for directing attention, guiding reasoning, and facilitating self-regulation in humans [10,11]. Recent work in robotics indicates that inner speech can enhance transparency and explainability of robotic behavior [12]. However, its potential as a computational mechanism driving appraisal processes has not yet been explored. Addressing this gap is essential for designing affective architectures capable of producing interpretable evaluations of emotionally relevant situations.

The present work investigates how this ability contributes to the robot's affective evaluation mechanisms, by focusing on the cognitive processes that make emotional appraisals interpretable to human observers. This research builds on prior work on robot inner speech, where some authors proposed a cognitive architecture for inner speech [13,14] and deployed it on a real robot capable of verbalizing its internal reasoning. Previous studies have shown that such capability improves transparency (i.e., traceability of decision processes), robustness in collaborative tasks, and human trust [12,15]. The present work investigates how this ability contributes to the robot's affective evaluation mechanisms.

The proposed work introduces the following main innovations: (i) a cognitive model integrating robot inner speech with appraisal-based affective evaluation, in which inner speech supports attention allocation, knowledge access, and the evaluation of salient events; (ii) a formalization of the appraisal mechanism enhanced by inner speech, including a structured mathematical representation of appraisal variables and their combination for generating affective evaluations.

These contributions extend current research on computational appraisal and strengthen the role of inner speech as a mechanism for cognitive transparency in robotic systems. The model has been implemented on a humanoid robot, where all internal processes are accompanied by verbalized inner speech, allowing observers to follow the robot's internal reasoning as it evaluates the current context. The affective evaluation capabilities of the implemented system have been assessed through two complementary experimental validations, demonstrating that the resulting affective behavior is coherent and interpretable to human observers.

The proposed model of inner speech and emotions finds inspiration from the modal model by Gross [16], which provides a structure highlighting the steps involved in the emotional self-regulation cognitive process, which are Situation, Attention, Appraisal and Response.

The general structure of the modal model is maintained in the proposed one: the perceptual and output channels from and to the Situation module model the typical cognitive perception and control functions of the robot from and to the environment. The strength of this approach is related to the process in the Attention block that enables the cognitive evaluations of the context, focusing on the relevant aspects useful for the different appraisal variable via inner speech.

Once a stimulus becomes a thought (i.e., a textual surface form is associated to the stimulus), the inner dialogue starts and enables the evaluation of further facts and events (which could be external, i.e. regarding the environment, or internal, i.e. regarding the inner state), simulating the cognitive evaluation of the context. The inner dialogue is a set of turns, and a new turn emerges in response to the previous one, leading to a rehearsal loop. The loop is repeated until no further facts to evaluate emerge, or all the needed information are inferred. At this point, the model computes the appraisal variables (the Appraisal block), that are specifically formalized so that the trends of such variables replicate the observed trends in healthy adults [17]. Then, the corresponding emotion associated with the appraisal variables, with a specific intensity, is evaluated by referring to the Russel's Circumplex Model of Emotion [18] (more popularly known as the Circumplex Model of Affect), and the emotion is elicited (the Response block).

Two different experimental setups were carried out, which are qualitative and empirical. The first experiments aims to evaluate the coherence of the proposed model with appraisal theory and the human's emotional behaviors. In this case, the robot was immersed in simulated stressful situations, and its emotional behavior was compared to the trends presented in [19]., where a strategy for evaluating the computational models of emotions is described. In particular, the strategy is defined for the evaluation of the EMA model (EMotion and Adaptation model [9]), and consists of simulating stressful situations, that are loss and aversive, and observing the values of appraisal variables provided by the model. These values are then compared with the emotional trends assessed by the standard SCPQ (Stress and Coping Process Questionnaire [17]) clinical tool in the same scenarios.

The evaluative strategy is based on a narration for abstracting the stressful episodes proposed by the SCPQ tool, and identifies four canonical situations. In this work, canonical situations are reproduced and the outputs, in terms of appraisal variables and emotional reactions, are observed and compared to the EMA's and SCPQ's trends. In the second kind of experiment, the model was empirically assessed by human observers. In this case, a set of scenes representing interactive acts were shown to the participants, and they were engaged in evaluating the emergent robot's emotional behavior. They evaluated the robot's behaviors by open-ended and multiple-choice questionnaires according to the protocol described in [20].

Results so far are promising. The mathematical formalization of the appraisal variables computes the expected appraisal patterns and emotions. Moreover, the results show improvements in many cases in respect to EMA.

This paper is organized as follow: the theoretical basis of the model, including the appraisal foundations, the Gross' model, and the role of inner speech in feeling emotions, are presented in Section 2. Section 3 describes the whole proposed model linking inner speech and emotions. In particular, this section details the inner speech process for cognitive evaluation of the context, and the proposed mathematical formalization of the appraisal variables. Moreover, given the values of the appraisal variables, the section shows how the emotion related to the computed appraisal pattern emerges with a specific intensity. A couple of use cases related to a collaborative task, involving the robot and a human partner, are detailed in Section 4 with the aim to show how the model works. The methodologies for the double evaluations of the model and the obtained results are discussed in Section 5. Related work and the current state of the art are reviewed in Section 6. Finally, Section 7 provides a general discussion on the model's strengths, limitations, ethical implications, and outlines directions for future research, while Section 8 draws conclusions.

Throughout the paper, references to the robot's *emotional experience* or *feeling emotions* are intended in a functional and representational sense. Specifically, emotional states denote structured internal representations derived from appraisal processes, which modulate the robot's behavior and expressive output. These states do not imply subjective feelings, consciousness, or phenomenological experience, but rather constitute computational constructs used to model and regulate affective behavior.

2. Theoretical background

2.1. Appraisal theories

Appraisal theories [21–23] posit that emotions arise from an individual's interpretation of situational events. Such interpretations are grounded in cognitive evaluation processes that assess the relevance, implications, and potential consequences of perceived events. These evaluations give rise to a meaning structure that organizes the elements contributing to the resulting emotional state.

A central assumption of appraisal theories is that emotions are elicited through patterns of appraisal variables, which quantify how aspects of the situation relate to the agent's goals and concerns. Although individual theories differ in the specific set and formulation of these variables, they share a bottom-up perspective: emotions emerge

Fig. 1. The modal model of emotion regulation proposed by Gross.

from context-derived appraisals rather than from predefined emotion categories.

Appraisal theory has become a standard framework in computational affective modeling due to its interpretability and the operational nature of appraisal variables. Classical categorical approaches [24,25] conceptualize emotions as discrete classes, but such taxonomies often struggle to account for contextual variability and the graded nature of emotional meaning. More recent models [26] emphasize continuous emotional dimensions and context-specific appraisal patterns, making them well suited for artificial agents operating in dynamic environments.

In this work, appraisal theory provides the conceptual foundation for modeling the robot's evaluation of its environment through inner speech. The appraisal process integrates internal operational conditions (e.g., battery level, joint temperature) and environmental characteristics (e.g., disorder, noise), enabling a context-sensitive affective assessment grounded in the robot's functional constraints.

2.2. The modal model of emotions

Gross's modal model [16] characterizes emotion as a process unfolding through four stages (Fig. 1):

1. Situation: an internal or external event with potential emotional significance;
2. Attention: selective focus on salient aspects of the situation;
3. Appraisal: evaluation of the situation, giving rise to an emotional state;
4. Response: behavioral, experiential, and physiological reactions.

The model encompasses automatic and controlled mechanisms and highlights the recursive influence between emotional responses and subsequent situational appraisals. This process-oriented view aligns naturally with the proposed architecture: the robot identifies a relevant situation, focuses on its key elements, interprets it through an appraisal driven by inner speech, and generates an affective state representation that modulates its behavior.

2.3. The function of self-talk in emotion

Vygotsky's interfunctional theory of emotion [5] underscores the reciprocal relationship between thought, language, and affect. Language is not merely a medium of expression; it shapes thought, thereby influencing how emotional experiences are structured and understood. Inner speech plays a central role in this mediation, providing the cognitive substrate through which individuals interpret and regulate emotional states. Empirical findings by Morin [27] further support this view, showing that inner speech frequently includes self-evaluative and affective content.

The proposed architecture adopts this perspective by employing inner speech as the mechanism through which the robot constructs an internal interpretation of the situation. Through cycles of linguistic reasoning, the robot evaluates contextual and internal information, yielding an appraisal pattern from which an emotional state representation is computed. This representation in turn guides expressive behavior and enables the agent to maintain an integrated cognitive-affective understanding of the situation.

3. Modeling emotions in robots with inner speech

Fig. 2 shows the proposed cognitive architecture of inner speech and emotions. As mentioned above, the backbone structure is the modal model, integrated by a rehearsal loop for inner speech. Moreover, a further block, modeling the robot's memory, contains the robot's knowledge and is used for the retrieval/analysis of further concepts correlated to the perceived one.

In simple terms, the architecture enables the robot to experience the emotional response when a stimulus occurs in the environment. The Situation block encodes the stimulus by associating a symbolic form to the perceived signal. The Attention block recalls the correlated concepts from long term memory in linguistic form: it represents the emergence of the thought corresponding to the stimulus, which triggers inner speech. The thought is re-heard by the Attention block, and it could require to retrieve further stimulus from the environment by the Situation block, which will search for them, or further concepts from memory. The retrieve/recall from environment/memory depends on the concepts in the turn, and on their possible matching with new concepts in the environment/memory. Inner speech enables the cognitive evaluation and provides the parameters to the Appraisal block which computes the appraisal pattern and the corresponding emotion to that stimulus. The emotion is then verbally externalized by the Response block. All the processes related to the cognitive evaluation with inner speech are detailed in Section 3.2.

The model is based on *general appraisal variables*, but *specific appraisal variables* could be added in each specific application scenario. The first variables are from appraisal theories, and they are generally recognizable in each context. The general appraisal variables included in the proposed model are:

- *Likelihood*, which measures how probable is the outcome of an event. Generally, it represents the probability of negative outcomes which lead to stressful situations. The higher the likelihood is, the more negative the outcome is, and the more stressful the event is;
- *Controllability*, which measures how an outcome of an event could be modified by directly acting on the context;
- *Changeability*, which measures how an outcome of an event could be altered by some other event or by somebody else;
- *Desirability*, which measures how desirable the outcome of an event is.

For each of these variables, the model proposes a mathematical formalization, as defined in Section 3.3.

The specific appraisal variables depend on the specific scenario in which the robot is immersed, and pertain to specific aspects that could affect its emotional state representation. Considering that it is impossible to anticipate all possible situation-dependent appraisal components, the model is designed to allow the introduction of additional variables whenever a refinement of the robot's affective behavior is desired. These variables are not predefined by the architecture: they may encode any contextual factor that the designer considers relevant, and they may take any mathematical form, from continuous functions to discrete indicators. The only requirement is to specify whether their contribution to the appraisal is positive or negative. How these specific appraisal variables could be integrated is discussed in the same aforementioned subsection.

Once the appraisal variables are computed, the corresponding emotion is inferred by applying the circomplex model of emotions [18], according to which the emotions result from the combinations of two dimensions, which are the *valence*, which explains how pleasant/unpleasant an emotion is, and the *arousal* which explains the level of physiological activation. The emotions are thus represented in a two-dimensional space. An emotion is a point in that space.

The circumplex model includes 28 emotions, but in the proposed architecture only the five basic emotions identified by Ekman [28] are

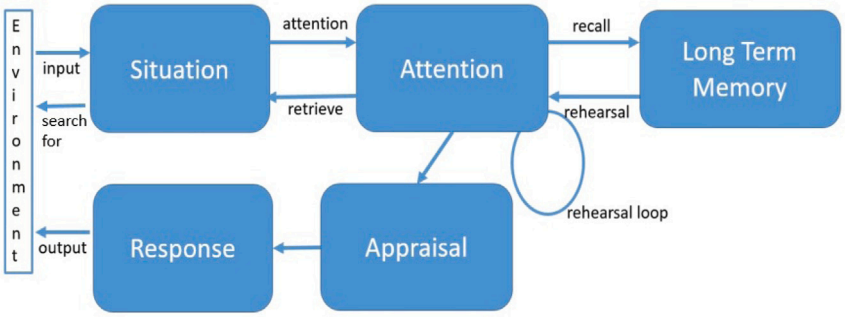


Fig. 2. The proposed cognitive architecture of inner speech and emotions.

considered: *happiness*, *sadness*, *fear*, *anger*, and *disgust*. The mapping between appraisal variables, valence-arousal values, and the corresponding emotion is detailed in Section 3.3. Moreover, the position of the emotion in the circumplex space allows assigning an intensity to the resulting affective state.

The choice to adopt Ekman's basic emotions is motivated by both methodological and practical considerations. Ekman's set constitutes a cross-culturally validated and widely accepted subset of affective categories in HRI, ensuring interpretability and comparability across studies. Moreover, in HRI contexts, the primary requirement is to provide the user with clear, unambiguous, and easily recognizable emotional cues. Restricting the emotion set to Ekman's five basic categories prevents over-fragmentation of the affective space and reduces the risk of ambiguous or overly nuanced emotional outputs, which may not be easily distinguishable by human partners during interaction.

A further advantage of using a compact emotional set is that it facilitates the work of external evaluators and human observers involved in assessing the robot's behavior. Simpler and more distinct emotional categories reduce interpretative ambiguity and support more reliable judgments about the robot's internal state during experimental validation. While the architecture is compatible with richer emotional taxonomies, using Ekman's basic emotions offers a principled and operationally efficient compromise between expressive power and interaction clarity.

3.1. The knowledge model

The knowledge of the robot models a situated cognitive agent operating in a structured environment, integrating perception, cognition, action, and contextual awareness. It provides a formal representation of cognitive and metacognitive processes, including perception, emotional and physical state assessment, planning, and behavioral control. Reactive and abstract actions are explicitly modeled and constrained by general feasibility rules that link actions to manipulable objects.

The physical domain is described through a detailed taxonomy of objects and utensils, enriched with material properties and etiquette-based spatial relations that encode conventional placement and usage constraints.

Additionally, the ontology captures interaction context through requests characterized by environmental, spatial, and sensory parameters. Overall, this ontology enables semantic reasoning over agent behavior, environmental conditions, and normative constraints, supporting adaptive decision-making in human–agent interaction scenarios.

Specifically, the knowledge model pertains to two worlds, which are the *external world* and the *inner world*. The first world consists of all the facts contained in the domain and it represents the generic knowledge about the environment the robot is in. The inner world represents the knowledge the robot has about itself, that is its physical conditions (for example the level of battery, the working conditions of the arms, joints, and so on). That knowledge is formalized by an ontology, the *KB ontology*, which defines the general concepts of the knowledge, with their attributes and the relations among them. The general concepts are the ontology's *classes* whose *instances* are the concrete entities in the environment (when the class models a concept of the external world) or the concrete state of the robot (when the class models a concept in the internal world).

Formally, the KB ontology is the tuple $O = \langle C_o, P_o, T_s, L, P_d \rangle$ defined according to the W3C technical report specification,[1] where C_o is the set classes or general concepts of the worlds, I_o is the set of individuals, that are the instances of the previous classes, P_o is the set of the object properties, linking two concepts, and P_d is the set of the datatype properties, linking a concept and a datatype value.

For example, the class `Person` represents a human, that is a concept in the external world, and an instance of such a class will model an individual perceived in the actual context; the `emo` and `age` attributes are examples of datatype properties of the class `Person`, representing the emotional state of the individual and his/her age respectively. Instead, a set of object properties could model the individual's position in respect to another perceived entity, which are the `left`, `right`, `up` and `down` object properties, linking the individual to the specific entity if he/she is at the left, right, up or down of that entity.

The ontology includes the robot's internal world knowledge. For example, the class `Emotion` with the sub-classes `Anger`, `Joy`, `Sadness` and so on, models the possible robot's emotions. The instances of these classes define the emotional state of the robot.

An example of complete ontology structure, used in this work, is available in OWL format in the project's GitHub repository (see Resource Availability section 8), enabling full inspection of its classes, properties, axioms, and hierarchical organization.

An excerpt of the ontology is represented in Fig. 3.

Importantly, the inner speech architecture does not depend on these specific domain classes or properties: it operates through generic ontological queries that can be applied to any domain-specific ontology structured according to OWL specifications.

The only architectural requirement is that the agent-related portion of the ontology — modeling the robot's internal states (emotions, battery level, motor conditions) and general cognitive processes (perception, action feasibility, appraisal variables) — must be integrated into any target domain ontology.

This agent-related component is relatively small and domain-independent, containing the classes and properties necessary for the robot to represent its own physical and affective state regardless of the application context. The domain independence of the approach means that deploying the system in a different application domain (e.g., healthcare, manufacturing, education) requires only merging this fixed agent ontology with a new domain-specific ontology describing the relevant objects, actions, and constraints of that domain. The inner speech mechanism remains entirely unchanged, making scalability concerns minimal: the agent ontology is reused across all domains,

[1] https://www.w3.org/TR/owl-ref/

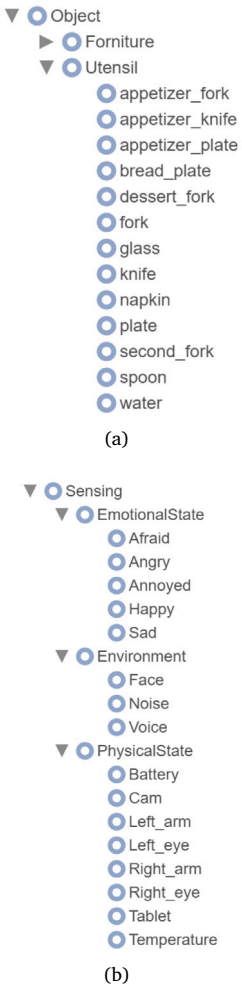


Fig. 3. Excerpts of the ontology modeling a generic domain (table setup objects) and the inner world of the robot.

and only the domain-specific knowledge needs to be modeled and integrated. This modular design ensures that the cognitive and affective architecture scales seamlessly to new application contexts without code modifications or architectural redesign.

3.2. Inner speech for cognitive evaluation of the context

The central idea of the proposed architecture is the *inner voice* of the robot with which it focuses its attention to the concepts of the situation which could affect its emotional state. The robot "internally reasons" about a stimulus, and focuses on the concepts that are related to that stimulus; the reasoning is symbolic, involving the linguistic surface form of the concepts. This is in line with the claim that *the words allow thinking about the emotions, and the emotions generate further words, or thoughts from the inner dialogue.* [6]

The inner dialogue enables the *cognitive evaluation* of the situation, by identifying the *meaning structure* of the context, and, according to appraisal theories, these processes provide the interpretation of the appraisal variables.

The mechanism is shown in Fig. 4, and it is inspired by [29]. Inner speech functions as an active cognitive mechanism that interacts with the robot's evaluative processes through a structured rehearsal loop composed of three sequential phases: *conceptualization*, *inner verbalization*, and *inner comprehension*. During conceptualization, the elements of the perceived situation are retrieved from the ontology, modeling the spreading activation of semantic processing [30]. The robot becomes "aware" of the perception and verbalizes the activated concepts, leading to the inner verbalization phase.

