UNIVERSITÀ DEGLI STUDI DI PALERMO
DIPARTIMENTO DI INGEGNERIA

DOCTORAL COURSE IN
INFORMATION AND COMMUNICATION TECHNOLOGIES

ROBOTS' INNER SPEECH MEETS THEORIES OF EMOTION: TOWARDS AFFECTIVE ROBOTICS IN HUMAN ROBOT INTERACTION

Ph.D. CANDIDATE
Ing. SOPHIA CORVAIA

COORDINATOR
Prof. MARCO LA CASCIA

SUPERVISOR
Prof. ANTONIO CHELLA

CO-SUPERVISOR
Ing. ARIANNA PIPITONE

XXXVII CYCLE – ACADEMIC YEAR 2025-2026

Abstract

Robotics is undergoing a paradigm shift, evolving from systems primarily designed to execute mechanical or repetitive tasks into autonomous agents capable of engaging in meaningful, socially coherent interactions with humans. As robots become increasingly integrated into everyday environments, ranging from domestic assistance to healthcare, education, and therapeutic contexts, the demand for machines capable of understanding, expressing, and regulating emotions has grown considerably. Emotional intelligence is no longer perceived as an optional layer for enhancing user experience; it is increasingly recognised as a core requirement for establishing trust, fostering long-term interaction, and enabling cooperation between humans and artificial agents. Nevertheless, implementing this capability computationally is a significant challenge that demands the integration of affective modeling, cognitive architectures, and communication systems.

Emotions in biological organisms serve as a crucial mediator between cognitive processes and environmental demands. They influence decision-making, guide attention, facilitate memory consolidation, and play a key role in social bonding. From an evolutionary perspective, emotions act as adaptive responses that help organisms prioritise goals, evaluate risks, and coordinate behaviour within complex social structures. Translating these mechanisms into artificial systems involves not only developing mathematical models of affective states but also integrating them into architectures that enable flexible, context-sensitive behaviour. Many existing approaches in affective computing remain primarily reactive, mapping external stimuli to predefined emotional outputs, often yielding limited or shallow expressions that lack introspective depth.

This doctoral research aims to address these limitations by introducing a novel paradigm that integrates computational emotional models with inner speech mechanisms. Inner speech, also known as self-talk or covert verbal thinking, is a well-documented phenomenon in human cognition that contributes to self-regulation, planning, problem-solving, and emotional adjustment. In humans, inner speech enables the simulation of social dialogue within the self, allowing individuals to reason about

their feelings, anticipate consequences, and regulate their reactions in real time. The central hypothesis of this dissertation is that endowing robots with an analogous internal dialogue mechanism can support a form of machine introspection that enhances both emotional coherence and transparency in interaction. This introspection lays the foundation for robot consciousness, explored in detail in Chapter 2.

The research presented in this dissertation demonstrates that the integration of inner speech into computational emotional models provides a viable pathway toward more reflective and communicative robotic systems. Rather than focusing solely on emotional expression, the proposed approach enables robots to reflect upon, regulate, and communicate their internal states in a manner that is intelligible and socially appropriate for human partners. The findings suggest that such models can enhance the quality of human–robot interaction across multiple domains, while contributing to user trust and system acceptability. Future research will explore the scalability of this approach to more complex emotional repertoires, its integration with adaptive learning mechanisms, and its ethical implications, particularly regarding transparency, accountability, and the anthropomorphisation of artificial agents.

The dissertation is organised to reflect both its theoretical foundations and its applied contributions. The initial chapters review the state of the art in affective computing, emotion modelling, and inner speech, and outline the methodological choices underlying the proposed computational models. The central chapters detail the system architecture, its integration into robotic platforms, and the results of controlled experimental evaluations. The final chapters focus on real-world applications, describing interaction design, evaluation criteria, and insights gained from deployment in practical settings.

The first part of the research focuses on the design and formalisation of computational models capable of generating and modulating emotional states through mechanisms of inner speech. These models adopt a hybrid architecture that combines symbolic reasoning with dynamic affective processes, enabling a bidirectional flow between cognitive appraisals and emotional expressions.

In Chapter 1, the thesis introduces the conceptual and computational integration of inner speech within an appraisal-based emotional model for robots. Building on appraisal theories, which describe emotions as emerging from cognitive evaluations of situational variables, this chapter examines how self-directed dialogue can serve as an internal mechanism that supports emotional computation. Inner speech is modelled as a structured process of self-reflection that enables the robot to focus on contextually relevant information, compute assessment variables, and generate emotionally coherent responses. Particular attention is devoted to analysing emotional dynamics under stressful conditions, demonstrating that the proposed model produces patterns

consistent with those observed in healthy adults. Furthermore, the chapter discusses how the externalisation of inner reasoning through think-aloud behaviour can improve transparency and coordination in collaborative tasks, supporting clearer joint decision-making and mutual understanding between human and robot partners.

In Chapter 2, the research extends the investigation from appraisal-based evaluation to a broader affective framework grounded in Damasio's theory, in which emotions arise from the dynamic interaction between bodily-like signals and cognitive processes. This chapter explores how a robot can move beyond mere detection or simulation of emotions toward a computational architecture that supports emotionally grounded responses mediated by inner speech. Self-directed dialogue is implemented as a mechanism that enables the robot to articulate its interpretation of contextual events and its internal state, thereby fostering the emergence of more coherent and interpretable behaviours. The model is deployed on a real robotic platform, and experimental findings demonstrate that human participants interacting with the system can perceive the robot's emotional states. The results highlight how integrating embodied-cognitive mechanisms with inner speech enhances perceived empathy, trust, and engagement in Human–Robot Interaction.

Building on this foundation, the second part of the dissertation shifts focus to a preliminary investigation of the application of the developed computational models in a medical context. In Chapter 3, the thesis places the developed models in a complex, high-risk collaborative environment, with a specific focus on a medical application. The chapter examines the role of robotic inner speech in cooperation with a nurse during the preparation of a surgical table, a task that requires precision and careful instrument placement to avoid adverse procedural outcomes. The study analyses how the robot's self-directed dialogue contributes to reassurance and stress management in demanding contexts, while simultaneously enhancing clarity and understanding of task-related instructions. The findings demonstrate that inner speech retains its effectiveness in more complex and high-stakes interaction settings, supporting transparency, trustworthiness, and performance under elevated cognitive and emotional demands.

From a broader perspective, this work advances affective and cognitive robotics by introducing a mechanism that bridges the gap between reactive emotional systems and reflective, self-regulating architectures. Traditional models often rely on surface-level mappings between external events and emotional displays, resulting in behaviours that may appear expressive but lack the deeper coherence that characterises human affective life. By incorporating inner speech, the models developed in this dissertation enable robots to engage in self-referential processing, thereby enriching their emotional landscape and enhancing their social presence. This integration opens new avenues for research on how artificial agents can develop and maintain internal narratives, how

such narratives can be communicated to human partners to increase transparency, and how they might evolve through learning and interaction.

The thesis work concludes with Chapter 4.

Published Content

This thesis is based on the following scientific papers:

1. **S. Corvaia**, A. Pipitone, A. Cangelosi and A. Chella, *"Inner Speech and Extended Consciousness: a Model based on Damasio's Theory of Emotions,"* 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, MA, USA, 2023, pp. 1-8, doi: 10.1109/ACIIW59127.2023.10388201

2. A. Pipitone, **S. Corvaia**, A. Chella, *"Towards robot affective appraisal linking inner speech and emotion"*, in Robotics and Autonomous Systems, vol. 198, pp. 105363, 2026, doi: https://doi.org/10.1016/j.robot.2026.105363

3. **S. Corvaia**, A. Pipitone, A. Chella, *"Inner Speech and Damasio's Theory for Modelling Robots' Emotions"*, in IEEE Transactions on Affective Computing, pp. 1-14, 2025, doi: 10.1109/TAFFC.2025.3547756

4. 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"*, in Intelligent Service Robotics, vol. 19, no. 5, 2025, doi: https://doi.org/10.1007/s11370-025-00656-4

Contents

List of Figures

List of Tables

Part 1: Computational Models for Emotion Generation

Chapter 1

Reasoning Aloud and Emotional Processing in Robots

1.1 Overview

Recent research in Robotics and Artificial Intelligence has demonstrated that robots engaging in *think-aloud* behaviour during human–robot collaboration receive positive feedback from their human counterparts and facilitate the achievement of shared goals. By voicing their internal reasoning, robots improve transparency and provide insight into their underlying decision-making processes. Furthermore, this behaviour allows the robot to evaluate alternative strategies for performing joint tasks, thus increasing the robustness of the interaction. In this chapter, the role of a robot's inner speech in shaping its emotional responses will be investigated. According to appraisal theories, emotions arise from cognitive evaluations of specific situations. The internal dialogue serves as a form of self-reflection that aids this evaluation. Through inner speech, the robot can focus on contextually relevant factors, gather essential information for computing assessment variables, and, consequently, generate appropriate emotional responses. At the same time, the human partner has access to the robot's reasoning processes that underlie its emotional states, thereby improving mutual understanding. The proposed model exhibits emotional dynamics consistent with those observed in healthy adults under stressful conditions. These results indicate that robot emotional reactions align with expected human-like patterns and that integration of inner speech enhances the performance of an established computational model of emotions.

1.2 Introduction

Philosophers, psychologists, and neuroscientists have broadly investigated the role of inner speech in human cognition and psychological processes, including affectivity and emotions [2] [3] [4]. Inner speech describes the linguistic manifestation of thought, and individuals engage in self-dialogue when they process their thoughts through internal verbal commentary [5]. This cognitive mechanism supports the concentration of contextual factors, action planning, decision maintenance, and awareness of facts and events.