Through inner verbalization, the robot formulates explicit internal linguistic representations about the meaning and relevance of the current event, producing them as overt internal speech. These representations are then reinterpreted during inner comprehension, a phase whose functional role is to internally re-hear the linguistic output generated in inner verbalization. This internally perceived utterance is treated by the system as a newly perceived stimulus, thereby reactivating the conceptualization phase and initiating a new cycle of semantic processing. Across these phases, the produced linguistic forms update and refine the meaning structure with additional inferred concepts.

The rehearsal loop operates iteratively until no additional concepts can be derived from the ontology or until a predefined semantic depth is reached. In this implementation, the depth is limited to two levels of semantic expansion beyond the concepts directly activated in the ontology. This choice offers a balance between cognitive plausibility and computational tractability. From a cognitive perspective, theories of spreading activation [31] indicate that the most affectively relevant inferences arise within the first two conceptual neighborhoods of a stimulus. From an implementation standpoint, limiting the expansion depth prevents propagation into conceptually remote or irrelevant regions of the ontology, avoiding noise and unnecessary computational load [32].

Through this iterative cycling, the robot incrementally expands and sharpens its internal representation of the situation, enabling increasingly structured and context-sensitive appraisal cues to emerge. The verbalized and reinterpreted inferences resulting from this process directly inform appraisal variables such as desirability, controllability, and likelihood, creating a tight functional coupling between cognition and affect, and making the emotional outcome both coherent and transparent to human collaborators.

From an operational point of view, the meaning structure is formalized by the couple $T = [sem(.), syn(.)]$, which associates the semantics $sem()$ with the syntax $syn()$ of a word or set of words. The elements in the $sem()$ and $syn()$ components are referred to as *chunks*. The $sem()$ component contains *meanings*, i.e., chunks corresponding to concepts in the knowledge base, while the $syn()$ component contains *words*, i.e., chunks corresponding to linguistic tokens. The meaning structure expands during processing as new meanings and words are added; each new element is merged into the appropriate component unless it is already present.

The three phases of the rehearsal loop are formalized as follows:

1. *Conceptualization* A percept p is the symbolic form of a perceived stimulus which could be a voice stream, the image of an object, one or more features of the object, and so on. In any case, the form of the percept is textual and syntactically describes the perceived stimulus (i.e., it is the text produced by the speech recognition routines, the text describing the object or its features by the image processing routines, and so on). The step finds correspondences between the syntax of the percept and the knowledge base of the robot. Being $S = \{s_1, s_2, \ldots, s_n\}$ the set of tokens of the percept p, pre-processed by removing non-meaningful words (such as articles, conjunctions, prepositions, and so on), the meaning structure is filled by the chunks of the percept, so that $T = [sem(), syn(s_1, s_2, \ldots, s_n)]$.
   To conceptualize the percept means *to find the concepts in the KB ontology that better match with the syntax of the percept*: the matched concepts in the ontology are the classes that conceptualize the percept. The match is implemented by computing the Jaro–Winkler distance [33] between the elements in S and the labels of the ontology. This distance is a measure about how syntactically similar two words are. Being $d_{jw}(w_1, w_2)$ the

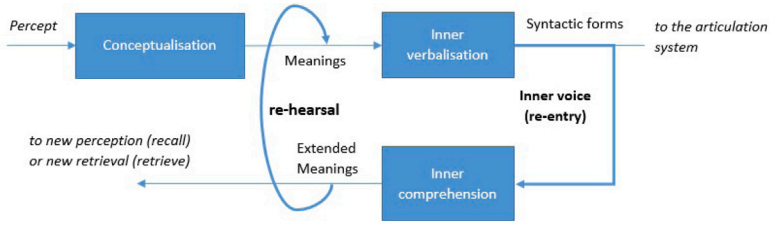


Fig. 4. The language re-entrance components: the syntactic forms from inner verbalization component are inputted to the inner comprehension one; further expansion of meanings allows to reason about beliefs and internal state thus identifying emotional relevant situation and defining attention.

function that returns the Jaro–Winkler distance between the words w_1 and w_2, and $couple(r)$ the function that returns the domain and the range of a property r, the match between S and the ontology O is modeled by the following functions:

- $a_c : S \rightarrow C_o$ that returns the set $a_c(s_i) = \{\ cl_k \mid d_{jw}(s_i,\ cl_k) > \alpha,\ cl_k \in C_o\}$ composed of the ontological classes such that the Jaro–Winkler distance between their labels and the words contained in S is higher than a threshold value α (the words are syntactically similar to the classes).
- $a_p : S \rightarrow P_o \cup P_d$ returning the set $a_p(s_i) = \{\ r_j \mid r_j \in P_o \cup P_d \wedge couple(r_j) \supset a_c(s_i)\}$ composed of the properties in the ontology for which the previous retrieved concepts are the domain or the range of these properties.

The threshold α used in the Jaro–Winkler matching was set to a relatively permissive value ($\alpha = 0.2$) based on preliminary exploratory trials. These trials were conducted to qualitatively verify that relevant ontological concepts were consistently retrieved during the conceptualization phase, while avoiding excessive spurious matches.
The goal at this stage is not to enforce strict lexical precision, but to favor semantic coverage by retrieving a set of candidate concepts that are subsequently filtered through ontological constraints (domain and range restrictions). Accordingly, the overall behavior of the architecture does not critically depend on the exact numerical value of the threshold. The result is a sub-ontology $O_m = \langle C_m, P_m \rangle$ where $C_m = \bigcup_S a_c(s_i)$, $P_m = \bigcup_S a_p(s_i)$. The O_m ontology will include all the concepts and the relations corresponding to the percept. The elements in O_m conceptualize the percept. The meaning structure becomes

$$T = [sem(m_1, m_2, \ldots, m_m), syn(s_1, s_2, \ldots, s_n)]$$

with $m_i \in C_m \cup P_m$.

2. *Inner verbalization* Once the meanings of the percept are inferred, they are verbalized. It means that the retrieved meanings, representing the emergence of knowledge related to the stimulus (i.e. the experience the robot has about the stimulus), a first thought is elicited. To produce the thought, the meaning structure is modified and consists of just the semantic component including the meanings, so that:

$$T = [sem(m_1, m_2, \ldots, m_m), syn()]$$

and the step rebuilds the structure by associating its syntax to each meaning (i.e., the original labels in the ontology), so that the meaning structure becomes

$$T = [sem(m_1, m_2, \ldots, m_m),$$
$$syn(label(m_1), label(m_2), \ldots, label(m_m))]$$

begin $label(c)$ the function returning the label of the class c in the ontology.
The labels are verbalized, and back-propagated to the inner comprehension component, i.e. they are re-heard. The important aspect is that the production could include additional syntactic tokens related to the meanings that the conceptualization step has merged to the structure, and that could involve a different set of concepts than the tokens of the initial percept.

3. *Inner comprehension* The new labels emerging from the previous verbalization represent new thoughts that could require other perception or other concepts from the knowledge base. The inner comprehension searches for the added labels in the meaning structure in the environment (retrieve) or in the knowledge base (recall). In particular, there are two cases:

   (a) The meaning is related to a property (i.e., $m_i \in P_m$). In this case, the module searches in the knowledge base for the other property's domain or range not contained in the meaning structure (recall), and adds it to the $sem()$ part of the structure, if it is not yet included;
   (b) The meaning is related to a class (i.e., $m_i \in C_m$). In this case, by considering that the classes could be from the previous point, and not from the a_c function which codified the percept, the module requires to search in the environment (retrieve) for the possible entity represented by m_i, and, if the perception returns the entity, the corresponding token is added in the $syn()$ part (and the conceptualization restarts), if it is not yet included, otherwise the module removes m_i. Moreover, the module searches for other properties of the class in the knowledge base (recall), and if they exist, it adds them in the meaning structure if not yet included.

At the end of each phase the meaning structure grows. The loop is repeated until no further chunks are added to the meaning structure. The robot stores the information about the context, including those about the inner world, and uses them for computing the appraisal variables and for triggering the affective evaluation output.

Table 1 shows a use case of the described method. For each involved phase of the loop, the table reports the meaning structure, and the process that modified such a structure. In that example, the robot receives a command by the partner, that is '*take the plate*', and that percept is processed to fill the $syn()$ part of the initial meaning structure. For each phase, the process column shows the concepts and the properties emerging from the knowledge base (recall), or from the environment (retrieve). These entities are highlighted in bold. The production of the sentence (produce) represents the inner verbalization. It is simple to observe how the new entities are added to the meaning structure, and how the structure grows in each next phase.

3.3. The appraisal variables and emotions

The meaning structure is the output of the cognitive evaluation the robot performs with inner speech. Once the loop ends, the values of the appraisal variables can be computed. The model proposes a new mathematical formalization of these variables, based on the trends they should display, as described in [17]. It is worth noting that the proposed mathematical formulations are intended as computational

Table 1
How inner speech works for cognitive evaluation. For each step of the loop, the table reports the meaning structure, the phase of the loop and the related process among recall, retrieve and produce. The loop phases are C for the Conceptualization step, IV for the Inner Verbalization step, and IC for the Inner Comprehension step. The robot perceives the audio stream *take the plate* by the partner. The meaning structure initially includes the chunks related to the meaningful tokens of the codified stream. The loop recalls from the ontology the related concepts and adds them to the structure, and the inner speech starts. All the emergent concepts in each phase are in bold.

Meaning structure	Loop phase	Process
[sem(), syn(take, plate)]	C	Recall - **C:request**, **C:action**, C:take, C:plate, **P:position**, **C:left_box**
[sem(C:request, C:action C:take, C:plate, P:position, C:left_box), syn()]	IV	Produce - Request for action take plate with position left box
[sem(), syn(request, action, take, plate, position, left_box)]	IC	Recall - C:request, C:action, C:take, C:plate, P:position, C:left_box, **P:by**, **C:left_arm**, **P:state**, **D:'hot'**
[sem(C:request, C:action C:take, C:plate, P:position, C:left_box, P:by, C:left_arm, P:state, D:'hot'), syn()]	IV	Produce - Request for action take plate with position left box by left arm with state hot
[sem(), syn(request, action, take, plate, position, left box, by, left arm, state, hot)]	IC	Recall - C:request, C:plate, C:take, C:action, P:position, C:left_box, P:by, C:left_arm, P:state, D:'hot', **P:likelihood**, **D:'20'**
[sem(C:request, C:action, C:take, C:plate, P:position, C:left_box, P:by, C:left_arm, P:state, D:'hot', P:likelihood, D:'20'), syn()]	IV	Produce - Request for action take plate with position left box by left arm with state hot. Event likelihood 20
[sem(), syn(request, action, take, plate, position, left box, by, left arm, state, hot, likelihood, 20)]	IC	Recall - C:request, C:plate, C:take, C:action, P:position, C:left_box, P:by, C:left_arm, P:state, D:'hot', P:likelihood, D:'20', **P:voice**, **C:voice**, **P:noise**, **C:noise**
[sem(), syn(request, action, take, plate, position, left box, by, left arm, state, hot, likelihood, 20, voice, tone)]	C	Retrieve - C:request, C:plate, C:take, C:action, P:position, C:left_box, P:by, C:left_arm, P:state, D:'hot', P:likelihood, D:'20', **P:voice**, **C:voice**, **D:'calm'**, **P:noise**, **C:noise**, **D:'no'**
[sem(C:request, C:action, C:take, C:plate, P:position, C:left box, P:by, C:left_arm, P:state, D:'hot', P:likelihood, D:'20', P:voice, C:voice, D:'calm' P:tone, C:tone, D:'no'), syn()]	IV	Produce - Request for action take plate with position left box by left arm with state hot Event likelihood 20 Voice calm No noise
[sem(), syn(request, action, take, plate, position, left box, by, left arm, state, hot, likelihood, 20, voice, tone)]	IC	Recall - **No new concept to explore** **STOP LOOP**

instantiations constrained by qualitative trends described in appraisal theory and empirically observed in SCPQ-based analyses, rather than as precise quantitative models of psychological variables. In particular, SCPQ characterizes appraisal variables in terms of directional and temporal trends across canonical stressful situations, including: (i) an increase of perceived likelihood as situations evolve towards negative outcomes and a decrease when situations resolve positively; (ii) an inverse relationship between likelihood and controllability; (iii) an increase of changeability in unstable or worsening situations and a decrease as the context stabilizes; and (iv) preservation of the relative ordering of appraisal variables across aversive and loss conditions.

Accordingly, the tuning process was aimed at verifying that the appraisal variables exhibit these SCPQ-consistent directional and temporal dynamics, rather than at fitting specific numerical values. As a

consequence, the behavior of the model depends primarily on relative variations and trends of the appraisal variables, while the specific numerical constants represent one possible implementation-level instantiation. Alternative parameterizations satisfying the same qualitative constraints would therefore lead to comparable appraisal and emotional dynamics.

The formalization of the appraisal variables is based on the particular environmental conditions because, when involving a robot, these conditions can heavily influence its whole evaluation of the context. For keeping in account this issue, the model defines the *environmental entropy* k, meant as an indicator about the lack of order in the environment. The higher the entropy is, the higher the environmental disorder is. It is not intended as a formal information-theoretic measure, but as a qualitative index of environmental disorder affecting the appraisal process. In SCPQ-based analyses, contextual instability and unpredictability are reflected indirectly through variations in appraisal variables such as controllability and changeability, rather than being modeled as independent quantitative constructs. The role of k is to influence the relative evolution of appraisal variables, rather than to provide an absolute measure of entropy. Alternative formulations capturing the same qualitative notion of environmental disorder would therefore be compatible with the proposed appraisal dynamics. To evaluate the entropy, the model takes into account the environmental noise (which could make it difficult to recognize the sound stimulus), and the emotional state of the partner (which could alter his/her approval rating, making the situation more unpredictable).

The robot's routines detect these specified features, and return the values to use in the entropy formalization.

Given e the set of the possible partner's emotions detectable by the robot's routines, they are classified as *negative*, *neutral* and *positive* emotions. The entropy depends on the following parameters: α that is the level of environmental noise, β that measures the partner's emotion inferred by his/her facial expression, and γ that measures the partner's emotion inferred by his/her vocal tone. The parameters are formalized as follow:

$$\alpha = \begin{cases} 0 & \text{noiseless environment} \\ 0.5 & \text{noisy environment} \\ 1 & \text{very noisy environment} \end{cases}$$

and

$$\beta = \begin{cases} -1 & \text{negative by face} \\ 0 & \text{neutral by face} \\ 1 & \text{positive by face} \end{cases} \qquad \gamma = \begin{cases} -1 & \text{negative by tone} \\ 0 & \text{neutral by tone} \\ 1 & \text{positive by tone} \end{cases}$$

By considering that the unpredictability of the environment grows when the noise grows, and when a partner's negative emotion is detected, the environmental entropy k is simply modeled by a linear combination of these parameters, that is:

$$k = -\alpha + \beta + \gamma$$

The linear combination adopted to compute k serves as a pragmatic operationalization of contextual factors that modulate the predictability of the situation. Qualitatively, the expression indicates that environmental disorder increases with higher noise (α) and negative partner affect (β for facial expression, γ for vocal tone). The variable k takes values in the range $[-3, 2]$ given the discrete definitions of α, β and γ. The minimum $k = -3$ corresponds to a highly noisy environment ($\alpha = 1$) combined with negative partner emotions in both face and voice ($\beta = \gamma = -1$), indicating high unpredictability. Conversely, the maximum $k = 2$ occurs in a noiseless setting ($\alpha = 0$) with positive partner emotions ($\beta = \gamma = +1$), reflecting a predictable and orderly context. This interval ensures that k meaningfully captures the continuum from chaotic, negatively charged situations to stable, positively charged ones, consistently aligning with the model's appraisal dynamics as illustrated in the presented use cases. Since k enters other appraisal variables as a modulator rather than an absolute quantity, only its relative variation influences appraisal dynamics. Its effect on subsequent equations is illustrated numerically in the application scenario described in Section 4. The entropy is next used in the formula of the appraisal variables, for making them dependent on the environmental conditions.

Table 2
The work conditions establishing the likelihood of a negative outcome. To a negative condition corresponds a high probability to fail.

Work Conditions		Likelihood L
State of the Component to use	Malfunctioning	80%
	Working	20%
Action feasibility	Not feasible	90%
	Feasible	10%
Rules infringment	Yes	90%
	Not	10%
State of Battery	Low	90%
	Good	10%

Likelihood L

The likelihood L defines the probability that the negative outcome of an event happens, leading to stressful situations [17]. A variation of this value is a sign about the evolution of the event. When it grows, the negative outcome becomes more possible, and the stress grows. On the contrary, when the likelihood decreases, the negative event is resolving, and the stress is reduced.

Thus, the likelihood L by definition lies within the interval $[0, 1]$, reflecting a continuum from no chance ($L = 0$) to certainty ($L = 1$) of a stressful outcome.

In the proposed model, the likelihood is hand encoded for the particular work conditions, because they influence the success/failure of the action to take. The considered work conditions and the corresponding likelihood values are shown in Table 2.

The work conditions include the robot's physical conditions, the valuations about the feasibility of a required action (i.e., is the involved object in the action reachable or visible?), and the admissibility of the action (i.e., does the action follow the rules?).

The likelihood is higher when it corresponds to negative work conditions, because they affect the outcome's feasibility. For example, when a physical component of the robot is not properly working, the likelihood of an event involving that component is fixed to 80% because the probability the event fails is high. If, for some reasons, another component is chosen to work, or the problem of the same component is resolved, the value decreases to 20%.

When more than one work condition occur simultaneously, the final likelihood is the weighted sum of the likelihood values of each occurred condition, with the weight sets to 1 when the likelihood corresponds to a negative condition, and sets to -1 when the likelihood corresponds to a positive condition.

The likelihood values associated with work conditions are not intended as precise probabilistic estimates, but as qualitative indicators of the perceived risk of a negative outcome. In SCPQ-based analyses, likelihood is characterized by its directional and temporal evolution across situations, rather than by absolute values. In particular, likelihood increases as situations evolve towards negative outcomes and decreases when situations resolve positively.

Accordingly, the specific numerical values adopted in the model serve to encode this relative ordering between conditions and to support the evaluation of likelihood variations over time. The appraisal dynamics primarily depend on changes in likelihood across successive evaluations, while alternative numerical instantiations preserving the same qualitative ordering would lead to comparable behavior.

Controllability C

The controllability C measures if the outcome of an event can be modified by directly acting on the context [17]. The controllability variable is not intended as an absolute or objective measure of the agent's ability to influence the environment, but as a qualitative indicator of the perceived possibility to actively modify the outcome of the situation. In SCPQ-based analyses, controllability is primarily characterized by its relational and temporal dynamics with respect to other appraisal variables, rather than by absolute values.