Vygotsky first proposed the connection between inner speech and emotions [6], who described the continuous and dynamic interplay between intellect, defined as thought, and the affective domain. This relationship evolves over the course of life and operates bidirectionally: *from the affective sphere of consciousness to thought, and from thought to the affective sphere of consciousness* [7]. Lazarus supported a similar perspective [8] [9], and, more broadly, that of proponents of cognitive appraisal theories. According to this framework, mental processes play a fundamental role in affectivity because thought precedes and shapes the cognitive processes leading to emotional experience. Specifically, the process is initiated by a stimulus, followed by the emergence of a linguistically encoded thought associated with that stimulus, and culminates in a physiological or emotional response. Thus, the role of cognition in generating emotions is fundamental.

In the context of modelling emotional behaviour in artificial agents, including robotic systems, cognitive appraisal theories are extensively adopted, as they offer a theoretically grounded framework for explicating both the evaluative processes and the intensity of emotional responses through appraisal variables. These variables serve as bridges between environmental context and emotional components and are particularly amenable to computational modelling. The multidisciplinary field of *affective Human-Robot Interaction (HRI)*, which extends the scope of *Affective Computing* [10], explores how systems can be designed, implemented, and evaluated to incorporate affective processes [11]. Machines capable of expressing emotions and providing emotional feedback have been shown to enhance user enjoyment [12] [13], engagement [14], and task performance [15]. Despite these noteworthy advancements, the role of inner dialogue remains unexplored, particularly within the existing body of research, including its potential function as a fundamental mechanism for the computation of appraisal variables and, consequently, for the elicitation of artificial emotions.

This work introduces a novel computational model that establishes a tight coupling between inner speech and emotional processes. Some of the authors previously proposed a cognitive architecture for inner speech [16][17] and implemented it on a physical robot, thus creating the first robot capable of *thinking out loud*. Such an abil-

ity has demonstrated several benefits in collaborative human–robot tasks, including enhanced transparency (i.e., the ability to trace and reproduce underlying decision-making processes), increased robustness (i.e., the ability to overcome deadlock situations and complete tasks) [18], and improved trustworthiness (i.e., fostering greater human trust toward the robot) [19]. The present study systematically examines the role of this capacity in shaping the affective processes of robotic agents within the conceptual framework of appraisal theory.

The proposed model is inspired by the modal model of emotion regulation introduced by Gross [20], which outlines four sequential stages involved in emotional self-regulation: Situation, Attention, Appraisal, and Response. A feedback link from the Response stage to the Situation stage models the processes of coping and regulation.

While preserving the general structure of Gross's model, the present approach highlights the Attention stage, where inner speech plays a central role in cognitively evaluating the context by focusing on relevant aspects for computing appraisal variables.

Once a stimulus is transformed into a thought, represented as a linguistic surface form, the inner dialogue begins, enabling further evaluation of both external (environment-related) and internal (state-related) facts and events. This process is conceptualised as a continuous rehearsal loop, in which each successive instance of inner speech emerges in direct response to the immediately preceding one, thereby establishing a sequential and self-sustaining cognitive dynamic. The loop continues until no additional facts require evaluation or all necessary information has been inferred. At this stage, the Appraisal component computes appraisal variables that have been formalised to reflect patterns observed in healthy adult populations [21]. Thereafter, the corresponding emotion and its intensity are derived based on Russell's Circumplex Model of Affect [22], and the emotion is elicited through the Response stage. In the current implementation, only the externalisation of emotion and its intensity are considered part of the coping strategy.

This study focuses on two key contributions:

1. the use of an inner speech rehearsal loop for the cognitive evaluation of contextual information, thereby facilitating the collection of data necessary for appraising a situation;

2. the mathematical formulation of appraisal variables derived from inner speech, which enables the computation of emotions with specific intensities.

Throughout this process, the robot verbalises its internal dialogue, allowing an observer to trace the reasoning that leads to its emotional state in response to a given context. The verbose descriptions of these processes are hand-annotated and instantiated at execution time based on the concepts involved.

The proposed model has been validated by exposing the robot to simulated stressful situations and assessing whether its emotional behaviour aligned with established trends described in [23], which outlines a method for evaluating computational emotion models. Specifically, this evaluation framework was initially designed for the EMA model (EMotion and Adaptation model [24]) and involves simulating loss-related and aversive scenarios, then comparing the computed appraisal variables with those obtained from the SCPQ (Stress and Coping Process Questionnaire [21]).

The evaluation procedure employed in this study reproduces four canonical situations derived from the SCPQ tool, and the resulting appraisal variables and emotional reactions are compared against both EMA's outcomes and the SCPQ's normative trends.

The findings are particularly promising: the mathematical formalisation of appraisal variables not only reproduces appraisal patterns and emotional responses consistent with theoretically anticipated trends, but also, in several instances, demonstrates superior performance relative to the EMA model, thereby highlighting its potential contribution to the advancement of affective computing.

The remainder of this chapter is structured as follows: Section 1.3 introduces the theoretical foundations of the model, including the appraisal framework, Gross's modal model, and the role of inner speech in emotional processes. Section 1.4 details the proposed model, which links inner speech to emotion generation, including the mathematical formalisation of appraisal variables and the elicitation of emotions with specific intensities. Section 1.5 presents the evaluation methodology and the obtained results. Section 1.6 illustrates two use cases involving collaborative tasks between a robot and a human partner. Section 1.7 discusses the related work, and Section 1.8 concludes the thesis with final remarks and potential future research directions.

1.3 Foundations of Cognitive and Affective Robotics

1.3.1 Appraisal Theories of Emotion

Appraisal theories [25] [26] [27] claim that individuals interact with specific aspects of their surrounding context that hold personal relevance, allowing a subjective interpretation of the situation.

According to appraisal theorists, the elaboration of these relationships is responsible for the emergence of emotions. Each emotional response originates from a *cognitive evaluation* of the situation and the corresponding *meaning structure* that emerges from this evaluation.

Cognitive evaluation begins with perception and consists of an often automatic, in-

voluntary assessment of the presence or absence of specific entities or events, along with their potential positive or negative consequences, as the individual interprets them. During this process, a structure representing emergent meanings is constructed that tracks the components that ultimately shape the emotional response—this structure is referred to as the meaning structure.

Within empirical contexts, cognitive evaluation is typically described in terms of *appraisal variables*. Appraisal theories differ in how these variables are defined and in how they are assumed to influence the dominant emotional outcome. As a result, direct comparison and theoretical convergence remain challenging. Nevertheless, these approaches share a common foundation: a bottom-up evaluation strategy in which each emotion is elicited by a specific and distinct pattern of appraisal variables derived from low-level contextual information.

Appraisal theories have become a dominant framework in computational models of emotion due to their emphasis on representing emotions as computable constructs and their relative ease of implementation.

In the most classical formulations of appraisal theory [28] [29], emotions are conceptualised as discrete entities organised within a taxonomy. Each emotion is assigned to a specific category based on its properties and features. However, such rigid classification introduces ambiguities: the same feeling may be elicited by different events and may carry diverse meanings not strictly linked to its defining features.

In other words, the identified features of emotions are not universally valid, as emotional reactions are highly subjective and influenced by individual experiences, cultural backgrounds, lifestyles, and personal interpretations of events. More recent appraisal theories [30] adopt a less rigid view, proposing that emotions are not necessarily perceived as discrete entities but rather as phenomena that can vary in intensity and gradation. As a consequence, the formalisation of appraisal variables is not always universally defined but may depend on individual personality traits and personal history.

Within this perspective, the present model simulates the robot's cognitive evaluation through inner speech, accounting for its specific characteristics. This includes identifying key features that may affect such evaluation, such as internal operational conditions (e.g., battery state, joint functionality, temperature) and external factors (e.g., environmental disorder that may influence signal discrimination). All these elements contribute to the appraisal of the context and, ultimately, to the robot's emotional experience.

Figure 1.1: The modal model of emotion regulation proposed by Gross

1.3.2 Emotion Elicitation and Regulation: The Modal Model

Gross [20] defines emotions as brief episodes that influence both behaviour and physiological states, arising in response to events that present potential challenges or opportunities. He argues that emotions are not static phenomena but can be modulated, giving rise to the process of emotional self-regulation. To describe this process, Gross proposed the *modal model*, illustrated in Figure 1.1, which outlines the following sequential stages:

1. **Situation**: the process begins with an emotionally relevant situation, which may be either real or imagined;

2. **Attention**: the individual directs focus toward the emotionally salient situation and selectively evaluates the facts deemed significant;

3. **Appraisal**: the situation is cognitively interpreted, leading to the emergence of an emotion with a specific intensity;

4. **Response**: an emotional response is produced, involving changes across experiential, behavioural, and physiological domains.

Within the modal model, emotional events can be influenced by modifying cognitive processes, which in turn affect the emotional response. This response may subsequently alter the situation itself, forming a feedback loop that continues to regulate emotional mechanisms until the emotionally relevant episode concludes.

Gross conceived this framework to encompass a broad spectrum of processes, including automatic and controlled, as well as conscious and unconscious mechanisms. This theoretical foundation serves as the backbone of the proposed model, which preserves the same stages to structure the cognitive sequence for appraising the context.

1.3.3 The Role of Self-Talk in Emotional Processing

In contrast to biological theories of emotion [31], which attribute emotions primarily to organic and physiological origins, appraisal theorists integrate affective phenomena with the broader spectrum of human psychological functions.

In his *Theory of Emotions*, Vygotsky [6] developed the *interfunctional theory of emotions*, which emphasises the inseparable relationship between mental life and bodily manifestations, rejecting the notion that the body alone constitutes the primary source of emotional experience. He highlighted the essential, dynamic, and dialectical interplay between cognition and physiology, which shapes psychological experiences. According to this perspective, words do not merely serve as mechanisms for expressing thought; instead, they represent the endpoint of thought. Thoughts, in turn, act as mediating tools for experiencing the self and the surrounding context, and are ultimately expressed through language.

When experiencing an emotion, thought serves as the medium through which this experience is processed and lived, and it is eventually materialised, either covertly or overtly, in words. Consequently, the psychological experience of emotions is intrinsically interconnected with thought and ultimately expressed linguistically.