In particular, SCPQ reports an inverse relationship between perceived likelihood of a negative outcome and controllability: as situations evolve towards negative or worsening outcomes, controllability tends to decrease, whereas it increases when situations become more manageable or resolve positively. Accordingly, the proposed formulation of controllability is designed to reproduce this qualitative inverse dependency and its evolution over time.

The specific functional form and parameter values adopted in the model therefore serve to encode this SCPQ-consistent relationship and to support the evaluation of changes in controllability across successive appraisal cycles. The resulting appraisal dynamics primarily depend on relative variations and trends of controllability, while alternative functional forms and parameterizations preserving the same qualitative inverse relationship would lead to comparable appraisal and emotional dynamics.

For formalizing C, the model considers that the higher the likelihood is (i.e., the higher the probability of a negative outcome is), the lower the possibility to modify the context and to resolve the negative evolution is. Similarly, the higher the environmental disorder is, the lower the possibility to control the situation is.

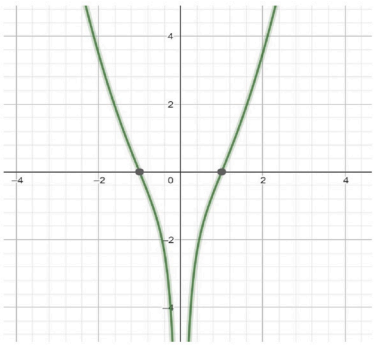
As a consequence, the defined formula for C modeling that trend is:

$$C(k,x) = -\frac{1}{|k|\,x} + x^2$$

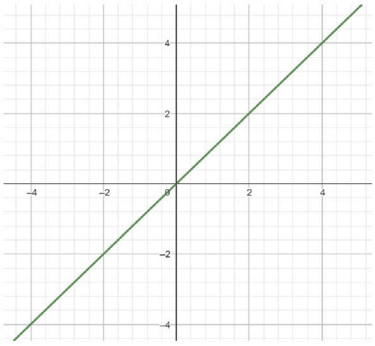
begin $x \in]0,1]$ the absolute value of the variation of the likelihood in two consecutive measurements, that is $x = \left|L_a - L_p\right|$, with $L_a \neq L_p$. L_a is the value of the likelihood in the antecedent measurement, and L_p is the value of the likelihood in the present measurement. L_a is set to zero when the measurement is the first. Since $L_a, L_p \in [0,1]$, it follows that x ranges from 0 to 1. The lower bound $x = 0$ indicates no change in perceived outcome probability, while $x = 1$ corresponds to a complete reversal, from certainty of a negative outcome ($L = 1$) to certainty of a positive one ($L = 0$), or vice versa.

More qualitatively, the parameter x shows how the event is evolving. When it decreases, it means that two consecutive likelihood measurements are close, and the situations are not evolving. As a consequence, the controllability, i.e. the possibility to change the situation, remains the same. When x changes, it means that there was a modification in the situation (the likelihood of the recent measurement differs from the previous one), and the possibility to intervene depends on the kind of the x variation. When x decreases for negative values or increases for positive values, it means that the likelihood is decreasing, and the situation is becoming less bad. As a consequence, it could be more controllable, and the controllability grows. On the contrary, when x decreases for positive values or increases for negative values, it means that the likelihood is growing and the situation is becoming less controllable. Thus, the controllability decreases.

Fig. 5(a) shows the trend of the C function. By changing k, the width of the projection of the C on the x-axis changes, that is the higher k is, the faster the controllability decreases, because the disorder makes the situation less controllable. However, the trend of the C chart remains the same independently from the k value. The independence of the trends of the C function from the k value is next detailed.



(a)



(b)

Fig. 5. (a) The curves of the $C(k,x)$ function varying x for $k = 1$. (b) The curve of the $M(k,x)$ function varying x with $k = 1$.

Changeability M

The changeability M measures if the outcome of an event can be modified by another external event, for which the control does not depend by the individual [17]. It represents a measure of how unpredictable the context is, because the outcome does not depend by the self. According to this definition, it is trivial to consider that the higher the probability of a negative outcome is, the higher the possibility that the situation becomes more unstable is. Similarly, the higher the entropy is, the more unpredictable the situation is. As a consequence, M is formalized as:

$$M(k,x) = |k| \star x$$

begin x defined as the previous case, that is $x = \left|L_a - L_p\right|$. The changeability is a straight line, as represented in Fig. 5(b). The k value influences the inclination of this line, which anyway remains in the first and fourth quadrants. The lower k is, the higher the inclination is (over the 90 grades), making the variation of the changeability faster. It is in line with the fact that when the entropy is high, the situation is more unforeseeable.

When x increases for positive or negative values, the changeability grows because the situation is aversively changing and the context is becoming more unpredictable. On the contrary, when the likelihood decreases for positive or negative values, it means that the context is becoming more stable, and the situation is less modifiable by unpredictable events. The independence of the M trend from the k value is next detailed.

Desirability D

The desirability D represents if an outcome of an event is desirable or not. It can be positive or negative [9], depending on the expected outcome.

Simply:

$$D = \begin{cases} 1 & \text{if the outcome is desirable} \\ -1 & \text{if the outcome is not desirable} \end{cases}$$

Integrating specific appraisal variables

Beyond the general appraisal variables (likelihood (L), controllability (C), changeability (M), and desirability (D)), the architecture allows the introduction of additional scenario-dependent appraisal variables. These variables are meant to capture contextual aspects that cannot be foreseen in the general formulation of the model. Each variable may be defined through any mathematical structure—continuous or discrete, linear or nonlinear-depending on the phenomenon it is intended to represent. The only requirement is to specify whether its contribution to the global valence is positive or negative.

Let us denote such a context-dependent variable generically as X, whose value is obtained by applying a function f to the relevant contextual parameters:

$$X = f(\theta_1, \theta_2, \dots, \theta_n),$$

where $(\theta_1, \theta_2, \dots, \theta_n)$ are domain-specific quantities selected by the designer. No assumption is made about the form of f: it may be a normalized distance, a counter, a rate, a probability, a penalty, or any other task-appropriate mapping. The architecture only prescribes how such variables are subsequently integrated into the affective evaluation.

Once defined, specific variables can be seamlessly incorporated into the global valence computation. If two such variables X_1 and X_2 are introduced in a given context and their qualitative contributions have been specified, the resulting valence can be written as:

$$\upsilon(L, C, M, D, X_1, X_2) = N(C - M + D + \alpha_1 X_1 + \alpha_2 X_2),$$

where α_1 and α_2 modulate the weight of each specific variable. This formulation illustrates that the architecture does not impose structural constraints on the mathematical form of the variables; rather, it prescribes how they are combined once defined, allowing the designer to refine the emotional dynamics of the robot with minimal effort.

Example instantiations of X. To illustrate the expressive flexibility of the formulation, three independent examples of specific appraisal variables and their integration through appropriate coefficients α are presented. These examples represent possible instantiations depending on design needs.

1. Distance-based variable: Let $\theta_1 = d$ be the Euclidean distance between the robot and a goal location, and ρ the maximum relevant distance. A possible specific appraisal variable increases as the robot gets closer to the target, thereby contributing positively to the valence. A possible definition is:

$$X_1 = f_1(d) = 1 - \frac{d}{\rho}.$$

A suitable weight may be assigned, as it contributes positively to the valence, e.g.,

$$\alpha_1 = 0.4.$$

The amount of this value depends on how its contribution is considered relevant by the designer.

2. Time-pressure/urgency variable: Let $\theta_1 = t_r$ be the remaining time to complete a task and T the maximum allowed duration. A designer may define:

$$X_2 = f_2(t_r) = 1 - \frac{t_r}{T},$$

which increases as time becomes scarce. A higher sensitivity may be appropriate here, such as:

$$\alpha_2 = -0.7.$$

3. Risk accumulation variable: Let $\theta_1 = k$ be a risk indicator, with $k_{\max}$ the maximum expected level. A simple formulation is:

$$X_3 = f_3(k) = \frac{k}{k_{\max}}.$$

If risk should strongly influence the emotional state, one may choose:

$$\alpha_3 = -1.0.$$

A concrete instantiation of two specific appraisal variables used in this study, together with their explicit mathematical definitions and integration into the valence function, is provided in Section 4.

Matching the appraisal variables to the emotions

Once the general appraisal variables are computed, they are be combined for inferring the corresponding emotion. For this purpose, the model refers to the bi-dimensional Russell's space [18], in which the emotions are point whose coordinates are the *valence* υ and the *arousal* a, each falling in the range $[-1, 1]$. The valence means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event. The arousal represents the physiological activation when facing an emotional experience. According to these definitions, the proposed model uses the values of the appraisal variables for computing the valence, because the modeled appraisal variables keep in account the conditions of the situation and hence its attractiveness/aversiveness. Instead, the model computes the arousal by referring to the inner states of the robot, representing the robot's physiology.

More specifically, a situation has a positive valence when the controllability is high, because it is an indicator about the possibility to control and change the context for improving the likelihood of a negative outcome. In the same way, the desirability of an outcome contributes positively to the valence. Instead, a high value of the changeability implies that the context is unpredictable, and it contributes to a negative valence.

In view of these considerations, the matching between the valence and the appraisal variables is modeled by a linear dependence, that is:

$$\upsilon(C, M, D) = N(C - M + D)$$

where C, M, and D are the controllability, changeability and desirability appraisal variables, that is $C = C(k, x)$, $M = M(k, x)$ and $G = g(\Delta)$, and $N(.)$ is the function that returns the normalized value in the range $[-1, 1]$ of its argument. In particular, the $N(.)$ function is formalized as follow:

$$N(arg) = \frac{(arg - r_1)(n_2 - n_1)}{r_2 - r_1} + n_1$$

begin r_1 and r_2 the original range extremes, and n_1 and n_2 the new range extremes into which the values must fall. The normalization enables to project the valence to the SCPQ and EMA spaces.

The arousal is modeled by the level of the battery and the inner temperature of the robot. These are just the two plausible physiological conditions of the robot. The arousal is more active when the level of the battery and the inner temperature are optimal, that is the battery is charged and the temperature is not hot. Both values are detectable by the typical robot's routines which monitor the robot's functioning. As a consequence, the arousal is:

$$a(B, T) = N(B - T)$$

where B is the level of the battery, and T is the inner temperature.

The model infers the emotion by projecting the *appraisal pattern* $\mathbf{a} = (\upsilon, a)$ in the Russell's space.

Recent studies have shown that mapping continuous valence-arousal representations to discrete emotion categories can be challenging, and that richer affective interpretations can be obtained through free-form semantic descriptions (e.g. [35]). In particular, such approaches demonstrate that continuous affective dimensions support nuanced affective inference when the output is not restricted to a fixed set of categorical labels.

Table 3
The matching of the values of the valence v and the arousal a with the labels $l(v,a)$ of the five basic emotions by Ekman in the Russell's space as defined at [34].

	Valence v	Arousal a	Emotion label $l(v,a)$
1	−0.40	0.79	*Angry*
2	0.89	0.17	*Happy*
3	−0.12	0.79	*Afraid*
4	−0.44	0.76	*Annoyed*
5	−0.81	−0.40	*Sad*

The present work is consistent with this perspective in that emotional states are represented internally in a continuous valence-arousal space. However, for the purposes of embodied robotic expression and human-centered evaluation, a restricted set of emotion labels (specifically, the basic Ekman emotions) is adopted as the output interface. While this choice necessarily limits expressivity, it enables robust and interpretable emotional behavior in the task-oriented interaction scenarios considered here. Extending the output layer towards semantic or free-form affective descriptions represents a natural direction for future work.

Being E the set of the labels of the Ekman's emotions (i.e., $E = \{Angry, Happy, Afraid, Annoyed, Sad\}$), and being R the set of coordinates in the Russell's space corresponding to each emotion in E, the element $\mathbf{r}_i \in R$, with $\mathbf{r}_i = (v_i, a_i)$, is the couple of coordinates corresponding to the ith emotion in E, as shown in Table 3. These coordinates are defined at [34].

The function $l(\mathbf{r}_i)$ returns the label of the emotion in E corresponding to the point in R passed as argument. The emotion $e_{\mathbf{a}}$ corresponding to an instance of the appraisal pattern $\mathbf{a}$ is the following:

$$e_{\mathbf{a}} = \{l_i(\mathbf{r}_i) \mid min\|\mathbf{r}_i - \mathbf{a}\|,\ \mathbf{r}_i \in R\}$$

that is, the emotion is the element in E whose coordinates in R are at the minimum Euclidean distance from $\mathbf{a}$.

The computation of the emotion's intensity

The emotion's intensity represents the level of feeling the emotion in a more or less vivid and profound way.

The model considers three levels to which correspond three different reactions in feeling emotions: an intensity of *high level* corresponds to emotions that are difficult to control and regulate, and leads, for example, to feeling bursts of joy or anger. The *medium level* pertains tp emotion that could be regulated and managed, while the *low* or *null level* represents little emotional involvement, almost indifference.

In the proposed model, the intensity of the emotion $e_{\mathbf{a}}$ corresponding to the appraisal pattern $\mathbf{a}$ is computed by referring to two threshold values, that are the high threshold t_h and the low threshold t_l. They are fixed respectively to the quarter and to the half of the Euclidean distance d in the Russell's space between the point $\mathbf{r}_i \in R$ corresponding to the emergent emotion $e_{\mathbf{a}}$ and the origin $\mathbf{o} = (0, 0)$ of the space.

More formally:

$$t_h = \frac{d(\mathbf{r}_i - \mathbf{o})}{4}, \quad t_l = \frac{d(\mathbf{r}_i - \mathbf{o})}{2}$$

The corresponding intensity will be:

- *High intensity*: when $\mathbf{a}$ falls within the circumference whose center is the point r_i corresponding to $e_{\mathbf{a}}$ and whose radius is t_h;
- *Medium intensity*: when $\mathbf{a}$ falls between the two circumferences centered in r_i with radius t_h and t_l respectively;
- *Low intensity*: when $\mathbf{a}$ falls outside the two circumferences above.

To each intensity corresponds a different response by the robot. The response is related to the vocalization of the state of feeling. In particular, the robot expresses the high intensity by the adjective *very* to the label of the emergent emotion, the medium intensity by using just the label, and the null intensity by one of the adjectives *a bit*, *little* and synonyms to the label.

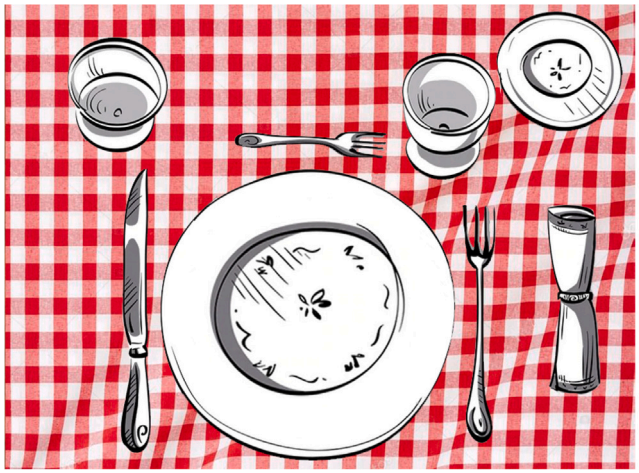


Fig. 6. The etiquette schema to follow for setting up the table. The human partner and the robot must place the utensils according to the schema.

4. An application: setting up the table with a human partner

4.1. Narrative description of the scenario

The robot and a human partner collaborate to set up a table according to formal etiquette rules. This scenario is the same analyzed by some of the authors in [12], where the first interesting results, about the robot's inner speech effects on human–robot cooperation, are discussed. It provides a concrete context for observing how the robot's inner speech drives affective appraisal in a structured cooperative task. The interaction unfolds through a tablet-based interface where utensils must be placed in specific positions following an etiquette schema (Fig. 6). The schema establishes precise placement rules for each item (plates, glasses, cutlery) defining what constitutes correct execution of the task.

The interactive dynamics

The human partner interacts with the system by dragging and dropping utensils onto a virtual tablecloth displayed on the tablet. Each placement represents a micro-event that the robot must evaluate. The partner can either follow the etiquette rules correctly or violate them by placing items in wrong positions. Additionally, the partner can verbally request that the robot perform specific actions, such as "take the plate" or "place the fork on the left". The robot's role is dual: it must execute actions when requested by the partner, and it must continuously appraise the unfolding situation to generate appropriate affective responses. Importantly, a constraint is built into the scenario: each utensil can be moved only twice, creating a temporal pressure that amplifies the affective significance of errors and corrections.

The robot used is Pepper by Aldebaran[2] and all the information related to the work conditions are inferred from the suitable Pepper's API.[3]

The APIs enable to detect the external elements (the tone voice and the facial expression of the partner, as well as the environmental noise conditions), and how the robot perceives them. The typical interval values representing the conditions are described in the APIs' documentation. For example, the class `ALMood`[4] allows to detect the mood of the partner and to evaluate it.

[2] https://www.aldebaran.com/it/pepper
[3] http://doc.aldebaran.com/2-5/naoqi/index.html
[4] http://doc.aldebaran.com/2-5/naoqi/core/almood-api.html

Inner speech in action

When an event occurs (for example, when the partner places the bread plate in an incorrect location) the robot's inner speech mechanism activates. The robot does not simply register the error as a binary fact; rather, it engages in a verbalized internal dialogue that explores the meaning and implications of what has happened.

The inner voice might articulate thoughts in sequence such as: *"Request action for bread plate in wrong location"*, *"Event rule infringement with likelihood 90%"*, *"Distance from correct location high"*, *"Event battery level good with likelihood at 10%"*, *"Update likelihood to 80%"*, *"Voice harsh. Little noise"*..

This stream of verbalized reasoning reflects the robot's attention moving through different aspects of the situation: the nature of the action itself, the physical distance from the correct position, the robot's own internal state (battery, motor conditions), and environmental factors (partner's emotional tone, ambient noise). Each element contributes to building a comprehensive appraisal of the context. The inner dialogue follows the rehearsal loop described in Section 3.2. Initial perception triggers conceptualization (i.e. retrieving concepts like "action take", "bread plate", "final position" from the knowledge base). These concepts are then verbalized, generating an internal utterance that is "re-heard" by the system, prompting further semantic expansion. New concepts emerge like: "expected position", "likelihood", "work conditions", "voice tone" each bringing additional information that refines the appraisal.

Specific appraisal variables in the table-setting context

The general appraisal variables defined in Section 3.3 (likelihood, controllability, changeability, and desirability) are computed based on the information gathered through inner speech. As discussed above, some specific appraisal variables can be added with the aim to consider the emotional contributions of specific aspects of the context. For example, in this scenario the following aspects should be kept in account:

1. Did the partner place the utensil in the correct final location?
2. If the partner was wrong, is the final position too far away from the correct one?
3. How many tentative actions remain for correctly placing the utensil?
4. Does the partner ask the robot to take a correct or a wrong action?

These aspects affect the global computed affective state, and they should be modeled as specific appraisal variables that contribute to the final emotion. An error made by the partner has a negative valence, as the position of the utensil is farther from the expected position, while the possibility to fix the situation, that is to make another tentative action, contributes positively. In the same way, when the partner requests for a correct/incorrect action from the robot, the contribution to the valence is positive/negative. The specific appraisal variables are:

- The Gradient Variable (G): This measures how far the incorrectly placed utensil is from its expected position. The greater the spatial error, the more negative the contribution to the overall affective state. A bread plate placed just slightly off-center has a different affective weight than one placed at the opposite end of the table.
- The Recovery Variable (R_u): This tracks how many attempts remain for correcting a placement error. Since each utensil can only be moved twice, this variable directly modulates hope or frustration. When one correction attempt remains, the situation feels recoverable; when no attempts remain, the error becomes irreversible, intensifying negative affect.

Emotional outcomes

Through the interplay of these variables, coherent emotional states emerge. Consider two contrasting situations:

1. Error with Harsh Tone: The partner places the bread plate in the wrong location (where the water glass should be) and speaks with a harsh tone. The inner speech process reveals high likelihood of a negative outcome (rule infringement), but also identifies that correction is still possible (one attempt remains) and that the robot's physical state is optimal (battery charged, motors functioning well). However, the harsh vocal tone of the partner contributes to environmental entropy, reducing perceived controllability. The resulting emotion is annoyance with medium intensity—a negative but not overwhelming affective response that reflects both the error and the remaining possibility for correction.
2. Successful Correction: The partner moves the same bread plate from its incorrect position to the correct one. Inner speech now identifies decreasing likelihood of negative outcomes, as the rule infringement is being resolved. Although environmental noise is high (reducing controllability somewhat), the desirability is strongly positive (D = 1), and the gradient reaches its maximum value since the utensil is now correctly placed. The recovery variable drops to zero (no further moves possible), but the correctness of the action dominates the appraisal. The resulting emotion is happiness with medium intensity—a positive affective state that acknowledges both the resolution of the error and the acceptable environmental conditions.

The role of physical and environmental state

Crucially, the robot's affective evaluation is not determined solely by task-relevant events. The robot's physical state (battery level, motor temperature) continuously feeds into the arousal dimension of the emotional space. A robot with low battery and overheating motors will experience higher arousal (physiological activation) even when task events are neutral. Similarly, environmental factors such as ambient noise and the partner's emotional expression (detected through facial and vocal cues) modulate the entropy parameter, which in turn affects how controllable and predictable the situation appears.

This integration of internal and external states through inner speech creates affective evaluations that are contextually sensitive, transparent to observers (since the inner speech is verbalized aloud), and grounded in a structured cognitive process.

4.2. Mathematical formulation and computational details

4.2.1. Ontology specification

The ontology used in this scenario provides sufficient coverage for the table-setting and human–robot interaction scenarios tested. In particular, it contains:

- $C_o = 94$ classes modeling entities in the external world (e.g., Person, Utensil, Location, Action) and internal robot states (e.g., Emotion subtypes, BatteryLevel, ArmState)
- $P_o = 33$ object properties defining spatial relations (left, right, up, down), functional relations (by, with), and domain-specific relations (e.g., position, state)
- $P_d = 8$ datatype properties linking concepts to numerical or categorical values (e.g., likelihood, temperature)

The complete ontology structure is available in OWL format in the project's GitHub repository (see Resource Availability section 8), enabling full inspection of its classes, properties, axioms, and hierarchical organization.

4.2.2. Specific appraisal variables

According to the general formulation introduced above in , additional appraisal variables can be instantiated as $X_1, X_2, \dots$ and integrated into the valence computation through the weighted contribution specified in the generic expression:

$$v = N(C - M + D + \alpha_1 X_1 + \alpha_2 X_2 + \dots).$$

In this scenario, we instantiate two specific appraisal variables by assigning:

$$X_1 \equiv G \quad \text{and} \quad X_2 \equiv R_u,$$

where G is the *gradient* variable and R_u is the *recovery* variable.

The gradient variable

The contribution of the wrong action by the partner to the valence, is formalized by evaluating the percentage of errors between the expected and the wrong positions. It measures how wrong the partner is in respect to the maximum possible error, that is the maximum distance between two locations according to the etiquette schema (i.e., the distance between the two furthest locations). From the beginning of the virtual table, the distance is computed by considering the coordinates of the pixel on the virtual table representation. The gradient variable is based on that percentage of errors, and is the following function:

$$G \equiv g(\Delta) = \begin{cases} \frac{\rho/2 - 1.2 * \Delta}{\rho} & \text{if } \Delta \neq 0 \\ G_{max} & \text{if } \Delta = 0 \end{cases}$$

where ρ is the maximum possible Euclidean distance the utensils could be positioned, and Δ is the Euclidean distance d between the final position P_f of the utensil as placed by the partner, and the expected correct position P_e. The function models that the further the utensil is placed from the correct position, the lower the gradient value becomes. When $\Delta = 0$, the utensil is correctly positioned and the gradient returns its maximum value G_{max}, which is set to 13 following the tuning phase of the architecture. The higher the gradient is, the higher the valence should be.

The choice of $G_{max} = 13$ is the result of a systematic calibration process. During the design phase, several candidate values were tested to determine the most suitable scaling factor for the gradient variable. The selected value had to satisfy three constraints simultaneously:

1. produce a sufficiently high reinforcement in perfectly correct configurations;
2. avoid dominating or saturating the combined contribution of the general appraisal variables (C, M, D); and
3. amplify the perceptual difference between minor and major deviations, so that even small placement errors remain affectively distinguishable.

Empirical evaluation showed that 13 is the smallest value that satisfies all these requirements, yielding an optimal balance between expressiveness and stability. It provides a clear affective separation between correct and incorrect states while maintaining numerical coherence with the rest of the appraisal model.

Importantly, this value is not tied to the specifics of the scenario, but rather to the internal calibration of the gradient function within the proposed appraisal architecture. Alternative implementations may choose different scales; however, in our formulation, 13 proved to be the most effective choice for ensuring both robustness and interpretability of the resulting valence.

The recovery variable

The recovery variable measures the remaining tentative actions for placing a utensil. It is simply the counter tracing the tentative action for each utensil, so it can assume a value in the set $R_u = \{1, 0\}$ begin u one of the utensils to place (1 when there is another tentative action to fix the position of the utensil, 0 when there are no further tentative action). The higher the recovery is, the higher the valence should be.

Contributions to the valence

Once the specific appraisal variables and their qualitative contributions have been defined, their effect on the global valence follows the weighted combination introduced in the general formulation. In this scenario, the two specific appraisal variables instantiated as $X_1 \equiv G$ (gradient) and $X_2 \equiv R_u$ (recovery) contribute with unitary weights. This choice reflects a design decision aimed at keeping the contribution of the scenario-dependent variables directly comparable to that of the general appraisal components. Setting the coefficients to one also avoids introducing additional tuning parameters at this stage, ensuring that the impact of G and R_u on the valence is transparent and solely determined by their mathematical definition. Nonetheless, the architecture permits assigning different weights in future extensions, should a finer affective modulation be required. Accordingly, the numerical coefficients are:

$$\alpha_1 = 1, \qquad \alpha_2 = 1.$$

Given that $C = C(k, x)$ and $M = M(k, x)$, the resulting valence becomes:

$$v(C, M, D, G, R_u) = N\left(C - M + D + 1 \cdot G + 1 \cdot R_u\right),$$

which corresponds to the formulation adopted in this work.

4.3. The model at work

The two situations previously introduced, with different conditions of the context and different events, are now computational detailed.

> *Use case I.* The partner asks the robot to make an incorrect action in a harsh voice tone. The environment presents little noise, and the inner state of the robot (including the battery and the inner temperature levels) is optimal (the battery is charged at 80% and the level of the temperature is 1 indicating a good state according to the APIs).
> *Use case II.* An object is already on the table in a wrong location. The partner moves that object and places it in the correct location. The environment is very noisy. The motors of the robot are a little overheated and the level of battery is at 75%.

Use case I

In the use case I, the incorrect action pertains to the bread plate. The user asks the robot to place it on the location intended for the water glass. The robot's inner speech processes a recall of the concepts related to the bread plate and its position, and infers the error. Moreover, the concepts related to the inner state and the external conditions (the partner mood and the environmental state) are retrieved as shown in Section 3.2.

For the use case I, the inner dialogue of the robot looks like the following sentences:

- *Request action for bread plate in a wrong location*
- *event rule infringement likelihood 90%*
- *distance from correct location high* (the distance is more than 456.13)
- *event battery level good likelihood at 10%*
- *update likelihood to 80%*
- *event work conditions good likelihood 10%*
- *update likelihood to 70%*
- *Voice harsh. Little noise* (the robot infers the entropy parameters from voice $\gamma = -1$ and from noise $\alpha = 0.5$)

It is important to notice that the numerical values of the parameters produced by inner speech are replaced by the qualitative form (high, medium, low). The qualitative representation depends on the intervals defined by the typical robot's API. In this way, the thoughts of the robot become more understandable to the partner. The emergent parameters

by the inner dialogue are then used for the entropy calculation, according to the mathematical formalization of the model. For the use case under investigation, the entropy k will be:

$$k = -\alpha + \beta + \gamma = -0.5 + 0 - 1 = -1.5$$

The expected position of the bread plate corresponds to the coordinates $P_e = (446, 51)$, while the wrong final position is at the coordinates $P_f = (902, 62)$. As consequence, the Δ value, representing the Euclidean distance d between these positions, will be

$$\Delta = d(P_e, P_f) = \sqrt{(446 - 902)^2 + (51 - 62)^2} = 456.13$$

and the gradient $g(\Delta) = -0.30556$ by considering the ρ parameter is the maximum possible distance in the virtual table, that is 679.471. The recovery variable R_u is 1 because the partner makes the request for the first time, and another possibility to fix the error exists. Given the likelihood value inferred by the inner speech, that is $L_p = 0.7$, and considering that the state is initial and $L_a = 0$, the x variable will be $x = \|L_a - L_p\| = \|0 - 0.7\| = 0.7$. As a consequence, given the controllability $C(k, x) = \frac{-1}{\|kx\|} + x^2$ and the changeability $M(k, x) = \|k\|x$, with these parameters they become respectively:

$$C(1.5, 0.7) = -\frac{1}{\| - 1.5 * 0.7\|} + 0.7^2 = -0.4623$$

and

$$M(1.5, 0.7) = \| - 1.5\| * 0.7 = 1.05.$$

The desirability D of the executed action is negative because it corresponds to a rule infringement, so $D = -1$ With these appraisal variables, the model projects in the Russel's space the values of the valence v and the arousal a that are:

$$v(C, M, D, G) = N(C(-1.5, 0.7) - M(-1.5, 0.7) +$$
$$+D + g(\Delta) + R_u) = N(-0.4623 - 1.05 - 1 - 0.3055 + 1) =$$
$$= N(-1.8179) = -0.5572$$

and

$$a = N(B - T) = N(0.8 - 1) = N(0.2) = 0.4$$

because $B = 80\%$ and $T = 1$ as inferred by the use case narration. The normalized values of valence and the arousal are then projected in the Russell's space and negative emotion *annoyed* with medium intensity emerges.

Use case 2

The partner fixes a rule infringement, moving an object from the incorrect location to the correct one. In particular, the use case is the continuation of the previous use case: the partner places the bread plate from the water glass location to correct position. The robot infers that the rule is now respected via inner speech, and a positive emotional state representation should emerge. The inner sentences look like:

- *Action execution for bread plate in the correct location*
- *event rule infringement likelihood 10%*
- *event battery level good likelihood 10%*
- *update likelihood to 20%*
- *event work conditions good likelihood 10%*
- *update likelihood to 30%*
- *Big noise* (the robot infers the α parameter that is $\alpha = 1$)

The entropy k becomes

$$k = -\alpha + \beta + \gamma = -1.0$$

and the Δ parameter is 0 because the final and the expected locations match, and the gradient $g(\Delta)$ corresponds to the maximum value G_{max} according to its mathematical formalization, that is $G_{max} = 13$.

The fact that the utensil is already on the table, means that the recovery variable R_u is 0, because the utensil was already moved on the table, and after this action no further tentative action exists. The appraisal variables are computed by considering the likelihood variations, that are $L_p = 30\%$ and $L_a = 70\%$. As a consequence, the x parameter is $x = \|0.7 - 0.3\| = 0.4$.

By applying the model's formulas as in the previous case, the controllability and the changeability are

$$C(-1, 0.4) = -\frac{1}{\| - 1.0 * 0.4\|} + 0.4^2 = -2.34$$

and

$$M(-1, 0.4) = \| - 1.0\| * 0.4 = 0.4.$$

The desirability of the action is positive because it fixes a rule infringement, so $D = 1$.

The final valence and arousal are:

$$v(C, M, D, G, R_u) = N(C(-1, 0.4) - M(-1, 0.4) + D +$$
$$+g(0) + R_u) = N(-2.34 - 0.4 + 1 + 13 + 0) = N(11.26) = 0.6537$$

and $a = N(B - T) = N(0.65 - 1) = N(-0.35) = 0.325$ The projection of the normalized values in the Russell's space corresponds to the *happy* emotion with medium intensity, as one might expect in the described scenario.

Table 4
The appraisal variables computed by the model for the proposed use cases.

Use case	*g(Δ)*	R_u	C	M	D
I	−0.30556	1	−0.1541	1.05	−1
II	13	0	−2.34	0.4	1

Table 5
The appraisal patterns and the corresponding emotions with intensity for the proposed use cases.

Use case	*v*	*a*	*l(v,a)*	*in*
I	−0.5572	0.4	Annoyed	Medium
II	0.6537	0.325	Happy	Medium

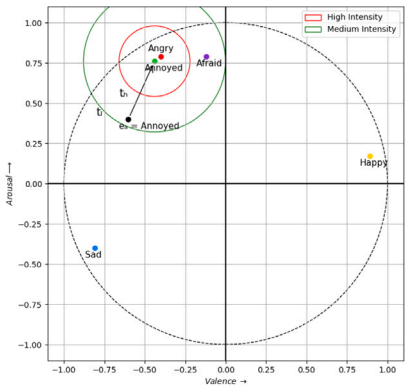
Assessing the robot's emotional behavior

The appraisal variables computed according to the proposed formalization in the uses are shown in Table 4. Fig. 7 shows the projections of the appraisal patterns in the Russell's space, and the emergence of the corresponding emotion with a specific intensity. The projected values, related to the appraisal variables, are summarized in Table 5. The controllability C is higher in the use case I than in the use case II because the conditions to fix the situation exist (the further tentative action and the optimal conditions of the robot). The same considerations regard the changeability M because the robot might expect that other event (for example, a initiative action by the partner) could fix the rule.

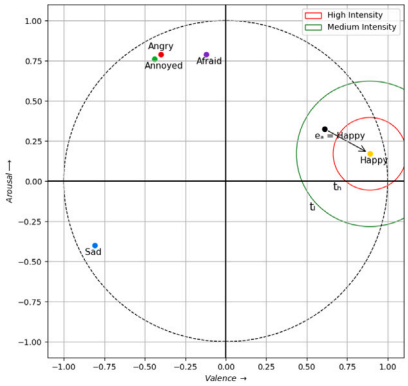
In the use case I the entropy is negatively influenced by the rule infringement, which also generates a low gradient which in turn negatively influences the valence. Moreover, the valence decreases due to the negative tone of the partner. The arousal is good, but for the low value of the valence, the medium annoyed reaction emerges, that is the expected emotive experience in a situation like this.

In the use case II, even if the controllability and the changeability are lower because the event cannot be changed, the emotional reaction becomes positive because the event is desirable (i.e., the executed action is correct) and the error is fixed, and the valence becomes more positive given the gradient contribution. The behaviors generated by the model in both the analyzed use cases are in line to the situations, and correspond to the expected ones according to common sense.

Many other examples were analyzed by simulating other kinds of situations in the cooperative scenario under investigation, and in each case the robot's reactions have been evaluated as corrects.



(a) Use Case I



(b) Use Case II

Fig. 7. The projections of the appraisal variables in the Russell's space, and the computation of the emergent emotion with its intensity.

5. Evaluations

Given the high subjectivity of feeling emotions, it is not trivial to validate and test a model implementing an emotional behavior. Moreover, the processes related to inner speech and emotions are not directly accessible and observable in humans, and to standardize the emotional responses remains (and probability will always remain) an open issue.

With the aim to make the evaluation reliable, two different experimental setups were carried out. The first avenue attempts to evaluate the model qualitatively with respect to its coherence with the appraisal theory and the human's emotional behaviors: in this case, the appraisal variables of the model were compared both to the human's emotional trends from observations of previous studies, and to the variables of one of the existing computational model of emotion. The stressful situations as described in the mentioned studies were reproduced and simulated, leading to the comparative evaluation. The second avenue evaluates the model empirically through actual human–robot interactions where the robot's behaviors were assessed by humans. In this case, a set of scenes representing interactive acts were shown to the participants, and they were engaged in evaluating the emergent robot's emotional behavior. Both experimental setups are next detailed.

5.1. The comparative evaluation

The authors in [19] proposed a strategy for evaluating the computational model of emotions, consisting of directly comparing the appraisal variables of the model to human data, that are collected with the Stress and Coping Process Questionnaire (SCPQ scale) [17]. The scale is an instrument for abstracting a general emotional human behavior in stereotypical stressful situations.

The scale aggregates the answers provided by the participants across canonical stressful situations. It abstracts a general profile of how the humans tend to appraise the context, tend to emotionally react to the appraised situation, and finally tend to apply a coping strategy. The comparison of the computational model's trends with these trends enables the model evaluation.

Applying this strategy has double benefits: the results are comparable to the trends of human's emotional behaviors when the robot operates in the same context, and also they are comparable to the results provided by EMA's authors that defined such an evaluation method. The way the scores are computed is next detailed.

5.1.1. The SCPQ scale

The SCPQ scale [17] by Perrez and Reicherts is a clinical instrument consisting of several narrative episodes and related questions for abstracting the trends of emotional behaviors in normal, healthy, adult humans when facing the stressful situations described in a set of episodes.

SCPQ episodes

The SCPQ test formalizes the episodes by a grammar that defines four prototypical situations. Each episode falls in one of these situations, which are *aversive-good*, *aversive-bad*, *loss-good*, and *loss-bad*. More specifically, in the aversive situations, some bad outcome has occurred but there is some potential to fix it. In the loss situations, a potential loss could occur in the future. Both kinds of situations can have a positive and negative resolution, defining the four aforementioned prototypical cases of the SCPQ test:

1. *aversive-good*: when the bad outcome is fixed;
2. *aversive-bad*: when the bad outcome occurs, and nothing fixed it.
3. *loss-good*: when the loss is averted;
4. *loss-bad*: when the loss occurs.