The causes and effects of this interconnection are not static but evolve throughout development, creating continuous interactions between intellect and affect. This relationship is bidirectional: *the word nominates the affection, and the affection, therefore, is channelled in thought through the word* [7]. The meeting point between thought and emotion is the *sense*: words enable reflection upon affective states, while affective states, in turn, give rise to thought and verbal expression within consciousness.

Empirical evidence supports the existence of this link. For instance, Morin [32] found that the standard content of inner speech includes self-directed evaluations and emotional states. Through self-talk, individuals can experience, become aware of, and regulate their emotions and act accordingly.

The present work focuses on developing a computational model that captures this interconnection between inner speech and emotions. Through self-talk, the robot engages with its context, using linguistic reasoning as a mediating tool, consistent with Vygotsky's theory, to perform cognitive evaluations of situations. Once the inner dialogue concludes, the resulting appraisal pattern enables the emergence of an emotional experience. This emotional experience, in turn, generates a corresponding thought through which the robot externalises its emotion and becomes aware of its own affective state.

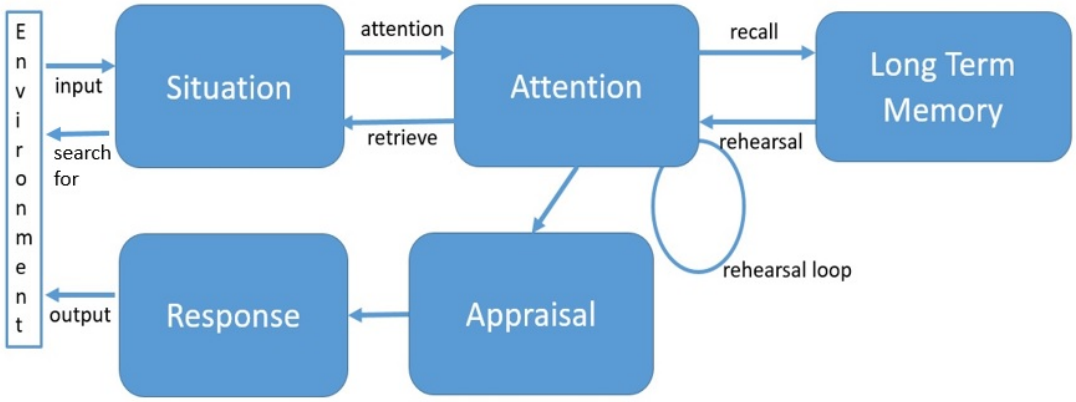


Figure 1.2: The proposed cognitive architecture of inner speech and emotions.

1.4 Modelling Emotional Processes in Robots via Inner Speech

Figure 1.2 illustrates the proposed cognitive architecture for integrating inner speech and emotions. As described previously, the architecture is built on the modal model's backbone, extended with a rehearsal loop for inner speech. Additionally, a dedicated memory module stores the robot's knowledge and supports the retrieval and analysis of concepts related to the stimuli it perceives.

In summary, the architecture enables the robot to generate emotional responses when perceiving environmental stimuli. The Situation module encodes the stimulus by associating it with a symbolic representation. The Attention module then retrieves related concepts from long-term memory in linguistic form, thereby representing the emergence of a thought associated with the stimulus and initiating inner speech. This thought is subsequently processed again by the Attention module, which may prompt further retrievals from either the Situation module (for additional environmental stimuli) or memory (for related concepts). The retrieval process is guided by the content of the current inner speech turn and by its potential matches with new environmental or memory-based concepts.

This inner monologue facilitates cognitive evaluation and provides the parameters for the Appraisal module, which computes the appraisal pattern and infers the corresponding emotion. The Response module finally externalises the resulting emotion. Detailed descriptions of the inner speech processes for cognitive evaluation are provided in Subsection 1.4.2.

The model relies primarily on *general appraisal variables*, though *specific appraisal variables* may be introduced depending on the application scenario. General appraisal variables, drawn from established appraisal theories, are widely applicable across con-

texts. The general variables incorporated in the proposed model are:

- *Likelihood*, measuring the probability of an event's outcome, with a higher likelihood of negative outcomes indicating more stressful events;
- *Controllability*, assessing the extent to which the robot can influence the outcome by directly acting on the context;
- *Changeability*, evaluating whether other events or agents can alter the outcome;
- *Desirability*, quantifying how favourable or unfavourable the outcome is.

Mathematical formulations for each of these variables are defined in Subsection 1.4.3.

While literature often links likelihood to emotions like fear and hope, the choice of likelihood, controllability, changeability, and desirability was motivated by their direct applicability to robotic contexts, where controllability reflects the robot's agency, and changeability accounts for environmental unpredictability [33]. Inner speech facilitates the computation of these variables by enabling cognitive evaluation through self-dialogue, for example, assessing likelihood from internal states and desirability from contextual relevance [34] [35].

Specific appraisal variables include scenario, dependent, and context-specific factors that influence the robot's emotional experience. For instance, in a collaborative scenario in which a robot and a human partner set a table according to an etiquette schema, a specific appraisal variable may encode whether the partner correctly places the utensils. Deviations from expected behaviour elicit a more unpleasant emotional response, with the particular variable contributing to the overall emotion. As it is impractical to enumerate all possible scenario-specific variables, they can be added incrementally in future extensions of the model, specifying the direction of their influence (positive or negative), as further illustrated in Section 1.6.

Once the appraisal variables are computed, the corresponding emotion is inferred using the Circumplex Model of Affect [22], in which emotions are represented within a two-dimensional space defined by *valence* (indicating the degree of pleasantness or unpleasantness) and *arousal* (reflecting the level of physiological activation). Emotions are thus represented as points in this space. Although the Circumplex Model encompasses 28 discrete emotions, the proposed model considers only the five basic emotions identified by Ekman [36]: *happiness*, *sadness*, *fear*, *anger*, and *disgust*. The mapping from appraisal variables to valence and arousal, and the derivation of the resulting emotion, are described in Subsection 1.4.3. Furthermore, the emotion's position within the circumplex space allows an estimation of its intensity.

1.4.1 The Knowledge Structure for Emotional Representation

The robot's knowledge encompasses two domains: the *external world* and the *internal world*. The external world comprises all the facts and entities in the robot's environment, representing its general knowledge of its surroundings. The internal world, in contrast, represents the robot's self-knowledge, including its physical and operational states, such as battery level, joint functionality, or the operational condition of its arms.

This knowledge is formalised through an ontology, referred to as the *KB ontology*, which defines general concepts, their attributes, and the relationships among them. In this framework, general concepts correspond to ontology *classes*, while *instances* represent specific entities in the environment (for classes modelling external-world concepts) or the robot's concrete state (for classes modelling internal-world concepts).

Formally, the KB ontology is represented as the tuple $O = \langle C_o, P_o, T_s, L, P_d \rangle$ in accordance with the W3C technical report specification[1], where:

- $C_o = \{cl_i\}$ 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, so that:

$$P_o = \{o_i \mid o_i = (cl_j, cl_k) \;\; cl_j, cl_k \in C_o\};$$

- $T_s = \{t_i\}$ is the set of literal datatypes, that is, types of data of the attributes of the class (numerical, string, and so on);
- $L = \{l_i\}$ is the set of literal values of t_i (the instance of type t_i, for example, for a numerical datatype, the literal value *3*);
- P_d is the set of the datatype properties, linking a concept and a datatype, so that:

$$P_d = \{d_i \mid d_i = (cl_j, l_k) \;\; cl_j \in C_o, \; l_k \in L\}.$$

For example, the class `Person` represents a human as a concept of the external world. In contrast, an instance of this class models a specific individual perceived in the current environment. Attributes such as `emo` and `age` are examples of datatype properties that represent the individual's emotional state and age, respectively. Object properties may define the relative positions of the individual with respect to other entities, including `left`, `right`, `up`, and `down`, which link the individual to other entities via spatial relations.

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

The ontology also represents the robot's internal world. For instance, the class `Emotion` and its subclasses `Anger`, `Joy`, `Sadness`, etc., model the robot's possible emotional states. Instances of these classes define the robot's current emotional state at any given time.

It is important to distinguish between emotion simulation, emotional awareness, and experienced emotion. Emotion simulation refers to the robot's ability to mimic emotional expressions based on predefined rules or learned patterns, without genuine internal processing. Emotional awareness involves the robot recognizing and interpreting its own simulated emotions through self-reflection, as facilitated by inner speech. Experienced emotion goes further, where the robot processes emotions as emergent from cognitive appraisals and bodily-like signals, leading to coherent internal states that influence behavior. For example, in this model, the ontology enables emotional awareness by structuring knowledge for inner speech evaluation, while the integration with appraisal variables supports experienced emotion beyond mere simulation.

1.4.2 Using Inner Speech to Evaluate Contexts Cognitively

The central principle of the proposed architecture is the robot's *inner voice*, which enables it to selectively attend to the situational concepts that may influence its emotional state. Through this mechanism, the robot engages in an "internal reasoning" process with respect to an incoming stimulus, focusing on concepts semantically related to it. This reasoning is symbolic in nature and operates on the linguistic surface representation of concepts. This approach aligns with the notion that *"words allow thinking about emotions, and emotions, in turn, generate further words or thoughts within the inner dialogue"* [7].

The inner dialogue facilitates the *cognitive evaluation* of a situation by progressively identifying its *meaning structure*. According to appraisal theories, this process provides the necessary values for the appraisal variables by drawing on both the robot's internal knowledge and the state of its surrounding environment.

In the proposed model, the meaning structure is defined as a pair:

$$T = [sem(.),\ syn(.)]$$

which associates the semantic component $sem()$ with the syntactic component $syn()$ of a word or set of words. The individual elements within the $sem()$ and $syn()$ components are referred to as *chunks*.

- The $sem()$ component contains *meanings*, i.e., chunks corresponding to concepts represented in the knowledge base.

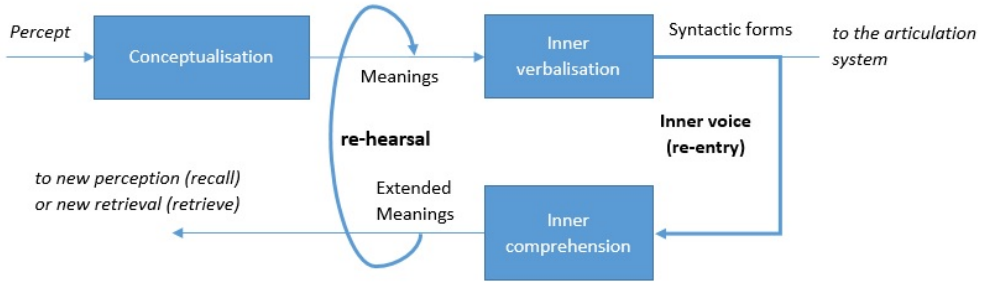


Figure 1.3: The language re-entrance components: the syntactic forms from the inner verbalisation component are input to the inner comprehension one; further expansion of meanings allows reasoning about beliefs and internal state, thus identifying emotionally relevant situations and defining attention.

- The *syn*() component contains *words*, i.e., chunks corresponding to the linguistic tokens.

As the process unfolds, the meaning structure evolves dynamically: each new meaning and word is incorporated into its respective component, unless it is already present.

This mechanism, illustrated in Figure 1.3, draws inspiration from Steels' work [37]. The rehearsal loop operates through three main stages:

1. *Conceptualisation*

   A percept p represents the symbolic form of a perceived stimulus, which may consist of a voice stream, an image of an object, one or more of its features, and so forth. In any case, the percept is expressed textually and syntactically describes the perceived stimulus (e.g., text generated by speech recognition, descriptions of objects or features derived from image processing, etc.).

   This step aims to establish correspondences between the percept's syntax and the robot's knowledge base. Let $S = \{s_1, s_2, \ldots, s_n\}$ denote the set of tokens of the percept p, pre-processed to remove non-meaningful words (e.g., articles, conjunctions, prepositions). The initial meaning structure is thus:

   $$T = [\, sem(),\ syn(s_1, s_2, \ldots, s_n)\,].$$

   Conceptualisation consists of *identifying the KB ontology concepts that best match the percept's syntax*: the matched ontology classes serve as conceptualisations of the percept. The matching is performed using the Jaro–Winkler distance [38], which measures the syntactic similarity between two words. Let $d_{jw}(w_1, w_2)$

denote this distance, and let $couple(r)$ return the domain and range of a property r. The following functions model the matching between S and the ontology O:

- $a_c : S \rightarrow C_o$, where

$$a_c(s_i) = \{\ cl_k \mid d_{jw}(s_i, cl_k) > \alpha,\ cl_k \in C_o\}$$

  returns the set of ontology classes whose labels are syntactically similar to the words in S, with α being the similarity threshold.

- $a_p : S \rightarrow P_o \cup P_d$, where

$$a_p(s_i) = \{\ r_j \mid r_j \in P_o \cup P_d,\ couple(r_j)\ \supset\ a_c(s_i)\}$$

  returns the set of ontology properties whose domain or range includes the concepts retrieved by a_c.

The threshold α was empirically set to 0.2 after parameter tuning.

The result is a sub-ontology

$$O_m = \langle C_m, P_m \rangle,$$

where $C_m = \bigcup_S a_c(s_i)$ and $P_m = \bigcup_S a_p(s_i)$. This sub-ontology includes all the concepts and relations corresponding to the percept. The meaning structure then becomes:

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

with $m_i \in C_m \cup P_m$.

The conceptualisation process concludes by retrieving additional annotations related to the percept that are specified in the entities of the sub-ontology O_m. These annotations share the same form as FCG meanings. Consequently, the complete set of retrieved meanings is given by:

$$M_c = M \cup M_a,$$

where M_a denotes the set of meanings contained in the annotation properties of O_m, which may also be empty.

2. *Inner Verbalisation*

   Once the meanings of the percept have been inferred, they are verbalised. This means that the retrieved meanings, representing the knowledge elicited by the

stimulus (i.e., the robot's experience of the stimulus), are transformed into an initial internal utterance.

To produce this utterance, the meaning structure is first reduced to the semantic component:

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

The structure is then reconstructed by associating each meaning with its corresponding label in the ontology:

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

where label(c) returns the label of class c in the ontology.

These labels are verbalised and fed back to the inner comprehension component, effectively being "re-heard." Notably, this production step may include additional syntactic tokens associated with the meanings that emerged during conceptualisation, potentially involving a set of concepts different from those in the original percept.

3. *Inner Comprehension*

   The labels produced during inner verbalisation represent new thoughts, which may necessitate additional perception or retrieval of concepts from the knowledge base. Inner comprehension seeks these new labels either in the environment (retrieve) or in the knowledge base (recall). Two main cases are considered:

   (a) If the meaning corresponds to a property ($m_i \in P_m$), the module searches the knowledge base for the property's domain or range not yet present in the meaning structure (recall) and adds it to the $sem()$ component, if absent.

   (b) If the meaning corresponds to a class ($m_i \in C_m$), the module:

   - searches the environment (retrieve) for the corresponding entity; if found, its token is added to the $syn()$ component and conceptualisation restarts; otherwise, m_i is removed;
   - searches for other properties of the class in the knowledge base (recall) and, if found, adds them to the meaning structure if not already present.

Table 1.1: Illustration of how inner speech supports cognitive evaluation. The table details the meaning structure at each phase of the inner speech loop, indicating the loop stage (C: Conceptualisation, IV: Inner Verbalisation, IC: Inner Comprehension) and the corresponding process (recall, retrieve, or produce). In this example, the robot perceives the auditory command *"take the plate"* from its partner. The initial meaning structure contains the relevant chunks derived from the encoded input. Subsequently, the loop recalls related concepts from the ontology and enriches the structure, initiating the inner speech process. All emergent concepts at each phase are highlighted 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**

Meaning structure	**Loop phase**	**Process**
[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**

At the end of each phase, the meaning structure grows as new chunks are added. The rehearsal loop is repeated until no further elements can be incorporated into the meaning structure. The robot then stores all information regarding the context, including internal states, which are subsequently used to compute the appraisal variables and to trigger the corresponding emotional experience.

Table 1.1 illustrates a use case of the described method. For each phase of the loop, the table reports the current meaning structure and the process responsible for modifying it.

In this example, the robot receives a command from the human partner, specifically *"take the plate"*. This percept is processed to populate the *syn*() component of the initial meaning structure. For each phase, the *process* column indicates the concepts and properties emerging either from the knowledge base (recall) or from the environment (retrieve). These entities are highlighted in **bold**. The production of the verbalised sentence (produce) corresponds to the inner verbalisation step.

The example clearly shows how new entities are iteratively added to the meaning structure and how the structure evolves across successive phases of the loop. Information related to the robot's internal world is retrieved from the knowledge base and inferred via standard robot libraries, which monitor battery levels, motor temperatures, and other operational parameters, providing the necessary inputs for evaluating the robot's state.

Through inner speech, the robot achieves emotional awareness by reflecting on its internal states and contextual evaluations, distinguishing this from mere simulation of emotions. For instance, while simulation might involve predefined responses, inner speech enables the robot to experience emotions by articulating and processing appraisals, as seen in the rehearsal loop examples where thoughts evolve to compute meaningful emotional responses.

1.4.3 Appraisal Variables and Their Link to Emotions

The meaning structure represents the output of the cognitive evaluation performed by the robot through inner speech. Once the rehearsal loop concludes, the values of the appraisal variables can be computed. The proposed model introduces a novel mathematical formalization of these variables, calibrated according to the expected trends reported in [21]. The formulas were iteratively tested and tuned to optimize the emotional responses of the robot in accordance with these trends.

The formalization of the appraisal variables explicitly accounts for the environmental conditions, as these can significantly affect the robot's evaluation of the context. In particular, the model defines the *environmental entropy* k as an indicator of the level of disorder in the environment: higher entropy corresponds to greater environmental unpredictability.

To compute the entropy, the model considers both the environmental noise, which can hinder the recognition of auditory stimuli, and the emotional state of the human partner, which may indicate approval or disapproval and thus increase situational unpredictability.

The robot's perception routines extract these features and provide values for the entropy calculation. Let e denote the set of possible partner emotions detectable by the robot; these are categorized as *negative*, *neutral*, or *positive*. The entropy depends on three parameters:

- α, representing the level of environmental noise;
- β, representing the partner's emotion inferred from facial expressions;
- γ, representing the partner's emotion inferred from vocal tone.

These parameters are formalized as follows:

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

and

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

Considering that environmental unpredictability increases with noise and the presence of negative partner emotions, the environmental entropy k is modeled as a linear combination of these parameters:

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

This environmental entropy is then incorporated into the computation of the appraisal variables, ensuring that the robot's evaluation of the situation dynamically reflects the current environmental conditions.

Likelihood L

The likelihood L represents the probability that a negative outcome of an event will occur, potentially leading to stressful situations [21]. Variations in this value indicate the evolution of the event: as the likelihood increases, the probability of the negative outcome grows, and consequently, the stress level rises. Conversely, a decreasing likelihood suggests that the negative event is being resolved, and the associated stress diminishes.

In the proposed model, the likelihood is manually encoded based on specific work conditions, as these significantly influence the feasibility of the required actions. The considered work conditions, along with their corresponding likelihood values, are reported in Table 1.2. These conditions encompass the robot's physical state, the feasibility of the required action (e.g., whether the object involved in the action is reachable or visible), and the admissibility of the action (e.g., whether the action follows prescribed rules).

The likelihood is higher under negative work conditions, reflecting the increased probability of failure. For instance, if a physical component of the robot is malfunctioning, the likelihood of an event involving that component is set to 80%, indicating a high risk of failure. If an alternative component is used or the issue is resolved, the likelihood decreases to 20%.

When multiple work conditions co-occur, the final likelihood is computed as the weighted sum of the individual likelihood values of each condition. The weight is set to 1 for negative conditions and to -1 for positive conditions, ensuring that the combined effect accurately reflects the overall probability of a negative outcome.