The narrative of the episodes is based on a canonical structure that models the temporal evolution of a stressful situation. Time is treated as discrete, and each episode unfolds through three successive phases:

> *Phase 1 - Start*: an initial situation is described and something could occur (among aversive or loss);
> *Phase 2 - Middle*: nothing happens, and the context does not change for some time;
> *Phase 3 - Good or Bad*: something happens, and the situation is resolved in a good or bad way.

The complete version of the SCPQ comprises 18 episodes, nine for each episode type, each articulated into three phases.

The different temporal granularity adopted for appraisal variables and valence in the SCPQ scale reflects their distinct theoretical roles in the emotion generation process. In appraisal theories of emotion, variables such as controllability and changeability represent cognitive evaluations of the situation that arise early during an episode and tend to stabilize once the individual has formed an initial interpretation of the event. Accordingly, the SCPQ scale operationalize these appraisal dimensions as relatively stable judgments that can be meaningfully assessed at the beginning of an episode and after an initial update, corresponding to the Start and Middle phases. Valence, by contrast, is not an appraisal dimension per se but a global affective outcome that emerges from the interaction of multiple appraisals and is strongly modulated by the unfolding and resolution of the event. As a consequence, valence evolves more continuously over time and is particularly sensitive to the final outcome of the episode. The inclusion of a third phase for valence therefore allows to capture outcome-dependent affective shifts, such as relief or disappointment, which would be missed by a two-phase representation.

Table 6
The mean values of the appraisal variables in the canonical situations, for adults as reported in the SCPQ book and for the EMA model, that are used as the baseline for comparison in the evaluation of the proposed model.

Aversive				
Controllability	*Start*	*Middle*		
EMA	3.25	1.6		
SCPQ	3.15	2.73		
Changeability	*Start*	*Middle*		
EMA	3.4	1.2		
SCPQ	1.76	1.51		
Valence	*Start*	*Middle*	*Good*	*Bad*
EMA	3.3	4.2	0	5
SCPQ	2.57	2.73	1.08	2.79
Loss				
Controllability	*Start*	*Middle*		
EMA	0	0		
SCPQ	2.57	2.08		
Changeability	*Start*	*Middle*		
EMA	2.1	1.2		
SCPQ	1.41	1.12		
Valence	*Start*	*Middle*	*Good*	*Bad*
EMA	2	3	0	5
SCPQ	2.75	2.65	2.99	0.9

Table 7
An example how inner speech works for cognitive evaluating the canonical loss-good situation. The initial meaning structure contains the syntactic node representing such a stressful event, that enables to retrieve the corresponding likelihood values by inner speech. The phase of the loop and the related process among recall, retrieve and produce are reported according to the notation shown in the sub-Section 3.2.

Meaning structure	Loop phase	Process
[sem(), syn(stressful_event, id2)]	C	Recall - C:stressful_event, C:id2, **P:likelihood**. **D:'50'**
[sem(C:stressful_event, C:id2, P:likelihood, D:'50'), syn()]	IV	Produce - Stressful event with likelihood 50
[sem(), syn(stressful_event id2, likelihood, 50)]	IC	Recall - C:stressful_event, C:id2, P:likelihood D:'50' **P:likelihood D:'0'**
[sem(C:stressful_event, C:id2, P:likelihood, D:'50', P:likelihood, D:'0'), syn()]	IV	Produce - Stressful event with likelihood 50 and likelihood 0
[sem(), syn(stressful_event id2, likelihood, 50, likelihood 0)]	IC	Recall - No new concept to explore STOP LOOP

SCPQ questionnaires

The SCPQ instrument defines questionnaires with two distinct *response forms*, corresponding to the two prototypical classes of episodes. Each response item is designed to operationalize a specific appraisal variable, and the numerical indexing of the responses directly identifies the variable being assessed (e.g., rating response no. 5 operationalises controllability). Each appraisal variable is assessed using a Likert-type scale. Participants provide ratings on ordinal scales (e.g., six-point Likert scales), and the numerical values of these responses constitute the basis for quantitative aggregation. Mean values for each appraisal variable are computed by averaging the Likert-scale ratings across the nine episodes belonging to the same prototype (aversive vs. loss/failure) and episode phase.

As a result, the expected appraisal profiles reported in the SCPQ correspond to averaged Likert-scale values. These mean values are reported in Table 6, as documented in the SCPQ reference manual [17]. In the same table, the appraisal values computed by the EMA [19] model are reported, supporting a direct qualitative comparison between the appraisal variables computed by computational models and the normative appraisal profiles reported for healthy adults, since the resulting values do not represent individual reactions but aggregated trends associated with each prototype and its temporal evolution.

Consistently with this design, the present evaluation does not aim to reproduce exact numerical scores, but rather to assess whether the appraisal trajectories generated by the proposed model follow the expected qualitative trends associated with aversive and loss prototypes as defined by the SCPQ. Moreover, to further operationalize this comparison, a quantitative distance measure between appraisal trajectories was also computed. The methods used to compare the model are described in .

SCPQ trends

The SCPQ qualitative patterns specify the expected behavior of appraisal variables and emotional responses in canonical situations. Between the whole trends, the following ones are considered for comparing the models:

a. the aversive condition should generate higher controllability and changeability than the loss condition;
b. the appraisal of controllability and changeability should decrease over the three phases;
c. the negative valence should increase over the three phases;
d. the negative valence and the positive valence should be strongly different in correspondence to a bad/good outcome;
e. the aversive condition should lead to more anger and less sadness.

5.1.2. Methodology

Encoding the canonical episodes and appraisal values

The evaluation strategy follows the methodology originally proposed in [19], in which the temporal evolution of stressful episodes is modeled by systematically varying the likelihood of future outcomes, while abstracting SCPQ episodes through a canonical grammar. This grammar encodes the underlying prototypical structure of stressful situations and allows each episode to be classified into one of four canonical conditions: aversive-good, aversive-bad, loss-good, and loss-bad.

According to this abstraction, all episodes are characterized by the presence of a goal whose attainment becomes progressively more or less likely across discrete phases of the episode. Specifically, the likelihood of goal attainment decreases during the middle phase and reaches either zero or one in the final phase, depending on whether a bad or good outcome is realized. This temporal structure mirrors the SCPQ design, in which appraisal variables are expected to evolve systematically as a function of situation type and phase.

For the aversive condition, the probability that a future corrective action succeeds is set to 66% in phase one, reflecting a relatively high degree of controllability and changeability. This probability is reduced to 33% in phase two, modeling the effect of time passing without improvement, and is finally set to either 0% or 100% in phase three, corresponding to bad or good outcomes, respectively.

For the loss condition, the probability of loss occurrence is initially set to 50% in phase one, increases to 75% in phase two as the loss becomes more imminent, and is finally set to either 0% or 100% in phase three.

To expose the robot to equivalent stressful situations, these canonical scenarios are explicitly modeled within the robot's knowledge base. In particular, a general concept `stressful_event` and four

Table 8
The parameters and the appraisal variables outputted by the proposed model corresponding to the aversive and loss situations. The likelihood values are from the EMA narration over the phases. The final normalized values of the appraisal variables that are reported in the graphs for the comparative evaluation are highlighted.

Aversive				
	Start	Middle	Good	Bad
k	1.0	1.0	1.0	1.0
L_a	0.0	0.66	0.33	0.33
L_p	0.66	0.33	1.0	0.0
$x(L_a, L_p)$	0.66	0.33	0.67	0.33
$C(k, x)$	−1.07	−2.92	−1.04	−2.92
$N(C(x, k))$	1.60	0.065	1.63	0.065
$M(k, x)$	0.66	0.33	0.67	0.33
$N(M(k, x))$	1.1	0.55	1.11	0.55
D	−1.0	−1.0	1.0	−1.0
$v(C, M, D)$	−2.73	−4.25	−0.71	−4.25
$N(v(C, M, D))$	1.88	0.623	3.57	0.623
Loss				
	Start	Middle	Good	Bad
k	1.0	1.0	1.0	1.0
L_a	0.0	0.50	0.75	0.75
L_p	0.50	0.75	0.0	1.0
$x(L_a, L_p)$	0.50	0.25	0.75	0.25
$C(k, x)$	−1.75	−3.0	−0.77	−3.0
$N(C(k, x))$	1.04	0.0	1.85	0.0
$M(k, x)$	0.50	0.25	0.75	0.25
$N(M(k, x))$	0.83	0.41	1.25	0.41
D	−1.0	−1.0	1.0	−1.0
$v(C, M, D)$	−3.25	−4.25	−0.52	−4.25
$N(v(C, M, D))$	1.45	0.625	3.73	0.625

corresponding subclasses `id_n` ($n \in \{0, 1, 2, 3\}$) are introduced, each representing one of the four canonical situations (aversive good/bad, loss good/bad). Each subclass is associated, via the object property `P:likelihood`, with the phase-dependent likelihood values derived from the EMA narrative. These values are then appraised through the robot's inner speech mechanism, as described in Section 3.2. An example of how inner speech operates in the modeled canonical episodes is provided in Table 7.

Once appraised, these parameters are used as inputs to the mathematical formulations of the appraisal variables defined by the proposed model. The resulting appraisal values are subsequently normalized to enable comparison across phases and conditions. Table 8 summarizes the parameter settings and the corresponding computed appraisal variables for all canonical situations.

For the reported analysis, the entropy k was set to 1. This was done because the k value does not influence the global trends. In fact, as shown in Fig. 8, the appraisal variables maintain the same reciprocal trends even if k varies. That is, when k changes, the trends of the controllability and changeability remain the same with different x values, that is, in a same canonical episode. As a consequence, the comparative evaluations of these trends do not depend on the different k values.

Model comparison methodology

The comparative evaluation does not aim at reproducing the exact numerical values reported in the SCPQ, which are derived from aggregated Likert-scale responses of human participants. Rather, it primarily assesses whether the appraisal trajectories generated by the proposed model match the expected normative trends observed in healthy adults, such as relative differences between aversive and loss conditions and systematic changes across the phases of each episode. Given the normative and aggregate nature of the SCPQ data, classical inferential statistics are not applicable in this context. For this reason, a distance-based metric was adopted to quantify the deviation between model-generated appraisal profiles and the expected normative trends. Such a metric measures the similarity between the temporal dynamics of the two models, and is quantified using the Root Mean Squared Error (RMSE). Specifically, for each combination of appraisal variable and episode type, the temporal trend is defined as the difference between the appraisal value at the Middle phase and the corresponding value at the Start phase:

$$\Delta = V_{\text{Middle}} - V_{\text{Start}}, \tag{1}$$

where V_{Start} and V_{Middle} denote the appraisal values at the Start and Middle phases, respectively. This definition isolates the early temporal dynamics of the appraisal process independently of absolute scale differences across models.

Valence is assessed over three phases (Start, Middle, and Outcome), reflecting its sensitivity to the resolution of the episode. To ensure comparability with appraisal variables, the trend for valence is computed using the same Start–to–Middle definition:

Let $\boldsymbol{\Delta}^A = \{\Delta_1^A, \ldots, \Delta_N^A\}$ and $\boldsymbol{\Delta}^B = \{\Delta_1^B, \ldots, \Delta_N^B\}$ be the trend vectors associated with two models A and B, computed over N appraisal–episode combinations, including valence trends defined as above.

The root mean squared error (RMSE) between the two trajectories is computed as:

$$\text{RMSE}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta_i^A - \Delta_i^B\right)^2}. \tag{2}$$

Lower RMSE values indicate a higher similarity between the temporal trends of the compared models.

This metric provides a compact measure of the average deviation between the model-generated appraisal trajectory and the normative human appraisal profile across phases. RMSE is particularly suitable in this setting because it does not assume any specific distribution of the underlying data and allows comparison between deterministic model outputs and aggregated human data. Moreover, by operating on phase-dependent trajectories rather than isolated values, it directly reflects the process-oriented nature of appraisal dynamics, which is central to both the SCPQ framework and the evaluation strategy originally adopted for EMA.

5.1.3. Comparative results and discussions

The comparative evaluation of the proposed model is articulated along two complementary dimensions: a qualitative analysis of appraisal trends and a quantitative assessment of their deviation from normative human profiles. This dual perspective reflects the nature of the SCPQ, which characterizes typical human appraisal dynamics in terms of prototypical patterns rather than individual-level ground truth trajectories.

Qualitative assessment

From a qualitative standpoint, the analysis focuses on whether the proposed model reproduces the expected structural properties of appraisal dynamics observed in healthy adults across canonical stressful situations. In particular, the comparison examines relative differences between aversive and loss episodes, as well as systematic changes of appraisal variables across the temporal phases of each episode. This trend-based evaluation is consistent with the diagnostic use of the SCPQ, where validity is established by conformity to normative response profiles rather than exact numerical correspondence.

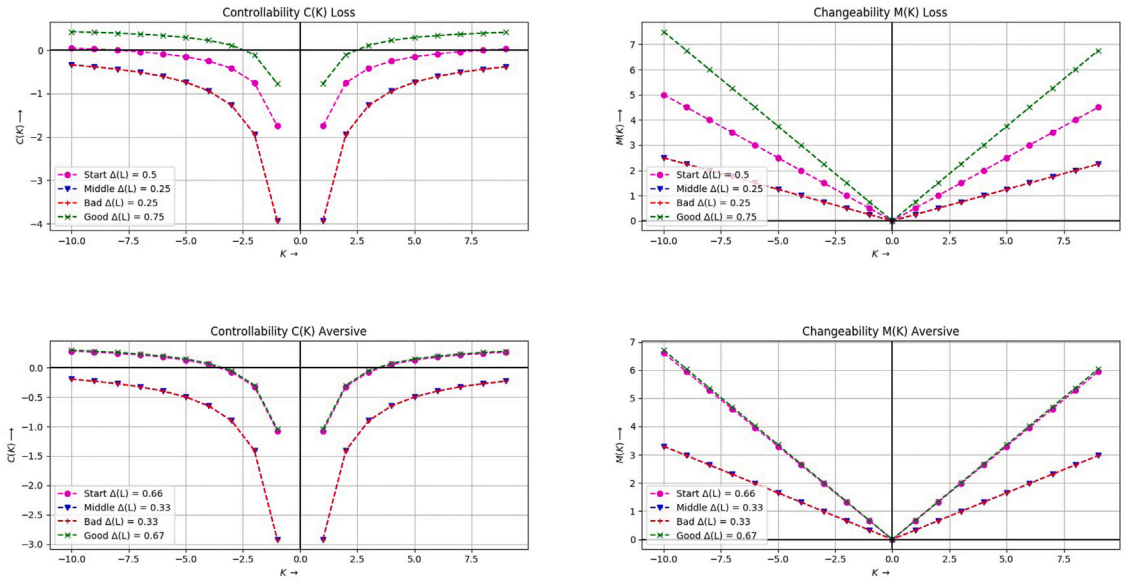


Fig. 8. The variation of controllability and changeability in correspondence of different entropy values. By fixing the likelihood, the controllability and the changeability vary maintaining the same trend. That is, in a same canonical episode, the appraisal variables follow the same trends under different environmental conditions.

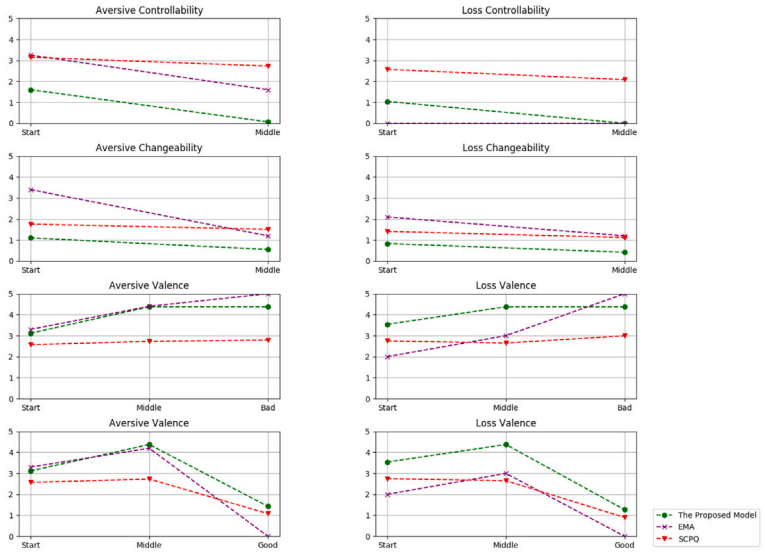


Fig. 9. The comparative evaluation of the appraisal variables with EMA and the SCPQ trends related to the four canonical stressful situations.

To facilitate qualitative comparison, Table 9 consolidates the EMA–SCPQ mean values reported in Table 6 with the outputs of the proposed model presented in Table 8. The same information is also shown in Fig. 9, which provides an immediate graphical overview, while Table 9 allows for a more detailed inspection of the numerical values.

From these observations, it emerges that the proposed model reproduces the main SCPQ trends. In aversive situations, controllability and changeability are consistently higher than in loss situations, reflecting a greater perceived potential for intervention. In both episode types, these variables decrease over time as the situation remains unresolved, capturing the progressive reduction of perceived agency. Notably, in loss scenarios, the proposed model exhibits a decreasing trend in controllability across phases, whereas EMA maintains a nearly constant profile, resulting in a closer alignment with the SCPQ expectations.

With respect to valence, the model captures both the increase in negative valence across phases and the marked divergence between good and bad outcomes in the final phase. This effect is particularly pronounced in loss episodes, in agreement with SCPQ findings. Moreover, Fig. 10 shows that in the good outcome, the positive emotions emerge over negative ones, which are predominant in each initial phase, because the simulated situations are stressful. Instead, when the situations end with a negative outcome, the predominant emotions remain negative.

Specifically, such a qualitative analysis could be synthesized by the following trend observations.

The trend *a.* is respected because in the aversive conditions, the controllability and changeability are higher than the loss conditions.

The trend *b.* is respected because the controllability and changeability decrease over the three phases. Moreover, the proposed model follows this trend in each situation, and in particular, for the loss situation, it presents a better trend than the EMA model, for which the controllability is constant in the loss situation.

The trend *c.* is respected because the negative valence grows over the phases, and the charts show this trend in the negative evolution

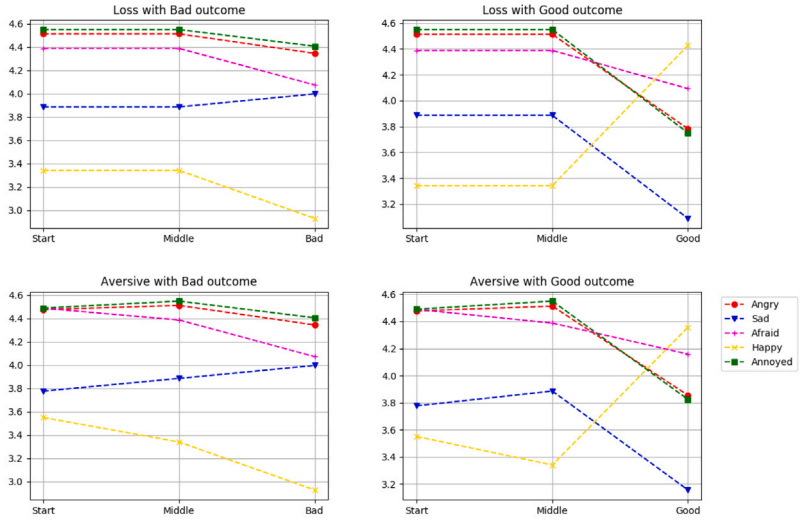


Fig. 10. The resulting emotions of the proposed model in the four SCPQ canonical episodes. The emotions are the expected ones according to the SCPQ trends.

of both aversive and loss conditions. Moreover, the trend follows more faithfully the SCPQ trend than EMA in the loss situation.