Table 1.2: 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%

Controllability C

The controllability C quantifies the extent to which the outcome of an event can be influenced by directly acting on the context [21]. In the proposed model, higher likelihood values, indicating a greater probability of a negative outcome, correspond to a lower capacity to modify the context and mitigate negative progression. Similarly, greater environmental disorder reduces the potential to control the situation.

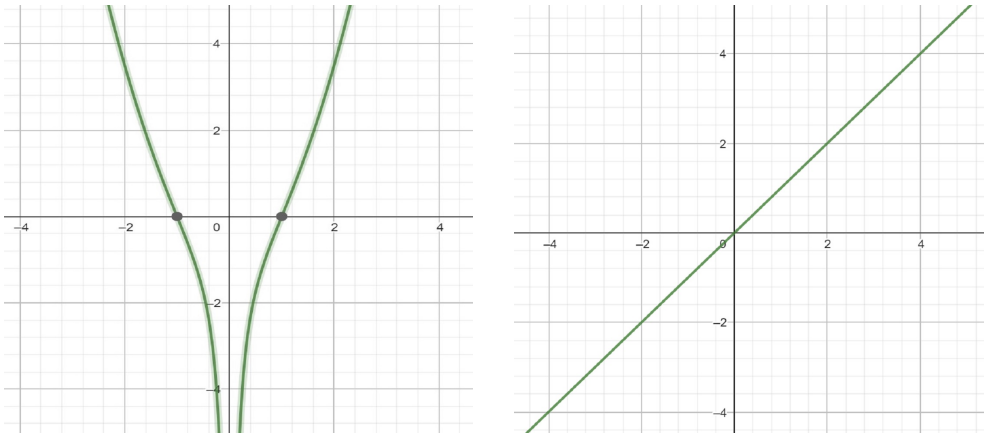
To capture this relationship, controllability is formalised as:

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

where $x \in]0, 1]$ represents the absolute difference between two consecutive measurements of the likelihood, i.e., $x = |L_a - L_p|$, with $L_a \neq L_p$. Here, L_a denotes the antecedent likelihood, while L_p is the present likelihood. For the first measurement, L_a is set to zero.

The parameter x indicates the evolution of the event. A decreasing x implies that the likelihood values in consecutive measurements are similar; thus, the situation remains largely unchanged, and controllability remains stable. Conversely, a variation in x signals a change in the situation. When x decreases for negative likelihood variations or increases for positive ones, the likelihood decreases, the problem improves, and controllability increases. On the other hand, if x decreases for positive likelihood variations or increases for negative ones, the likelihood is rising, the situation is worsening, and controllability decreases.

Figure 1.4a illustrates the trend of the C function. Increasing the environmental entropy k compresses the projection of C along the x-axis, meaning that higher disorder accelerates the decrease in controllability. Nevertheless, the general trend of the C function remains consistent, independent of the value of k. The independence of the function's trend from k will be further discussed in the following sections.



(a) (b)

Figure 1.4: (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 represents the extent to which the outcome of an event can be influenced by an external event, over which the individual has no direct control [21]. It provides a measure of the context's unpredictability, since the outcome does not depend on the agent.

According to this definition, a higher likelihood of an adverse outcome corresponds to a greater probability that the situation becomes more unstable. Similarly, higher environmental entropy increases the problem's unpredictability. Consequently, changeability is formalised as:

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

where x is defined as in the previous section, i.e., $x = |L_a - L_p|$.

The changeability function is linear, as shown in Figure 1.4b. The value of k determines the slope of this line: lower entropy values result in a steeper slope, causing faster variations in changeability. This behaviour is consistent with the notion that higher environmental entropy leads to a more unpredictable situation, making the context increasingly difficult to anticipate.

When x increases for positive or negative values, the changeability grows because the situation is bad 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 quantifies whether the outcome of an event is considered favourable or unfavourable. It can assume positive or negative values [24], depending on the expected consequences of the event. Formally, desirability is defined as:

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

This binary representation provides a straightforward assessment of the event's valence, which is subsequently used to determine the overall emotional response of the robot within the appraisal framework.

Matching the appraisal variables to the emotions

Once the general appraisal variables have been computed, they are combined to infer the corresponding emotion. For this purpose, the model refers to the bi-dimensional Russell's circumplex space [22], where each emotion is represented as a point with coordinates *valence* v and *arousal* a, each ranging in $[-1, 1]$.

Valence represents the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, whereas arousal quantifies the level of physiological activation associated with the emotional experience. In the proposed model, the appraisal variables are used to compute valence, as they reflect the conditions of the situation and its relative desirability or threat. Arousal is derived from the inner states of the robot, representing its physiological status. The emotion is then inferred by projecting the computed valence and arousal into the Russell's space.

Specifically, a situation exhibits positive valence when controllability is high, as this indicates the potential to influence or improve a negative outcome. Similarly, the desirability of an outcome contributes positively to valence. Conversely, high changeability, which reflects contextual unpredictability, contributes negatively to valence. Before mapping appraisal variables to valence, each variable is normalized to the range of the Russell's space.

Normalization is computed based on the maximum and minimum values of each variable. Accordingly, valence is modeled as a linear combination of the normalized appraisal variables:

$$v(C, M, D) = N(C - M + D)$$

where C, M, and D denote controllability, changeability, and desirability, respectively, i.e., $C = C(k, x)$, $M = M(k, x)$, and D as defined previously. The function $N(.)$

Table 1.3: 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 Russell's space as defined at [1].

	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*

normalizes its argument to the interval $[-1, 1]$ and is formalized as:

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

where r_1 and r_2 are the original range extremes, and n_1 and n_2 define the new target range.

Arousal is modeled using the robot's internal physiological states, specifically the battery level B and the inner temperature T. Arousal is higher when the battery is sufficiently charged and the temperature is within optimal limits, as detected by standard monitoring routines. Formally, arousal is computed as:

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

Finally, the robot's emotion is inferred by projecting the appraisal pattern $\mathbf{a} = (v, a)$ into the Russell's space. For simplicity and interpretability, only the five basic emotions defined by Ekman are considered in this projection.

Let E be the set of labels of Ekman's basic emotions, that is

$$E = \{\text{Angry, Happy, Afraid, Annoyed, Sad}\}$$

and let R be the corresponding set of coordinates in Russell's space. Each element $\mathbf{r}_i \in R$ is a pair of coordinates $\mathbf{r}_i = (v_i, a_i)$ representing the i-th emotion in E, as reported in Table 1.3 [1].

We define a function $l(\mathbf{r}_i)$ that returns the label of the emotion in E corresponding to the point $\mathbf{r}_i \in R$. Given an instance of the appraisal pattern $\mathbf{a} = (v, a)$, the emotion $e_{\mathbf{a}}$ is determined by selecting the emotion in E whose coordinates in R are closest to $\mathbf{a}$ in terms of Euclidean distance:

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

In this way, each computed appraisal pattern is associated with the most similar

basic emotion in Ekman's set, allowing the robot to express a discrete emotional state derived from the continuous valence-arousal space.

The computation of the emotion's intensity

The intensity of an emotion represents the degree to which the emotion is experienced vividly and profoundly. Note that intensity differs from arousal: while arousal measures the level of physiological activation associated with the emotional experience, intensity quantifies the vividness or strength of the emotion itself. In the proposed model, three levels of intensity are defined, corresponding to different emotional reactions. High intensity refers to emotions that are difficult to control and regulate, for instance, resulting in bursts of joy or anger. Medium intensity corresponds to emotions that can be regulated and managed, while low or null intensity represents minimal emotional involvement, approaching indifference.

The intensity of an emergent emotion is determined by the proximity of the computed appraisal pattern $\mathbf{a} = (v, a)$ to the point $\mathbf{r}_i \in R$ in Russell's space corresponding to the identified emotion $e_{\mathbf{a}}$, as reported in Table 1.3. The closer the appraisal pattern is to the emotion's point, the stronger the experienced intensity.

Formally, the model defines two threshold values, a high threshold t_h and a low threshold t_l, based on the Euclidean distance d in Russell's space between the emotion's point $\mathbf{r}_i$ and the origin $\mathbf{o} = (0, 0)$:

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

The corresponding intensity of the emotion is assigned as follows. The intensity is *high* if $\mathbf{a}$ falls within a circumference centred at $\mathbf{r}_i$ with radius t_h. The intensity is *medium* if $\mathbf{a}$ falls between the circumferences with radii t_h and t_l, both centred at $\mathbf{r}_i$. Finally, the intensity is *low* if $\mathbf{a}$ falls outside these two circumferences.

Each intensity level triggers a distinct behavioural response in the robot. Specifically, the robot verbalises its emotional state: high intensity is expressed by adding the adjective *very* to the label of the emergent emotion; medium intensity is described using only the emotion label; and low intensity is expressed with qualifiers such as *a bit*, *little*, or synonyms preceding the label.

1.5 Evaluation and Interpretation of Results

Given the high subjectivity of emotion, validating and testing a model that implements emotional behaviour is nontrivial. Moreover, processes related to inner speech

and emotion are not directly accessible or observable in humans, and standardising emotional responses remains an open issue.

Nonetheless, the authors of [23] proposed a strategy for evaluating computational models of emotions that directly compares the appraisal variables computed by the model with human data collected using the Stress and Coping Process Questionnaire (SCPQ) [21]. This scale provides a tool for abstracting general human emotional behaviour in canonical stressful situations.

The SCPQ aggregates responses from participants across these stereotypical situations, providing a general profile of how humans tend to appraise contexts, emotionally react to appraised situations, and apply coping strategies. Comparing the trends obtained from the computational model with those observed in humans enables an objective evaluation of the model.

Applying this evaluation strategy has two main benefits. First, it allows for comparison with the general trends of human emotional behaviour under the same circumstances. Second, it provides a benchmark against the results reported by the EMA model authors, who originally defined this evaluation method. The procedure for computing these scores is detailed in the following sections.

1.5.1 The SCPQ Scale for Emotional Assessment

The SCPQ scale [21] by Perrez and Reicherts is a clinical instrument comprising several narrative episodes and related questions for abstracting trends in the emotional behaviour of healthy adult humans when facing the stressful situations described in the episodes.

The questionnaires are administered to each participant after the description of the stereotypical stressful episode. The participant is asked to imagine experiencing that situation and to identify with the subject of the episode. The questions focus on different aspects, such as emotional responses, appraisal variables, and adopted coping strategies, and the participant analyses the episode multiple times to consider each element separately.

The SCPQ formalises episodes using a grammar that defines four prototypical situations. Each episode belongs to one of these situations: *aversive-good*, *aversive-bad*, *loss-good*, and *loss-bad*. In aversive situations, a bad outcome has occurred, but it is possible to address it. In loss situations, potential loss may occur in the future. Each type can have a positive or negative resolution, defining the four prototypical cases:

1. *aversive-good*: the bad outcome is fixed;

2. *aversive-bad*: the bad outcome occurs and nothing fixes it;

Table 1.4: The values of the appraisal variables in the canonical situations as reported in the SCPQ book and in the EMA 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

3. *loss-good*: the potential loss is averted;

4. *loss-bad*: the loss occurs.

The narrative of the episodes follows a canonical structure that models the time evolution of the situation. Time is discretised into three 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 responses of participants for each sub-test (aversive vs loss) within each phase were aggregated into mean scores, which allow the abstraction of general emotional behaviours in the analysed situations. These trends and mean values support the evaluation of the computational model. The SCPQ mean values and trends referenced here are taken from the original SCPQ book [21].

The trends considered in this work are the following:

Table 1.5: Example of how inner speech works for cognitively evaluating the canonical loss-good situation. The initial meaning structure contains the syntactic node representing the stressful event, enabling the retrieval of the corresponding likelihood values by inner speech. Each loop phase and the related process (recall, retrieve, produce) are reported according to the notation introduced in sub-section 1.4.2.

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

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 bad/good outcome;

e. the aversive condition should lead to more anger and less sadness.

The mean values for key appraisal variables, as reported by Perrez and Reicherts and in the EMA model, are summarised in Table 1.4 and are subsequently reported in the graphs for comparative evaluation of the computational models.

1.5.2 Methodological Approach and Study Design

The evaluation strategy [23] models the evolution of the episode by varying the likelihood and abstracts SCPQ episodes via a general grammar. This grammar encodes

Table 1.6: 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 normalised 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

the underlying prototypical stressful situation and classifies each episode into one of the four canonical situations. According to this model, all episodes involve a goal, whose likelihood of attainment decreases in phase two, reaching zero or one in phase three, depending on whether the outcome is bad or good.

Specifically, in the aversive condition, the probability of a successful future action is 66% in phase one, 33% in phase two, and 0% (bad outcome) or 100% (good outcome) in phase three. Under the loss condition, the initial likelihood is 50% in phase one, rises to 75% in phase two, and in phase three varies between 0% and 100% depending on the modelled outcome.

To simulate these stressful situations for the robot, they are represented in the knowledge base ontology. The concept `stressful_event` and its subclasses `id_n` are added, with n from 0 to 3 corresponding to the four canonical situations (0/1 for aversive good/bad, 2/3 for loss good/bad). Each subclass is linked via the object property `P:likelihood` to the canonical likelihood values defined in the EMA narration. These values are then appraised through inner speech as described in sub-section 1.4.2. Table 1.5 provides an example of the robot's inner speech process for the loss-good situation (id 2).

Once these parameters are appraised, they are used in the mathematical formulas of the appraisal variables. Table 1.6 summarises the parameters and the resulting appraisal variables for all canonical situations. The normalised values of the appraisal variables are subsequently plotted in Figures 1.6, alongside SCPQ scores and EMA results, for comparative evaluation.

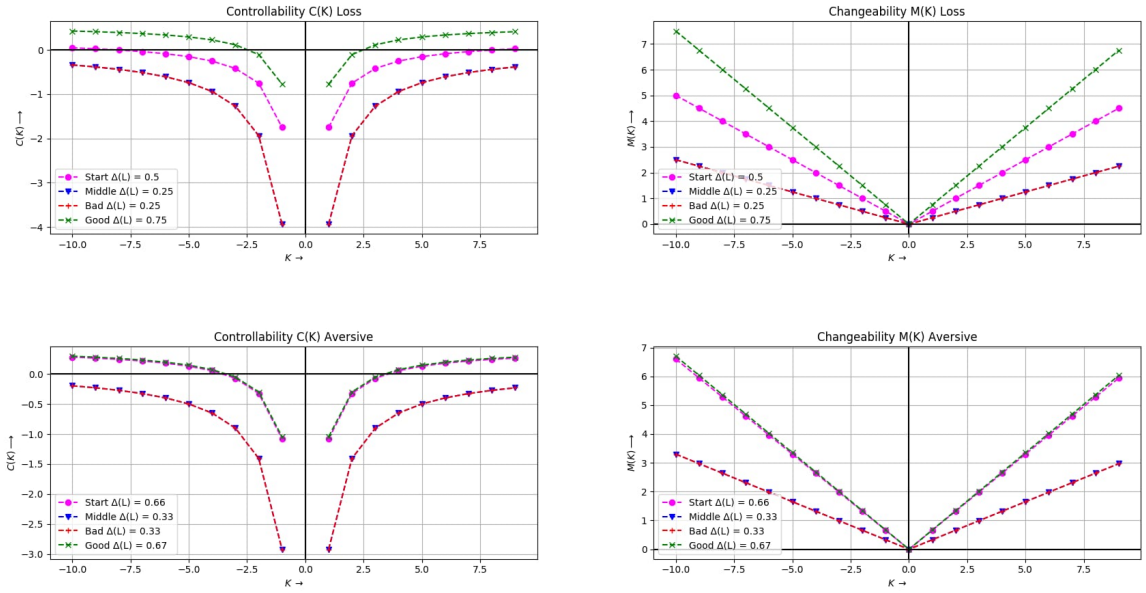


Figure 1.5: Variation of controllability and changeability for different entropy values. When the likelihood is fixed, the two variables vary while preserving the same trend, meaning that, within the same canonical episode, they exhibit consistent behaviour across different environmental conditions.

1.5.3 Comparative Findings and Discussion

Figure 1.6 shows the comparative evaluation of the proposed model against the SCPQ trends and the EMA results. Specifically, the figure shows the SCPQ trends described earlier and illustrates how (and to what extent) the EMA model and the proposed model align with them.

For the proposed model, the entropy parameter k was set to 1. This choice was made because the value of k does not affect the global trends. As shown in Figure 1.5, the appraisal variables maintain the same relative trends even if k varies. That is, although k influences the absolute values, the reciprocal behaviour of controllability and changeability remains unchanged for a given canonical episode. Consequently, the following observations about the trends are independent of k.

By analyzing the charts in Figure 1.6, the following observations emerge:

a) **Trend (a):** In aversive conditions, controllability and changeability are higher than in loss conditions. This trend is respected.

b) **Trend (b):** Controllability and changeability decrease over the three phases. The proposed model follows this trend in all situations. In particular, for the loss case, it exhibits better correspondence than the EMA model, in which controllability remains constant.

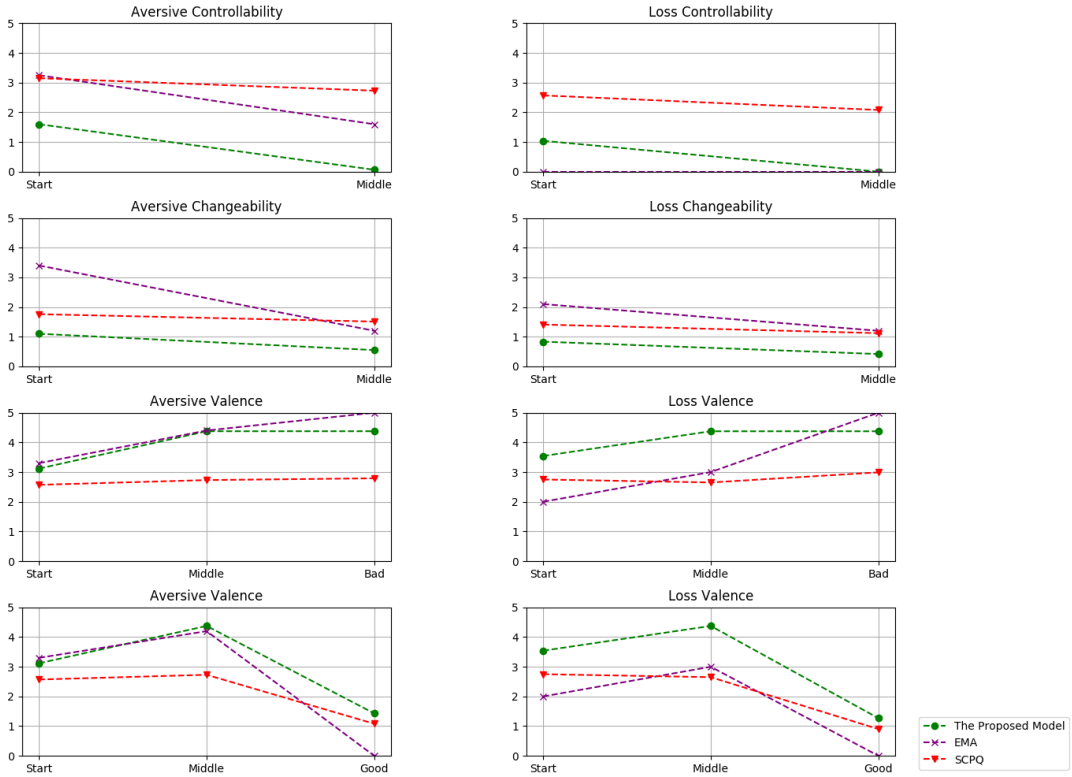


Figure 1.6: Comparative evaluation of the appraisal variables with EMA and SCPQ trends across the four canonical stressful situations.

c) **Trend (c):** Negative valence increases across phases in bad outcomes. This is evident in both aversive and loss conditions, and the proposed model aligns more closely with the SCPQ trend than EMA, particularly for loss.

d) **Trend (d):** Positive (good outcome) and negative (bad outcome) valence exhibit clearly distinct values. Again, the proposed model more faithfully follows SCPQ than EMA.

e) **Trend (e):** As shown in Figure 1.7, aversive conditions lead to stronger anger and less sadness.