The trend *d.* is respected because the positive valence (good outcome) and the negative valence (bad outcome) have strongly different values. In this case, the trend of the proposed model follows more faithfully the SCPQ trend than EMA.

The trend *e.* is respected because, as shown in Fig. 10, the aversive condition leads to more anger and less sadness.

These qualitative observations establish that the proposed inner-speech-driven appraisal mechanism preserves the core normative structure of human appraisal dynamics. With the aim to provide a more fine-grained comparison, a quantitative evaluation of the deviation between model-generated trajectories and SCPQ normative profiles complements this analysis.

Quantitative assessment

Based on the values reported in Table 9, a quantitative comparison between the normative SCPQ profiles and the computational models can be performed by analyzing their early temporal dynamics.

Table 10 reports the temporal trends of each appraisal variable, defined as the difference between the Middle and Start phases ($\Delta = V_{\text{Middle}} - V_{\text{Start}}$), for both episode types. These trends make explicit the vectors used for the quantitative comparison and highlight systematic differences in the magnitude and direction of early appraisal changes across models. While SCPQ exhibits relatively small and stable variations, both EMA and the proposed computational model display larger temporal shifts, with EMA generally producing the most pronounced changes, particularly for controllability and changeability.

To summarize these differences in a single distance-based measure, Table 11 reports the Root Mean Squared Error (RMSE) values computed as described in , over the full set of appraisal–episode combinations, with $N = 6$, corresponding to the 2 episode types (Aversive, Loss) $\times$ 3 appraisal variables (Controllability, Changeability, Valence). As discussed, RMSE provides a compact estimate of the overall deviation between the temporal trend vectors generated by each model and the normative SCPQ trends.

The results show that the proposed computational model is more closely aligned with SCPQ than EMA, as indicated by the lower RMSE value. In contrast, the largest divergence is observed between SCPQ and EMA, reflecting substantial differences in their predicted temporal dynamics. The RMSE between EMA and the proposed model (PM) further confirms that the two approaches generate markedly distinct appraisal trajectories.

Table 9
Integration of the mean values of the appraisal variables as reported in Table 6 extended with the corresponding outputs of the Proposed Model (PM).

Aversive				
Controllability	*Start*	*Middle*		
EMA	3.25	1.60		
SCPQ	3.15	2.73		
PM	1.60	0.07		
Changeability	*Start*	*Middle*		
EMA	3.40	1.20		
SCPQ	1.76	1.51		
PM	1.10	0.55		
Valence	*Start*	*Middle*	*Good*	*Bad*
EMA	3.30	4.20	0.00	5.00
SCPQ	2.57	2.73	1.08	2.79
PM	1.88	0.62	3.57	0.62
Loss				
Controllability	*Start*	*Middle*		
EMA	0.00	0.00		
SCPQ	2.57	2.08		
PM	1.04	0.00		
Changeability	*Start*	*Middle*		
EMA	2.10	1.20		
SCPQ	1.41	1.12		
PM	0.83	0.41		
Valence	*Start*	*Middle*	*Good*	*Bad*
EMA	2.00	3.00	0.00	5.00
SCPQ	2.75	2.65	2.99	0.90
PM	1.45	0.63	3.73	0.63

Table 10
Temporal trends ($\Delta = V_{\text{Middle}} - V_{\text{Start}}$) for each appraisal variable, model, and episode type.

Episode	Appraisal	EMA Δ	SCPQ Δ	PM Δ
Aversive	Controllability	−1.65	−0.42	−1.53
	Changeability	−2.20	−0.25	−0.55
	Valence	+0.90	+0.16	−1.26
Loss	Controllability	0.00	−0.49	−1.04
	Changeability	−0.90	−0.29	−0.42
	Valence	+1.00	−0.10	−0.82

Table 11
RMSE values computed over the trend vectors of the compared models SCPQ, EMA and the Proposed Model (PM).

Comparison	RMSE
SCPQ vs EMA	1.13
SCPQ vs PM	0.83
EMA vs PM	1.42

A closer inspection of the trend values in Table 10 helps to interpret these differences. EMA tends to exhibit sharper and less differentiated temporal variations across variables, whereas the proposed model more consistently preserves the direction of the normative SCPQ trends while allowing for larger but structured changes in magnitude. In this sense, the proposed model can be seen as mediating between the strongly data-driven dynamics of EMA and the smoother, aggregate nature of the SCPQ profiles.

Overall, the quantitative analysis corroborates the qualitative comparison of temporal trends. While EMA shows strong deviations from the normative trajectories, the proposed model achieves a closer match to SCPQ when considering the full set of appraisal–episode combinations. These results support the claim that integrating inner speech and generative appraisal mechanisms yields temporally coherent and normatively grounded appraisal dynamics, while maintaining interpretability with respect to the cognitive processes underlying such evaluations.

5.2. Empirical validation

The empirical validation adopts an observational design in which human participants evaluate the robot's affective responses by watching predefined scenes. Although participants do not interact directly with the robot, this method is widely used in affective HRI to assess whether a robot's appraisal-driven behavior is interpretable and judged as coherent by human observers [20].

A related study [36] evaluated a system in which a robot expressed reactive emotions based on multimodal emotion recognition. Their user evaluation with approximately ten participants demonstrated that even relatively small samples can yield meaningful and reliable insights into how humans interpret a robot's affective behavior. For this reason, and consistent with prior work, our empirical validation relies on a comparable number of participants, who observed a series of scenes involving a human actor and the robot. In each scene, an emotionally relevant event occurred, and the robot's inner speech driven appraisal model processed the situation, generating the corresponding affective evaluation. Despite the limited sample size, the collected observations provide evidence that human participants were able to interpret the affective evaluations produced by the model, supporting the hypothesis that inner speech contributes to generating transparent and readable robot behavior.

5.2.1. Methods

A series of scenes, representing interactive acts between the robot and a human actor, were observed and assessed by humans. During each scene, an emotional relevant fact happened, and the robot's emotional model by inner speech appraised the situation leading to the robot's emotional state representation. The observers watched the scene and, at the end of the scene, completed two questionnaires, which are open-ended and multiple-choice questionnaires. Following the same hypothesis described in [20], an assumption was that in an open-ended questionnaire the observers would spontaneously attribute emotional states to the robot after observing an interaction between it and a human. The plain text description of the situation would assess the evidence of the robot's emotional behavior. Another assumption was that observers would attribute corresponding emotions to the robot if they could choose from a set of emotions. For this reason, a multiple-choice questionnaire was administered with the open-ended one, allowing to assess if the recognized emotion was correct and considered proper for the situation represented in the scene.

The experiment was carried out at the RoboticsLab of the University of Palermo, and involved both the Pepper robot and the Nao robot by Aldebaran. The participants attended a set of scenes representing interactive acts between the robot and a human actor, and assessed the robot's emotional behavior emerging from the interaction. The observers did not know that the robot is able to talk to itself. The inner voice had specific parameters (speed, double effect, volume) that differed from the standard voice's parameters of robots, so it could be spontaneously detected by the participants.

Preparation for the scenes

The stimuli consisted of a set of predefined interaction scenes presented in two different ways: as live demonstrations or as recorded videos. In both cases, the robot was running the full appraisal and inner-speech architecture in real time. No emotional responses were scripted or post-produced.

In the first type of interaction scenes, participants observed a live interaction between the Pepper robot and a human actor. The actor was the same across all scenes in order to reduce variability in participants' observations.

Fig. 11 illustrates the experimental framework adopted for the live interactions.

In this case, the represented situation corresponded to the use case described in Section 4, with the associated appraisal variables. The robot and the human actor collaborated in setting up a virtual lunch table, implemented through a MIT App Inventor[5] application running on an external tablet. The actor dragged and dropped utensils onto the virtual tablecloth. While the table-setting task represents a functionally constrained activity, the evaluation of appraisal mechanisms in such contexts remains methodologically relevant. In fact, this scenario was inspired by prior research [37], which demonstrated that robot's inner speech significantly influences human perceptions of trust and anthropomorphism even in low-stakes collaborative tasks. These prior findings suggest that transparent cognitive processes influence human–robot interaction dynamics independently of task complexity. Building on them, the present study extends the investigation to appraisal-driven inner speech, examining whether affective evaluation mechanisms maintain interpretability and coherence when transparently communicated in the same controlled setting, and establishes foundational mechanisms whose effects on trust, transparency, and relational quality have already been shown to matter even in low-stakes collaboration.

As shown in Fig. 11, the robot detected the events occurring on the tablet surface via a client–server bridge and appraised them using the proposed model, generating an affective evaluation output. An event occurred whenever the human actor placed an object: while objects were expected to follow an etiquette-based schema, the actor could violate these rules with varying degrees of error, as described in Section 4. Throughout the interaction, the robot overtly verbalized its inner speech while appraising the context.

In the second type of interaction scene, participants watched a video showing an interaction between the Nao robot and a human actor in a more general context. These scenes were recorded in the same laboratory and involved the following two situations:

- a noisy environment in which the actor asked the robot to perform an action (e.g., to sit down); due to the noise, Nao was unable to understand the command, leading to a negative computed affective state according to the model;

[5] http://appinventor.mit.edu/

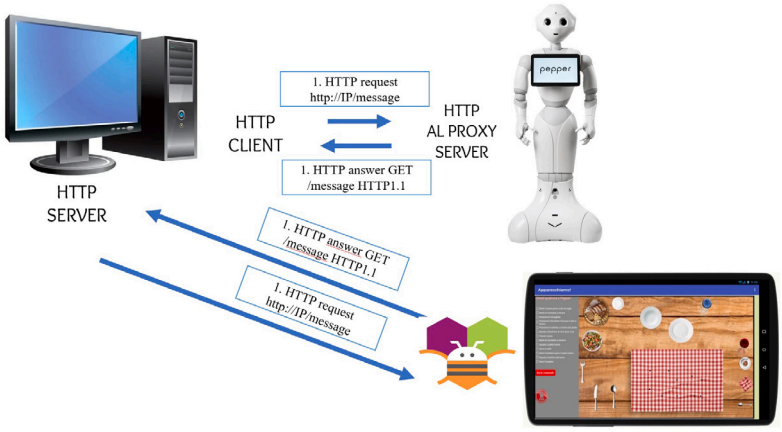


Fig. 11. The framework used in the live interactions. An actor prepared the virt the event detection by Pepper who appraised the action. An emotional state representation emerged from the proposed model.

- a situation in which the actor issued the same request using a harsh tone; in this case, Nao correctly understood the command, but the actor's tone contributed to a negative emotional state representation as predicted by the proposed model.

In both situations, Nao overtly verbalized its inner speech, allowing observers to hear the reasoning process used to appraise the context and leading to the negative affective state. The videos are available in the project's GitHub repository.[6]

Given that not all participants may have had prior direct experience with robots, the inclusion of both live and video-based interactions was intended to support emotion attribution independently of the robot's physical presence, thereby promoting a more general and observer-based evaluation of the robot's emotional behavior.

Participants

Table 12 summarizes the recruitment strategy. Participants were recruited among students at the University of Palermo. The average age was 20, and they were uniformly distributed for sex. With the purpose of making them blind about the study, they were contacted personally by visiting the courses of the university, and none of them registered to the test on his/her own.

The ability of the robot to talk to itself was not mentioned during the visits in order to have participants completely unaware of what would be happening during the test.

Participants present on the day of experiment were $N = 54$, and they were grouped into 5 small groups of 10/11 participants. A group of participants attended the scenes together. Some of participants did not complete the questionnaires (see below), and they were not included in the results. All participants declared to never have experiences with humanoid robots.

Trials and questionnaires

For each group of participants, the experimental session consisted of 7 trials. In each trial, each group attended the interactive scenes. In particular, in the first 5 trials, participants attended the live interaction. In the last 2 trails, participants watched the videos. The scenes were numbered according to the order of presentation.

During the viewing of the scenes, an experimenter was always present in order to confirm that participants complied with the rule and were not discussing the task. Moreover, the actor interacting directly with the robot was with the experimenter.

[6] https://github.com/RoboticsLab-Unipa/Robot-s-inner-speech-and-emotions

Table 12
Overview of participant recruitment and characteristics in the human-observer study.

Recruitment strategy	Direct personal contact during university courses; no self-registration.
Participants (N)	$N = 54$
Demographics	Mean age: 20 years; sex approximately uniformly distributed.
Rationale	Observer-based evaluation with participants naïve to humanoid robots, aimed at assessing the interpretability of the robot's emotional behavior.
Inclusion criteria	Adult participants (18+), with no prior experience with social or humanoid robots.

At the end of each trial, participants were asked to assess the scene using the open-ended and the multiple-choice questions, as previously described. The experimenter explained that they had to complete by plain text and free words, and after completing this part, they had to complete the forced question corresponding to the same scene. Participants had to evaluate the scenes one by one, sequentially. Before the first scene, participants received no information about the aim of the study or what they would see.

The instruction for the open-ended questions was: *In the following you will see a set of interactions between the robot and a human. Please, write down briefly what is happening in each scene, at the end of the scene...".* The question was followed by 5 blank rows per scene, for a total of 35 blank rows. Each block of 5 rows was numbered with the number of scene. In this part, the participants were free to write what they wanted.

Participants completed the multiple-choice question after completing the corresponding question in the open-ended one. They could choose from a list of the 5 basic emotions. In addition, participants were asked to choose a marker (+ or −) for evaluating the congruence of the emotion with the context. The instruction for this part was: *Now, choose from the following emotions the one which best describes what you see. Then, mark the + or the - sign near the list whether the emotion that emerged is congruous with the scene or not...* The list of emotions with the markers was repeated 7 times, for enabling the choice of the emotions and its congruence for each scene. Each repetition was marked with the number of the scene as in the open-ended questionnaire.

At the end of each scene, the experimenter gave time for completing both questionnaires, highlighting the number of the scene. Once all participants in the group indicated completion of the part of both questionnaires, the experimenter gave the start for the next scene.

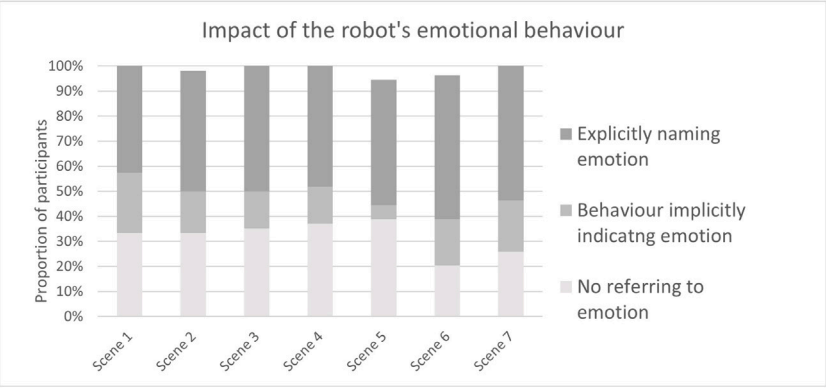


Fig. 12. The impact of the robot's emotional behavior on observers, as emerging from their answers to the open-ended questions. The distribution of the three scores shows in what proportion the participants assessed the scenes by giving a formal behavior description (score 1); or implicitly indicating an emotion (score 2); or explicitly naming an emotion (3). A very little percentage of participants did not provide an answer just for the scenes 2, 5, and 6, and the corresponding bars are reduced.

Data analysis method

The data analysis targeted three complementary outcome measures, following the established observational protocol proposed by [20]: (i) emotional attribution and salience, assessed through open-ended responses; (ii) emotion recognition accuracy, assessed via forced-choice selections; and (iii) perceived contextual appropriateness of the recognized emotion.

According to the mentioned protocol, answers for the open-ended questions were scored depending on how explicitly participants referred to the robot's emotional behavior. In particular, the score of the answer was:

1: if the answer was a formal description of the observed robot's behavior (i.e., it moves its head, raises its arm);
2: if the answer implicitly indicated some inner emotional state but without explicitly naming a concrete emotion (i.e., focuses the attention, tries to fix the situation)
3: it the answer referred explicitly to an emotion either as a verb, a noun or an adjective (i.e., happy, the robot is afraid).

With the aim to analyze the impact of the robot's emotional behavior, the score distribution over the participants was computed for each scene. Scores 2 and 3 denote a form of emotional impact on the participants, that is evident with score 3.

Data were analyzed descriptively by computing score distributions and proportions across participants for each scene. No inferential statistical tests were applied, in line with the exploratory and observational nature of the study.

Regarding the multiple-choice questions, the participants were rate, for each scene, according to the correctness of the recognized emotion. For the participants who provided a correct choice, the congruence of the recognized emotion with the scene was considered in the computation of the emotion's appropriateness.

Inter-rater reliability was not computed, as forced-choice responses did not require coding and open-ended answers were scored using a predefined scheme. Consistency was evaluated by examining convergence between qualitative descriptions and quantitative recognition patterns.

5.2.2. Discussions

Fig. 12 shows the impact of the robot's emotional behavior on human observers. Participants provided free-text descriptions of each scene, and only a portion of them ($M = 32\%$) did not refer to emotions. Half of the participants ($M = 50\%$) explicitly mentioned the robot's emotional state, and another group ($M = 16\%$) implicitly referred to emotions. These results indicate that hearing the robot's verbalized inner speech made its affective evaluation salient for observers: in 66% of the cases, participants referred to emotional content when describing the scene, suggesting that the robot's emerging emotional state was intelligible and perceptually accessible.