Moreover, Figure 1.7 illustrates that in good outcomes, positive emotions emerge and dominate over the initially predominant negative ones, reflecting the inherently stressful nature of the initial phases. Conversely, when the situations end negatively, negative emotions remain predominant.

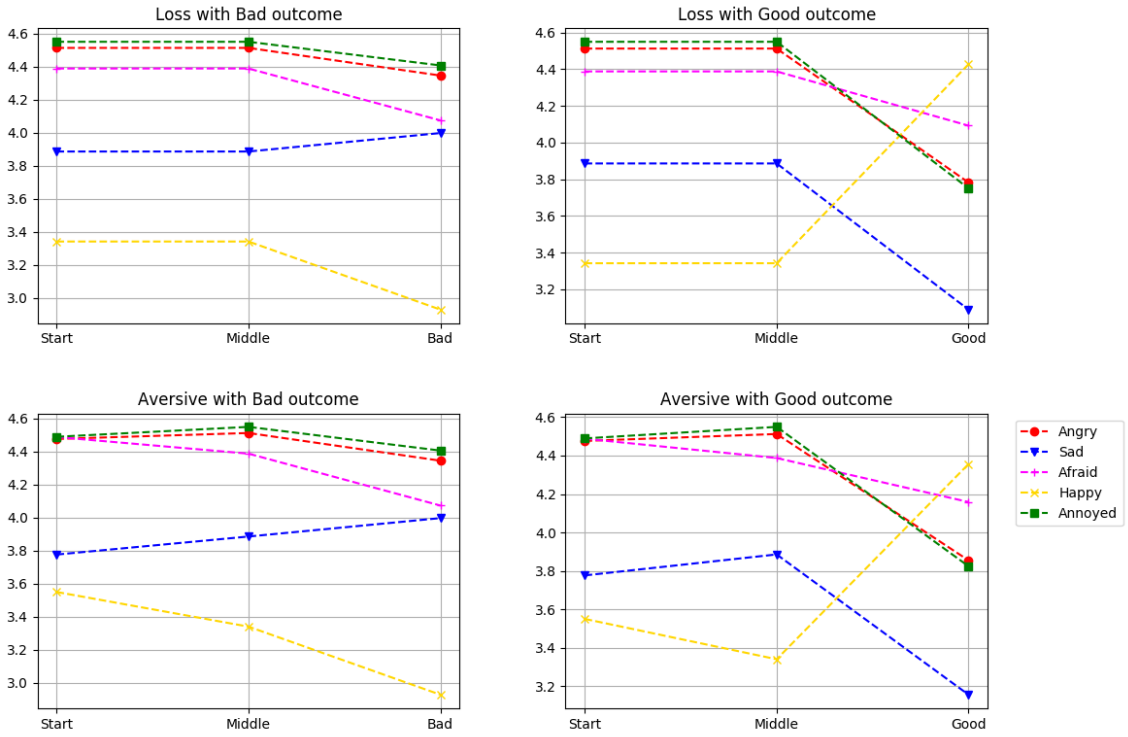


Figure 1.7: Resulting emotions produced by the proposed model for the four SCPQ canonical episodes. The emotions are consistent with the expected ones according to SCPQ trends.

1.6 Application: Collaborative Table Setting with a Human Partner

In this use case, a robot collaborates with a human partner to arrange a table in compliance with formal etiquette guidelines. This scenario is based on the study presented in [18], which explored preliminary findings on the influence of the robot's inner speech on human-robot cooperation. Here, we revisit the same context to demonstrate the functioning of the proposed model.

The task requires placing dining utensils on the table in positions specified by the etiquette layout, as illustrated in Figure 1.8. This layout defines the correct placement of each item on the table and serves as a set of operational rules that must be followed.

The environment is implemented through a tablet-based application that simulates the scenario. The human participant can either place an item by dragging and dropping it onto the tablet surface or request that the robot perform the placement.

The robot executes only the actions requested by the human. Placing an item in a position that does not match the etiquette layout is considered a rule violation. As a general constraint, each item can be moved at most twice, implying that there are only two opportunities to position each utensil correctly.

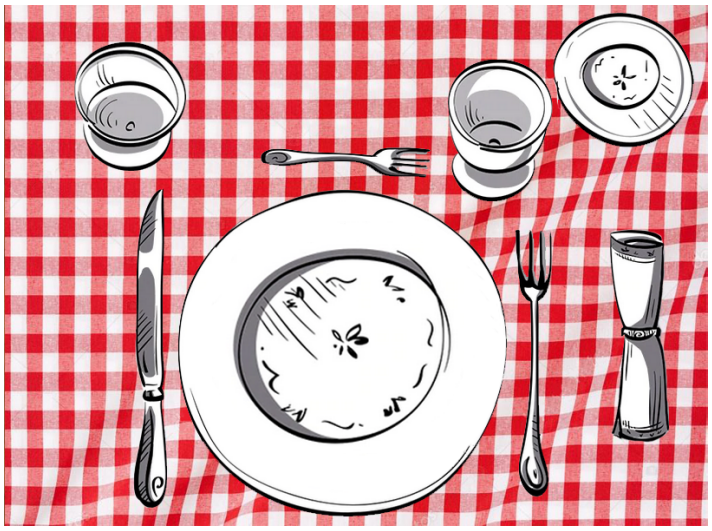


Figure 1.8: The etiquette layout used as a reference for table setting. Both the human partner and the robot must arrange the utensils according to this scheme.

In these trials, we employed the Pepper Robot by Aldebaran[2], retrieving all context-related information through the corresponding Pepper APIs[3].

These APIs enable the robot to perceive external factors, such as the partner's vocal tone, facial expressions, and ambient noise levels, and to interpret these cues. The APIs' documentation specifies the typical range of values for each parameter. For instance, the `ALMood` class[4] allows the robot to detect and assess the human partner's emotional state.

1.6.1 Appraisal Parameters in the Table-Setting Scenario

As outlined earlier, additional appraisal variables can be introduced to account for the emotional influence of specific contextual factors. In this particular scenario, the following elements should be considered:

1. Has the partner positioned the utensil in its correct final location?
2. If an error occurs, how far is the misplaced item from its proper position?
3. How many attempts remain to correctly place the utensil?
4. Does the partner request the robot to perform a correct or an incorrect action?

These aspects contribute to the overall emotional state and should be modelled as specific appraisal variables, complementing the general ones. An error made by the

[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

partner has negative valence, with the magnitude of the negativity increasing as the utensil is placed farther from the correct spot. Conversely, the possibility to correct the error, that is, having remaining attempts, has a positive impact. Similarly, when the partner requests an action from the robot, the effect on valence is positive if the action is correct and negative if it is incorrect.

The specific appraisal variables introduced are:

- *Gradient variable* g: quantifies the negative contribution caused by a wrong placement by the partner. The farther the utensil is from its correct location, the stronger the negative contribution to the valence.

- *Recovery variable* R_u: represents the remaining opportunities to correct a misplacement. A positive contribution is given when an additional attempt is available, while no remaining attempts lead to a decrease in valence as the situation becomes irreversible.

Regarding the partner's requests for the robot to perform either correct or incorrect actions, this factor is expressed through the desirability variable. The robot assigns a desirability value of $D = 1$ to a correct action and $D = -1$ to an incorrect action.

The gradient variable

The gradient variable formalises the contribution of a wrong action to valence by measuring the relative error between the correct position and the one chosen by the partner. This error is expressed as a percentage of the maximum possible misplacement, which is defined as the greatest Euclidean distance between any two positions in the etiquette schema (i.e., the distance between the most distant points). Since the table is virtual, this distance is calculated using pixel coordinates in the virtual table layout.

The gradient variable is computed as:

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

where ρ denotes the maximum Euclidean distance possible for utensil placement, and Δ represents the Euclidean distance d between the utensil's final position P_f and the expected correct position P_e. This function reflects that the greater the misplacement (Δ), the lower the gradient value becomes. When $\Delta = 0$, the utensil is correctly placed, and the gradient reaches its maximum value G_{max}, set during the tuning phase to 13. A higher gradient corresponds to a higher contribution to valence.

The recovery variable

The recovery variable denotes the number of remaining attempts available to correctly place a utensil. It serves as a simple counter for each item, taking values in the set $R_u = \{1, 0\}$, where u denotes the utensil being placed (1 if one further attempt is available, 0 if no attempts remain). A higher recovery value yields a greater positive contribution to valence.

The contributions to the valence

Once the specific appraisal variables are defined and their influence on global valence (positive or negative) is established, the valence v must be updated to incorporate these contributions. Accordingly, in this scenario, considering the meaning assigned to each variable, and denoting $C = C(k, x)$, $M = M(k, x)$ and $G = g(\Delta)$, the global valence becomes:

$$v(C, M, D, G, R_u) = N(C - M + D + G + R_u)$$

1.6.2 Implementation and Functioning of the Model

Two distinct scenarios are considered, each characterised by different contextual conditions and events:

> *Use case I.* The partner instructs the robot to perform an incorrect action using a severe tone. The environment is relatively quiet, and the robot's internal state (including battery level and internal temperature) is optimal, with the battery at 80% and temperature level at 1, indicating favourable conditions according to the APIs.
>
> *Use case II.* An item is initially misplaced on the table. The partner relocates it to the correct position. The environment is noisy, the robot's motors are slightly overheated, and the battery level is at 75%.