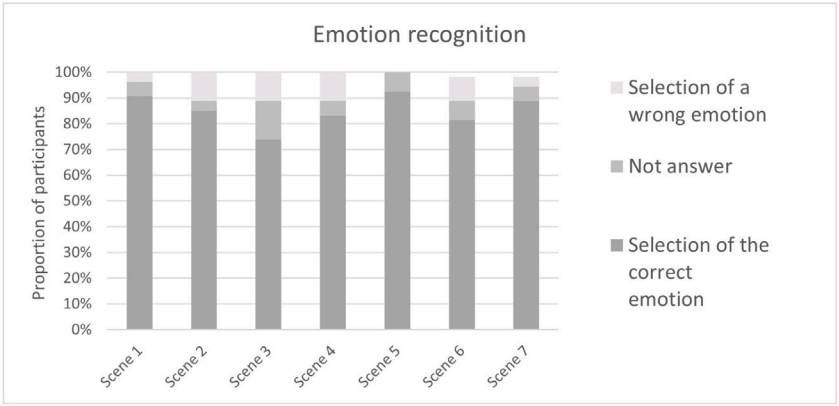
Moreover, the emotion produced by the proposed model was correctly recognized by a large majority of participants. As shown in Fig. 13(a), 80% of participants identified the intended emotion in each scene, with Scene 3 being the only exception, although still above 70%. Among those who selected the correct emotion, most considered it contextually appropriate ($M = 68\%$), as illustrated in Fig. 13(b). Only a small proportion of participants judged the emotion as incongruent ($M = 10\%$), while the remaining non-selections were primarily missing responses rather than explicit disagreement. This omission likely reflects a misunderstanding of the interface rather than a substantive judgment, as the same participants tended to omit this evaluation across several scenes. Overall, the data confirm that the robot's emotional responses were perceived as coherent with the situational context by the majority of observers.

Together, these quantitative observations show that the proposed model enables the robot to generate affective responses, through inner speech–driven appraisal, that are recognizable, interpretable, and contextually appropriate within real-world scenarios.

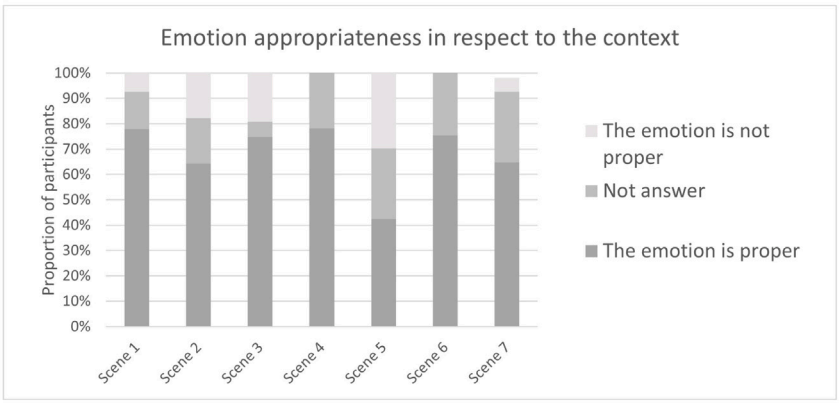
To complement the quantitative evaluation, we conducted a qualitative analysis of the open-ended descriptions provided by participants for each observed scene. The responses exhibited several systematic and theoretically relevant patterns. Participants frequently articulated the robot's internal state using affective descriptors that closely aligned with the emotional categories produced by the appraisal model, suggesting that the expressive cues were sufficiently transparent to support reliable emotional interpretation. Moreover, many observers explicitly grounded their descriptions in the situational dynamics depicted in the scene, indicating that they perceived a coherent mapping between contextual triggers and the robot's affective responses. A minority of participants reported instances in which the emotional expression appeared muted or insufficiently differentiated, thereby identifying potential avenues for refining the expressive bandwidth of the system. Taken together, these qualitative observations substantively reinforce the quantitative results and are consistent with prior evidence showing that human observers can meaningfully interpret appraisal-driven affective behaviors in artificial agents. They thus provide additional empirical support for the interpretability and contextual sensitivity of the proposed architecture.

6. Related works

The integration of robots' cognitive skills with the affective sphere is considered fundamental. In the last 30 years, there has been a significant increase in work that deepens the development of the emotional



(a) Emotion recognition



(b) Congruence of the recognized emotion

Fig. 13. The plots report the participants distribution for each scene in recognizing an emotion and its appropriateness for the scene by the multiple-choice questions. A high percentage of participants recognized the correct emotion in all scenes (subplot Fig. 13(a)), and among all correct answers only a small proportion considered this emotion not proper for the context: for more than half of the participants who provided the correct emotion, this emotion was considered congruent with the context (subplot Fig. 13(b)).

behavior in robots, highlighting the keen interest in this field [38]. Robots that rely on cognitive reasoning (planning, memory, task optimization) often fail in real-world settings because they ignore users' emotions, stress, and engagement, which are crucial for cooperation, trust, and learning [39].

Different research focuses emerge when studying the emotional component in robots. While some works focused on the development of computational models of emotions, to instill emotional behavior in the robot (that is elicited through the robot's gestures, phrases or music production [40–43]), other works focused on the social effects of such a behavior, to observe and test the sorted effects in human–robot interaction [44], in different social contexts, such as school [45] or hospital [46–48].

In a recent study [49], the cognitive and emotional processes were deeply linked and complementary, suggesting that the robot's affective behavior improves a human's confidence when collaborating with it. Human–robot collaboration can both increase role breadth self-efficacy and foster positive emotions, thereby improving service performance [50]. For example, a digital robot capable of autonomously expressing emotions was used to support teachers during lessons [51], demonstrating the benefits of an artifact with emotions relative to an artifact without them. Systematic reviews of social robots in pediatric care show consistent decreases in anxiety, distress, and stress, while their impact on pain itself is mixed or only small. [52,53].

Socially assistive and companion robots for older adults increasingly integrate emotion recognition, affective dialogue and cognitive support to address loneliness, depression and cognitive decline [54]. In [55], a pilot study examined a socially assistive robot for older adults with depression and dementia, comparing an empathic version equipped with multimodal emotion recognition and affective dialogue management to a non-empathic, scripted version. Results showed mood improvements in both conditions, with participants finding the empathic robot more engaging and likeable. This work highlights the potential value of artificial emotional intelligence in enhancing interaction quality in elder care settings.

Another example of a robotic system capable of using emotions during a communication with humans is presented in [56], where the robot becomes able to perceive and recognize the emotion of its interlocutor through audio sensors and video, to process this information and respond through actions that induce a positive emotional effect on the human being. For example, if the robot perceives a stressful situation in its partner, it responds by playing relaxing music.

A study on the empathy elicited by humanoid robots is discussed at [57] during a storytelling activity. In this case, the robot interprets the story's characters by enriching the narration with automatically generated emotional expressions that correspond to the dialogue. The system has been evaluated by comparing a simple narrative modality with an enhanced one, where an introspective dialogue is adopted to explain and make transparent the internal reasoning processes of the characters. The results show that storytelling activity significantly affected the cognitive component of empathy, particularly through the advanced narrative mode. Other studies, by contrast, focus on how

a robot can understand a human collaborator's emotion and elicit it through the actuators at its disposal.

The work in [58] presented a model that explicitly integrates a robot's inner speech with Damasio's theory of emotions and extended consciousness, resulting in SUSAN. By merging Damasio's framework with self-talk on a physical robot, the model allowed the robot to reason about and express its internal states. Experiments indicated that users perceived the robot as experiencing emotions, which increased empathy and emotional connection.

The architecture design that models emotion processing centrally was examined in [59], which investigated how large language models (LLMs) can improve the emotional intelligence of digital artificial agents. A new Chain-of-Emotion framework was introduced explicitly to model emotions through psychological evaluation. In three experiments, this method outperformed typical LLM architectures on user experience and content analysis metrics, providing preliminary support for the development of affective agents that utilize language-based cognitive mechanisms.

A different implementation was used in [60], where a Markov emotional model is proposed that accounts for transitions between emotional states while considering both the previously memorized emotion (internal to the robot) and the desired emotion of the robot's human interlocutor (external to the robot). The Markov emotional model was applied during interaction with the humanoid robot NAO to assess a human's personal affinity towards this type of machine.

AI and Machine Learning systems often prompt users to question how and why algorithms make specific decisions, particularly in sensitive domains such as medicine and healthcare. This has led to rapid growth in the field of Explainable Artificial Intelligence (XAI) in recent years. [61,62].

Often, expert systems do not provide additional information to support decisions, making them non-transparent, as if they were black boxes [63,64]. XAI aims to define rules that make a system transparent to users, providing explanations of its final decisions so that users can understand them without expert knowledge.

7. General discussion

While the experimental results presented in this work assess the coherence of the proposed model and its interpretability, a broader reflection on their implications is warranted. To fully frame the contribution of this research, it is essential to consider the theoretical and computational strengths of the architecture, the ethical dimensions associated with affective behavior in robots, and the current methodological limitations of the approach. The experimental scenarios adopted in the present study should be interpreted as controlled testbeds for validating the transparency and interpretability of appraisal-driven affective mechanisms, rather than as exhaustive demonstrations of their impact in socially complex interactions.

7.1. Strengths of the proposed approach

The proposed architecture presents several notable strengths that contribute to its relevance within the fields of affective computing and cognitive robotics. A first distinguishing feature is the innovative integration of inner speech with appraisal-based affective evaluation, which enables emotions to emerge from a transparent and linguistically grounded cognitive process rather than from predefined affective labels. This functional coupling between inner dialogue and appraisal aligns with established cognitive theories and provides a principled mechanism for directing attention, retrieving knowledge, and interpreting salient aspects of the situation. The system's reliance on an ontological representation, combined with spreading-activation dynamics, allows the robot to construct contextually rich and interpretable meaning structures, while the controlled recursive functioning of the rehearsal loop ensures a balance between cognitive plausibility and computational tractability. Furthermore, by verbalizing its internal inferences, the robot exposes the reasoning processes underpinning its affective evaluations, thereby enhancing transparency and explainability — key factors for trustworthy and collaborative human–robot interaction. Importantly, this verbalization also reduces the risk of unintentional deception, as it clarifies that the robot's emotional states stem from symbolic computations rather than subjective experience. The two experimental evaluations further demonstrate that the emotions generated through this mechanism are perceived by human observers as coherent and interpretable, suggesting that the proposed model can effectively support naturalistic and meaningful affective behavior in interactive scenarios.

7.2. Ethical considerations

The introduction of an appraisal-based affective mechanism raises important ethical considerations. Although the robot in the proposed model does not possess genuine phenomenological emotions, the production of verbalized inner speech and emotion-like evaluations may invite anthropomorphic interpretations. Prior work has shown that users tend to humanize technology when machines are augmented with additional anthropomorphic features [65], over-attribute mental states and emotional capabilities to artificial agents, which can affect trust calibration and expectations in interaction [66]. For this reason, it is crucial to clarify that the robot's affective states are functional representations derived from appraisal computations, not indicators of subjective experience.

Emotional expressiveness in robots should be carefully designed with respect to the task context, as inappropriate or unnecessary emotional displays may have unintended adverse effects on human–robot interaction [67]. It must satisfy not only believability but also social acceptability [68]: the risk is that existing computational models of robot emotion often prioritize expressive plausibility while neglecting mechanisms that determine whether an emotional display is contextually appropriate or should be inhibited, requiring some form of ethical reasoning to regulate the affective expressions. Within the proposed model, the inner speech mechanism contributes to this regulatory requirement by explicitly verbalizing and re-evaluating the situational context before an emotion is produced. While not a complete ethical reasoning module, inner speech provides a structured introspective layer that enhances transparency and could supports more socially acceptable affective responses.

Artificial systems capable of empathic behavior is highly contested, with many researchers arguing that such forms of empathy are either impossible or ethically problematic [69]: as a consequence, a more nuanced and empirically informed discussion is needed to assess both the risks and the potential benefits that such systems may offer for human well-being, ongoing study aimed at examining people attitudinal barrier and evaluating how serious a challenge it poses for the development and acceptance of empathic artificial systems.

A key challenge concerns human attitudes: people may be inclined to dismiss empathic responses from artificial agents simply because they originate from a non-human source. Other people know robots lack genuine emotions, but they still feel and attribute emotions to them because interactions activate scripts and schemata that prescribe how agents ought to act and feel [70]. Emotional attributions to robots are not mere cognitive errors but normatively guided responses [71], raising concerns about misplaced emotional commitments and assumptions of moral standing. Generally, the emotional deception and emotional attachment towards robot with affective behavior remained limited, although not entirely absent [72]. An affective robot is perceived more as a social entity, indicating a mild form of emotional deception, while emotional attachment ranged from low to moderate, although potential risks for vulnerable users can arise. Overall, while affective behaviors in robots can subtly influence users' perceptions, their ethical

impact depends strongly on individual differences, exposure time, and contextual factors, confirming the importance of further investigation.

In the proposed model, verbalized inner speech, by exposing the computational steps underlying the robot's affective evaluation, serves as an anti-deceptive mechanism. By explicitly revealing the symbolic reasoning that leads to the appraisal outcome, inner speech reduces the likelihood that observers interpret the robot's behavior as evidence of real computed affective state, thereby representing a way to mitigate the aforementioned ethical risks associated with artificial affectivity.

7.3. Limitations

Although the proposed architecture demonstrates coherent and interpretable affective behavior, some limitations must be acknowledged. First, the appraisal mechanism relies on a symbolic ontology whose coverage necessarily constrains the range and granularity of emotional evaluations that the robot can generate. While the two-level semantic expansion adopted in the rehearsal loop offers a balance between cognitive plausibility and computational efficiency, it also limits the depth of inference and may prevent the emergence of more nuanced or contextually subtle affective responses.

The experimental validation is restricted to scripted scenarios and observer-based evaluations. The study involved 57 participants, a number that aligns with, and often surpasses, sample sizes commonly adopted in affective HRI research [20,36]. This provides a solid basis for capturing reliable perceptual trends in how humans interpret a robot's affective evaluations. Nonetheless, a generalization is influenced not only by the number of participants but also by the structure of the experimental context. Participants evaluated the robot's behavior as observers rather than as active interlocutors, which may limit the generalizability of the findings to interactive contexts. Additional observations involving dynamic, bidirectional human–robot interaction are required to assess how inner speech-driven appraisal behaves in more naturalistic conditions. For example, some implementation challenges are orthogonal to the proposed work, and include robustness to noisy speech input, automatic speech recognition (ASR) errors, and large-vocabulary disambiguation. These issues are well known in the ASR and HRI communities and represent active areas of research. While they do not affect the appraisal formulation, they are critical for real-world deployment and will need to be addressed when integrating the architecture with state-of-the-art speech processing pipelines. However, the proposed framework is intentionally modular and compatible with any ASR system and domain-specific ontology, enabling straightforward incorporation of future advances in speech recognition and conceptualization.

The interpretation of the robot's emotions by human observers may be influenced by anthropomorphic biases. Although inner speech enhances transparency by revealing the symbolic reasoning that produces the affective state, the presence of verbalized self-reflection may still inadvertently reinforce the perception of agency or subjective experience. This raises important considerations regarding the boundary between functional affective models and user attributions, and suggests the need for further investigation into how design choices modulate emotional inferences.

Another limitation is that the table-setting task represents a low-risk and functionally constrained activity, which does not fully capture the affective and relational complexity of more socially demanding human–robot interactions. This choice was deliberate, as it enabled to isolate and validate the core appraisal and inner-speech mechanisms under controlled conditions, without confounding factors related to safety, escalation, or task failure. However, the perceptual dimensions assessed in the study (such as emotion attribution, behavioral plausibility, and contextual coherence) reflect core evaluative processes that are known to generalize across different task domains and participant groups.

Finally, the current implementation does not include an explicit regulation mechanism for modulating or suppressing emotional expressions when contextually inappropriate. While inner speech contributes a preliminary layer of meta-cognitive evaluation, more sophisticated regulatory components may be required to ensure social acceptability in complex interaction scenarios. Integrating ethical, normative, or task-oriented constraints into the appraisal cycle represents an important direction for future research.

7.4. Future work and research directions

The present work demonstrates that robot inner speech can effectively support context-sensitive affective evaluation by guiding attention, structuring conceptual interpretation, and providing transparent access to the robot's internal reasoning. While the results are promising, several avenues for further development emerge from this research.

A first direction concerns extending the emotional repertoire produced by the model. Although the current implementation focuses on a subset of basic emotions, the appraisal framework naturally scales to the full 28-emotion configuration of Russell's circumplex space. Incorporating this broader set would enable finer-grained affective distinctions and increase ecological validity in socially complex scenarios.

A second line of future work involves enriching the linguistic generation capabilities underlying inner speech. In the current prototype, inner speech is partly hand-crafted. Integrating advanced natural language generation systems or Large Language Models (LLMs) would enable the robot to verbalize richer, more contextually appropriate internal reasoning and to adaptively expand its expressive domain beyond predefined ontological concepts. This enhancement could improve the naturalness of the robot's explanations.

Future studies should also investigate the integration of multimodal perceptual cues (such as prosody, facial expression, or body posture) to complement the symbolic appraisal mechanism. Combining these cues with inner-speech-driven reasoning may produce more nuanced and robust affective evaluations.

From a social perspective, further empirical validation is needed to assess the broader impact of inner-speech-mediated affective behavior in realistic interactions. Future experiments will involve direct human–robot collaboration tasks, administered longitudinally, to measure changes in trust, transparency, user understanding, emotional alignment, and behavioral adaptation over time. Administering questionnaires before and after interaction sessions, as originally planned, may help quantify how inner speech influences user perception and social cognition.

Additionally, individual differences in users' sensitivity to emotional cues, anthropomorphic tendencies, or attachment styles should be explored. Tailoring the robot's expressive behavior and transparency level to user profiles could mitigate ethical risks such as over-trust, emotional dependence, or inadvertent deception. Finally, the scalability of the conceptualization and appraisal pipeline should be examined. As robots operate in increasingly open and dynamic environments, the ontology and reasoning mechanisms must adapt accordingly. Hybrid symbolic-subsymbolic architectures or LLM-assisted conceptualization may help achieve this flexibility while preserving the interpretability that inner speech provides. Moreover, such LLM integration could extend this architecture to more affectively demanding contexts. Recent evidence [73] demonstrates the potential of this extension: an LLM-powered robot facilitated cognitive reappraisal across multiple sessions with participants, yielding significant improvements in emotion self-regulation and expressiveness. Integrating explicit emotion regulation mechanisms would enable the proposed architecture to dynamically adapt robot behavior in response to users' affective states during more socially complex interactions, thereby meaningfully shaping interaction dynamics beyond low-stakes collaborative tasks. While such an approach relies entirely on LLM-driven generation, the proposed architecture provides a complementary contribution by grounding appraisal

processes in structured symbolic reasoning that maintains interpretability through inner speech. The integration of these approaches (combining the flexibility of LLMs with the transparency of symbolic appraisal) represents a promising direction for future work.

The present contribution can be situated within a broader research programme by referencing a recently published study [74], which investigates inner-speech mechanisms in higher-stakes interaction contexts, specifically surgical table preparation. While that work focuses solely on verbalized thinking, it demonstrates the applicability of inner speech architectures in safety-critical domains. This broader scope suggests that the appraisal and reappraisal framework validated here could be extended not only to affectively complex interactions but also to high-stakes operational settings where cognitive transparency and adaptive reasoning are paramount. Collectively, these research directions aim to refine the theoretical grounding, increase the expressive and perceptual capabilities of the system, and ensure that inner-speech-driven affective architectures evolve into transparent, contextually appropriate, and socially responsible robotic technologies.

8. Conclusions

This work introduced a computational model that integrates robot inner speech with appraisal-based affective evaluation, demonstrating how a self-generated verbal reasoning process can guide attention, structure contextual interpretation, and support the emergence of transparent and contextually grounded emotional states. By formalizing the appraisal variables and embedding them within a recursive inner-speech rehearsal loop, the model provides a cognitively inspired mechanism through which a robot can interpret emotionally relevant situations and verbalize the evaluative steps leading to its affective responses.

The implementation on a humanoid robot showed that the proposed approach is viable in real scenarios. The two complementary experimental evaluations demonstrated that the resulting emotional behavior is both coherent with appraisal theory and interpretable by human observers. Participants were generally able to recognize and understand the robot's emotional states, and their qualitative descriptions aligned with the internal reasoning expressed through inner speech. These findings confirm that coupling appraisal mechanisms with verbalized internal reasoning enhances the transparency and readability of the robot's affective processes, strengthening the foundation for trustworthy interaction.

Overall, the model contributes to the field of affective robotics by offering a novel integration of cognitive transparency and emotional evaluation, positioning inner speech as a functional bridge between perception, appraisal, and communication. The results highlight the potential of this approach to support more intelligible and socially aligned robot behavior, marking a significant step towards affective robotic systems capable of explaining not only what they feel, but why.

CRediT authorship contribution statement

Arianna Pipitone: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sophia Corvaia:** Validation, Software, Resources, Investigation, Data curation. **Antonio Chella:** Writing – review & editing, Validation, Supervision, Investigation, Funding acquisition.

Resources availability

The code produced for implementing the model, the necessary resources for running it and the presented results are available at the GitHub repository at https://github.com/RoboticsLab-Unipa/Robot-s-inner-speech-and-emotions.

Ethical approval

The study was approved by tthe Ethics Committee of the University of Palermo.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Special acknowledgments to Prof. Alain Morin[7] for his suggestions and improvements to the paper.

The material was supported by the co-funding of the European Union - ERDF or ESF, PON Research and Innovation 2014–2020 - Ministerial Decree 1062/2021.

Moreover, it is supported by the Air Force Office of Scientific Research, United States under award number FA9550-19-1-7025

Data availability

Data will be made available on request.

References

[1] D. Gregory, Inner speech: New voices, Analysis 80 (1) (2020) 164–173, http://dx.doi.org/10.1093/analys/anz096, arXiv:https://academic.oup.com/analysis/article-pdf/{80}/{{1}}/{164}/{32219089}/anz096.pdf.

[2] P. Langland-Hassan, A. Vicente, Inner Speech: New Voices, Oxford University Press, USA, 2018.

[3] M. Perrone-Bertolotti, L. Rapin, J.-P. Lachaux, M. Baciu, H. Lœvenbruck, What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring, Behav. Brain Res. 261 (2014) 220–239, http://dx.doi.org/10.1016/j.bbr.2013.12.034, https://www.sciencedirect.com/science/article/pii/S0166432813007833.

[4] A. Morin, Inner Speech, Oxford Scholarship Online, 2018.

[5] L. Vygotsky, Théorie des émotions: étude historico-psychologique, théorie des émotions, 1998, pp. 1–416.

[6] P. Fossa, R.M. Pérez, C.M. Marcotti, The relationship between the inner speech and emotions: Revisiting the study of passions in psychology, Hum. Arenas 3 (2) (2020) 229–246.

[7] R.S. Lazarus, Emotion and Adaptation, Oxford University Press, 1991.

[8] R.S. Lazarus, Cognition and motivation in emotion., Am. Psychol. 46 (4) (1991) 352.

[9] S.C. Marsella, J. Gratch, Ema: A process model of appraisal dynamics, Cogn. Syst. Res. 10 (1) (2009) 70–90, http://dx.doi.org/10.1016/j.cogsys.2008.03.005, modeling the Cognitive Antecedents and Consequences of Emotion. https://www.sciencedirect.com/science/article/pii/S1389041708000314.

[10] A. Morin, Inner speech, 2009.

[11] B. Alderson-Day, C. Fernyhough, Inner speech: development, cognitive functions, phenomenology, and neurobiology., Psychol. Bull. 141 (5) (2015) 931.

[12] A. Pipitone, A. Chella, What robots want? hearing the inner voice of a robot, Iscience 24 (4) (2021) 102371.

[13] A. Chella, A. Pipitone, A cognitive architecture for inner speech, Cogn. Syst. Res. 59 (2020) 287–292.

[14] A. Chella, A. Pipitone, A. Morin, F. Racy, Developing self-awareness in robots via inner speech, Front. Robot. AI 7 (2020) 16.

[15] A. Geraci, A. D'Amico, A. Pipitone, V. Seidita, A. Chella, Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions, Front. Robot. AI 8 (2021) 66.

[16] J.J. Gross, Handbook of Emotion Regulation, second ed., Guilford Press, New York, NY, US, 2014.

[17] M. Perrez, M. Reicherts, Stress, coping, and health: A situation-behavior approach: Theory, methods, applications, 1992.

[18] J.A. Russell, A circumplex model of affect., J. Pers. Soc. Psychol. 39 (6) (1980) 1161.

[7] http://morin.socialpsychology.org/

[19] J. Gratch, S.C. Marsella, Evaluating the modeling and use of emotion in virtual humans, in: International Conference on Autonomous Agents and Multiagent Systems, AAMAS, New York, NY, 2004.

[20] M. Gácsi, A. Kis, T. Faragó, M. Janiak, R. Muszyński, Ádám. Miklósi, Humans attribute emotions to a robot that shows simple behavioural patterns borrowed from dog behaviour, Comput. Hum. Behav. 59 (2016) 411–419, http://dx.doi.org/10.1016/j.chb.2016.02.043, https://www.sciencedirect.com/science/article/pii/S0747563216300954.

[21] R.S. Lazarus, Psychological stress and the coping process, 1966.

[22] G.L. Clore, A. Ortony, Appraisal theories: How cognition shapes affect into emotion, 2008.

[23] N.H. Frijda, The Laws of Emotion, Psychology Press, 2017.

[24] D. Irons, Prof. james' theory of emotion, Mind 3 (9) (1894) 77–97.

[25] K.R. Scherer, Appraisal Theory, Wiley-Blackwell, 2005, pp. 637–663, http://dx.doi.org/10.1002/0470013494.Ch30, (Chapter 30). arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013494.ch30. https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch30.

[26] K.R. Scherer, A. Schorr, T. Johnstone, Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, 2001.

[27] A. Morin, B. Hamper, Self-reflection and the inner voice: activation of the left inferior frontal gyrus during perceptual and conceptual self-referential thinking, Open Neuroimaging J. 6 (2012) 78.

[28] P. Ekman, Basic emotions, Handbook of cognition and emotion, vol. 98, (45–60) 1999, p. 16.

[29] L. Steels, et al., Language re-entrance and the 'inner voice', J. Conscious. Stud. 10 (4–5) (2003) 173–185.

[30] A.M. Collins, E.F. Loftus, A spreading-activation theory of semantic processing., Psychol Rev 82 (6) (1975) 407.

[31] D.A. Balota, R.F. Lorch, Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision., J. Exp. Psychol. [Learn. Mem. Cogn.] 12 (3) (1986) 336.

[32] A.S. Rotaru, G. Vigliocco, S.L. Frank, Modeling the structure and dynamics of semantic processing, Cogn. Sci. 42 (8) (2018) 2890–2917.

[33] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of tampa, Florida, J. Am. Stat. Assoc. 84 (406) (1989) 414–420.

[34] G. Paltoglou, M. Thelwall, Seeing stars of valence and arousal in blog posts, IEEE Trans. Affect. Comput. 4 (1) (2013) 116–123, http://dx.doi.org/10.1109/T-AFFC.2012.36.

[35] V. Mehra, G. Laban, H. Gunes, How large language models classify and semantically explain facial expressions from valence-arousal values, in: Proceedings of the 7th ACM Conference on Conversational User Interfaces, CUI '25, Association for Computing Machinery, New York, NY, USA, 2025, http://dx.doi.org/10.1145/3719160.3737618.

[36] Y. Li, C. Ishi, K. Inoue, S. Nakamura, T. Kawahara, Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction, Adv. Robot. 33 (2019) 1–12, http://dx.doi.org/10.1080/01691864.2019.1667872.

[37] A. Pipitone, A. Geraci, A. D'Amico, V. Seidita, A. Chella, Robot's inner speech effects on human trust and anthropomorphism, Int. J. Soc. Robot. 16 (6) (2024) 1333–1345.

[38] R. Savery, G. Weinberg, A survey of robotics and emotion: Classifications and models of emotional interaction, in: 2020 29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN, 2020, pp. 986–993, http://dx.doi.org/10.1109/RO-MAN47096.2020.9223536.

[39] R. Gervasi, F. Barravecchia, L. Mastrogiacomo, F. Franceschini, Applications of affective computing in human–robot interaction: State-of-art and challenges for manufacturing, Proc. Inst. Mech. Eng. Part B: J. Eng. Manuf. 237 (2022) 815–832, http://dx.doi.org/10.1177/09544054221121888.

[40] L. Li, Z. Zhao, Designing behaviors of robots based on the artificial emotion expression method in human–robot interactions, Machines (2023) http://dx.doi.org/10.3390/machines11050533.

[41] S. Jo, S. Hong, The development of human–robot interaction design for optimal emotional expression in social robots used by older people: Design of robot facial expressions and gestures, IEEE Access 13 (2025) 21367–21381, http://dx.doi.org/10.1109/access.2025.3534845.

[42] X. Liu, J. Dong, M. Jeon, Robots' woohoo and argh can enhance users' emotional and social perceptions: An exploratory study on non-lexical vocalizations and non-linguistic sounds, ACM Trans. Human-Robot Interact. 12 (2023) 1–20, http://dx.doi.org/10.1145/3626185.

[43] N.T.V. Tuyen, A. Elibol, N. Chong, Learning bodily expression of emotion for social robots through human interaction, IEEE Trans. Cogn. Dev. Syst. 13 (2020) 16–30, http://dx.doi.org/10.1109/tcds.2020.3005907.

[44] M.R. Fraune, B.C. Oistad, C.E. Sembroski, K.A. Gates, M.M. Krupp, S. Šabanović, Effects of robot–human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact, Comput. Hum. Behav. 105 (2020) 106220, http://dx.doi.org/10.1016/j.chb.2019.106220.

[45] A. Kouroupa, K. Laws, K. Irvine, S. Mengoni, A. Baird, S. Sharma, The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis, PLoS ONE 17 (2022) http://dx.doi.org/10.1371/journal.pone.0269800.

[46] N.L. Robinson, T. Cottier, D. Kavanagh, Psychosocial health interventions by social robots: Systematic review of randomized controlled trials, J. Med. Internet Res. 21 (2019) http://dx.doi.org/10.2196/13203.

[47] D. Logan, C. Breazeal, M. Goodwin, S. Jeong, B. O'Connell, D. Smith-Freedman, J.A.J. Heathers, P. Weinstock, Social robots for hospitalized children, Pediatrics 144 (2019) http://dx.doi.org/10.1542/peds.2018-1511.

[48] A.K. Triantafyllidis, A. Alexiadis, K. Votis, D. Tzovaras, Social robot interventions for child healthcare: A systematic review of the literature, Comput. Methods Programs Biomed. Updat. (2023) http://dx.doi.org/10.1016/j.cmpbup.2023.100108.

[49] L. Tan, C. Liu, Y. Wang, Y. Li, J. Zhao, S. Wang, B. Zhong, The effect of human–robot collaboration on frontline employees' service performance: A resource perspective, Int. J. Hosp. Manag. (2025) http://dx.doi.org/10.1016/j.ijhm.2024.104033.

[50] P. Weis, C. Herbert, Do i still like myself? human–robot collaboration entails emotional consequences, Comput. Hum. Behav. 127 (2021) 107060, http://dx.doi.org/10.31219/osf.io/6cmxd.

[51] F. Jimenez, T. Yoshikawa, T. Furuhashi, M. Kanoh, An emotional expression model for educational-support robots, J. Artif. Intell. Soft Comput. Res. 5 (2015) 51–57, http://dx.doi.org/10.1515/jaiscr-2015-0018.

[52] B. Littler, T. Alessa, P. Dimitri, C. Smith, L.D. de Witte, Reducing negative emotions in children using social robots: systematic review, Arch. Dis. Child. 106 (2021) 1095–1101, http://dx.doi.org/10.1136/archdischild-2020-320721.

[53] M.J. Trost, A.R. Ford, L. Kysh, J. Gold, M. Matarić, Socially assistive robots for helping pediatric distress and pain: A review of current evidence and recommendations for future research and practice, Clin. J. Pain 35 (2019) 451–458, http://dx.doi.org/10.1097/ajp.0000000000000688.

[54] J.A.R. Arango, C. Marco-Detchart, V.J.J. Inglada, Personalized cognitive support via social robots, Sensors (Basel, Switzerland) 25 (2025) http://dx.doi.org/10.3390/s25030888.

[55] H. Abdollahi, M. Mahoor, R. Zandie, J. Siewierski, S. Qualls, Artificial emotional intelligence in socially assistive robots for older adults: A pilot study, IEEE Trans. Affect. Comput. 14 (2022) 2020–2032, http://dx.doi.org/10.1109/taffc.2022.3143803.

[56] J.H. Lui, H. Samani, K.-Y. Tien, An affective mood booster robot based on emotional processing unit, in: 2017 International Automatic Control Conference, CACS, 2017, pp. 1–6, http://dx.doi.org/10.1109/CACS.2017.8284239.

[57] A. Augello, Unveiling the reasoning processes of robots through introspective dialogues in a storytelling system: A study on the elicited empathy, Cogn. Syst. Res. (2022).

[58] S. Corvaia, A. Pipitone, A. Chella, Inner speech and damasio's theory for modelling robot's emotions, IEEE Trans. Affect. Comput. (2025) 1–14, http://dx.doi.org/10.1109/TAFFC.2025.3547756.

[59] M. Croissant, M. Frister, G. Schofield, C. McCall, An appraisal-based chain-of-emotion architecture for affective language model game agents, PLOS ONE 19 (2023) http://dx.doi.org/10.1371/journal.pone.0301033.

[60] Y. Maeda, Human–robot interaction experiment based on markovian emotional model, in: 2018 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, 2018, pp. 1–6, http://dx.doi.org/10.1109/FUZZ-IEEE.2018.8491469.

[61] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, 2020, ArXiv abs/2006.00093.

[62] M.R. Islam, M.U. Ahmed, S. Barua, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, Appl. Sci. 12 (3) (2022) http://dx.doi.org/10.3390/app12031353, https://www.mdpi.com/2076-3417/12/3/1353.

[63] A. Rai, Explainable ai: from black box to glass box, J. Acad. Mark. Sci. 48 (12) (2019) http://dx.doi.org/10.1007/s11747-019-00710-5.

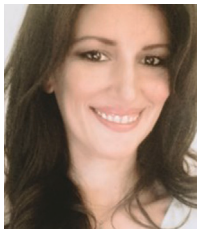
[64] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (02) (2018) http://dx.doi.org/10.1145/3236009.

[65] A. Waytz, J. Heafner, N. Epley, The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle, J. Exp. Soc. Psychol. 52 (2014) 113–117.

[66] P.A. Hancock, D.R. Billings, K.E. Schaefer, J.Y. Chen, E.J. De Visser, R. Parasuraman, A meta-analysis of factors affecting trust in human–robot interaction, Hum. Factors 53 (5) (2011) 517–527.

[67] D. Becker, D. Rueda, F. Beese, B.S.G. Torres, M. Lafdili, K. Ahrens, D. Fu, E. Strahl, T. Weber, S. Wermter, The emotional dilemma: influence of a human-like robot on trust and cooperation, in: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN, IEEE, 2023, pp. 1689–1696.

[68] S. Ojha, M.-A. Williams, B. Johnston, The essence of ethical reasoning in robot-emotion processing, Int. J. Soc. Robot. 10 (2) (2018) 211–223.
[69] J. Stenske, A. Tagesson, The prospects of artificial empathy: A question of attitude? Frontiers Artificial Intelligence Appl. 397 (2025) 149–156.
[70] N. Spatola, O.A. Wudarczyk, Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism, Comput. Hum. Behav. 124 (2021) 106934, http://dx.doi.org/10.1016/j.chb.2021.106934, https://www.sciencedirect.com/science/article/pii/S0747563221002570.
[71] L. Berio, Normativity and scripts in human robot interaction, in: Social Robots with AI: Prospects, Risks, and Responsible Methods, IOS Press, 2025, pp. 139–148.
[72] A. Van Maris, N. Zook, P. Caleb-Solly, M. Studley, A. Winfield, S. Dogramadzi, Designing ethical social robots—a longitudinal field study with older adults, Front. Robot. AI 7 (2020) 1.
[73] G. Laban, J. Wang, H. Gunes, A robot-led intervention for emotion regulation: From expression to reappraisal, 2025, arXiv preprint arXiv:2503.18243.
[74] A. Pipitone, G. Cataldo, S. Corvaia, A. Chella, The robot and the nurse prepare for surgery: early insights into the impact of robot's inner speech, Intell. Serv. Robot. 19 (1) (2026) 5.



Arianna Pipitone is a Research Fellow at the University of Palermo and the National Research Council (CNR) in Italy. Her research interests encompass Computational Linguistics, Natural Language Processing (NLP), Cognitive Robotics, Social Robotics, Artificial Intelligence, Artificial Consciousness and Affective Computing. She investigates the role of self-dialogue in trustworthy human–robot interactions, focusing on inner speech in robot selfconsciousness. She has authored and co-authored over 30 publications in Robotics and AI and has been featured in numerous national and international newspapers for her work on robots' inner speech.

Sophia Corvaia is a Ph.D. Student at RoboticsLab laboratory at the University of Palermo. Her studies are focused on Artificial Intelligence, Cognitive Robotics, Emotional Intelligence and Artificial Consciousness. Currently, her research topic focuses on the design and implementation of a cognitive model that can give the robot the ability to simulate emotions.

Antonio Chella is a Professor of Robotics at the University of Palermo, head of the Robotics Laboratory and director of the Computer Engineering Section of the Department of Engineering. He has been director of the Department of Computer Engineering and of the Interdepartmental Center for Knowledge Engineering. He is an Honorary Professor in Machine Learning and Optimization at the School of Computer Science, University of Manchester, UK. He is a member of the Italian National Academy of Sciences, Letters and Arts in Palermo, Italy. He received the James S. Albus Medal of the BICA Scientific Society. He coordinated research projects from Italy, Europe, and the United States. He is author or co-author of more than 200 publications in the fields of Robotics and AI. His current research topic concerns the interdisciplinary field of AI and machine consciousness. Prof. Chella's scientific research activities have been the subject of articles and interviews that have appeared in national and international journals and newspapers.