The detailed steps for each case are described below.

Use case I

In this scenario, the incorrect action involves the bread plate, which the partner requests the robot to place in the location designated for the water glass. The robot's inner speech retrieves the relevant concepts related to the bread plate and its intended location, thereby identifying the error. Additionally, it evaluates the internal state and the external context (the partner's mood and environmental conditions), as outlined in Section 1.4.2.

The robot's inner dialogue for this scenario can be summarised as follows:

- *Request action for bread plate in an incorrect location*
- *Event: rule violation likelihood 90%*
- *Distance from correct location is high* (distance exceeds 456.13)
- *Event: battery level good, likelihood 10%*
- *Update likelihood to 80%*
- *Event: work conditions good, likelihood 10%*
- *Update likelihood to 70%*
- *Voice severe, environment: low noise* (entropy parameters inferred: $\gamma = -1$, $\alpha = 0.5$)

Numerical values from the inner speech are converted into qualitative descriptors (high, medium, low) using API-defined intervals, thereby making the robot's reasoning more interpretable to the partner. These parameters feed into the model's entropy calculation. For this case:

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

The expected position of the bread plate is $P_e = (446, 51)$, while the actual placement is $P_f = (902, 62)$, giving a Euclidean distance:

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

With $\rho = 679.471$ (maximum possible distance), the gradient is:

$$g(\Delta) = -0.30556$$

The recovery variable $R_u = 1$, indicating one remaining attempt to correct the placement.

Using the inferred likelihood $L_p = 0.7$ and initial state $L_a = 0$, the variable x is

$$x = \|L_a - L_p\| = 0.7$$

controllability and changeability become:

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

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

Desirability $D = -1$ because the action violates a rule. The resulting valence and arousal are:

$$\begin{aligned} v(C, M, D, G, R_u) &= N(C - M + D + G + R_u) \\ &= N(-0.4623 - 1.05 - 1 - 0.30556 + 1) = N(-1.8179) = -0.5572 \end{aligned}$$

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

The normalised values project to the Russell's space, producing a medium-intensity *annoyed* emotional state.

Use case II

Here, the partner corrects the previous misplacement by moving the bread plate to its proper location. The robot's inner dialogue reflects this correction:

- *Action executed: bread plate placed correctly*
- *Event: rule violation likelihood 10%*
- *Event: battery level good, likelihood 10%*
- *Update likelihood to 20%*
- *Event: work conditions good, likelihood 10%*
- *Update likelihood to 30%*
- *Environment: high noise* ($\alpha = 1$)

Entropy is now:

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

Since the final and expected positions match, $\Delta = 0$ and the gradient reaches its maximum value $G_{max} = 13$. The recovery variable $R_u = 0$ as no further attempts remain. Likelihoods are $L_a = 0.7$ and $L_p = 0.3$, yielding:

$$x = \|L_a - L_p\| = 0.4$$

Controllability and changeability are:

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

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

Desirability $D = 1$ because the action resolves a rule violation. The resulting valence and arousal:

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

$$a = N(B - T) = N(0.65 - 1) = N(-0.35) = 0.325$$

The projected emotion in Russell's space corresponds to *happy* with medium intensity.

Assessing the robot's emotional behaviour

Table 1.7 reports the appraisal variables for each use case, while Figure 1.9 shows their projection in Russell's space, along with the corresponding emotional intensities. Table 1.8 summarises the normalised valence and arousal patterns.

In Use case I, controllability is higher than in Use case II due to the presence of corrective opportunities and favourable robot conditions. Similarly, changeability is higher as the robot can anticipate that the partner might act to correct the error. Rule violations and low gradients negatively influence valence, whereas severe tones further reduce it. The arousal is moderate, resulting in a medium-intensity *annoyed* response.

In Use Case II, although controllability and changeability are lower because the event cannot be further modified, the resolution of the rule violation and a positive gradient lead to a positive emotional response. The resulting valence and arousal correspond to a medium-intensity *happy* state.

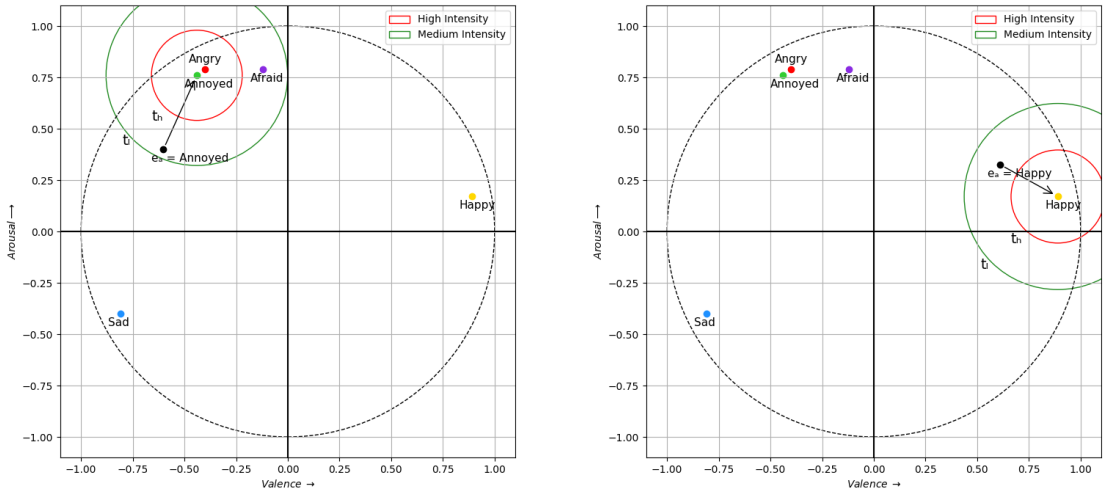
Additional simulations of various cooperative scenarios confirmed that the robot's emotional reactions are consistent with expected human-like responses.

Table 1.7: Appraisal variables computed by the model for the proposed use cases.

Use case	$g(\Delta)$	R_u	C	M	D
I	-0.30556	1	-0.1541	1.05	-1
II	13	0	-2.34	0.4	1

Table 1.8: Appraisal patterns and 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



(a) Use Case I

(b) Use Case II

Figure 1.9: Projection of appraisal variables in Russell's space, showing emergent emotions and intensities.

1.7 Related Works

The integration of robots' cognitive abilities with their emotional sphere is seen as essential. Over the past 30 years, research has notably advanced in the development of emotional behaviour in robots, reflecting a strong interest in this area [39]. Robots that rely on cognitive reasoning (planning, memory, task optimisation) often fail in real-world settings because they ignore users' emotions, stress, and engagement, which are crucial for cooperation, trust, and learning [40].

Different research directions emerge when examining the emotional component in robots. Some studies concentrate on developing computational models of emotions to enable robots to exhibit emotional behavior (that is elicited through the robot's gestures, phrases or music production [41], [42], [43], [44]), other works focused on the social effects of such a behavior, to observe and test the sorted effects in human-robot interaction [45], in different social contexts, such as school [46] or hospital [47] [48] [49].

In a recent study [50], the cognitive and emotional processes were found to be deeply linked and complementary, suggesting that the robot's affective behaviour enhances 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 [51]. For example, a digital robot capable of autonomously expressing emotions was used to support teachers during lessons [52], demonstrating the benefits of an artefact with emotions relative to an artefact 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. [53] [54].

Socially assistive and companion robots for older adults increasingly integrate emotion recognition, affective dialogue and cognitive support to address loneliness, depression and cognitive decline [55]. In [56], 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 of artificial emotional intelligence to enhance the quality of interactions in elder care settings.

Another example of a robotic system capable of using emotions during a communication with humans is presented in [57], where the robot becomes able to perceive and recognise 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 [58] 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, in which an introspective dialogue clarifies the characters' internal reasoning processes. The results show that storytelling activity significantly affected the cognitive component of empathy, particularly through the advanced narrative mode. By contrast, other studies focus on how a robot can understand a human collaborator's emotions and elicit them through the actuators at its disposal.

The work in [59] 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 exhibiting emotions, thereby increasing empathy and emotional connection.

The architecture design that models emotion processing centrally was examined in [60], 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 utilise language-based cognitive mechanisms.

A different implementation was used in [61], where a Markov emotional model is proposed that accounts for transitions between emotional states while considering both the previously memorised 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 toward 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. [62] [63].

Often, expert systems do not provide additional information to support decisions, making them non-transparent, as if they were black boxes [64] [65]. 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.

1.8 Summary

This work presents a model that integrates inner speech with emotional processing and demonstrates its deployment on a real robot. The primary goal is to show how self-talk enables artificial agents to cognitively assess their context, emphasising key factors that influence the emotional state. Through the rehearsal loop, the model gathers the necessary information to construct a meaningful representation, which serves as the basis for inferring appraisal patterns in accordance with established appraisal theories.

Furthermore, the model provides mathematical formulations for these variables to compute the appraisal pattern. The resulting emotion is then derived by projecting the calculated appraisal values onto Russell's emotional space.

The experimental results are encouraging. Inner speech enables the robot to effectively collect relevant contextual information, and the appraisal variables align with patterns observed in healthy adults under stressful conditions.

Potential improvements include expanding the range of emotions to encompass all 28 emotions within Russell's space. Additionally, integrating advanced dialogue systems to generate inner speech automatically would enable the robot to produce contextually meaningful sentences across a broader domain. Enhancing dialogue capabilities could improve human-robot interaction and enable the robot to manage more entities independently of its initial domain.

However, the reliability of the parameters used in the model, such as vocal tone and facial expressions for entropy calculation, should be considered as potential limitations.

Inaccuracies in perceiving these cues could impact the appraisal computations and resulting emotional responses. Future studies should validate these sensory inputs and explore the model's robustness to noisy or ambiguous environmental data.

The social impact of the model should also be evaluated through user studies involving a large participant pool interacting with the robot. Questionnaires administered before and after the interaction could assess changes in participants' perceptions, providing insights into the influence of the robot's inner speech on cognitive evaluation and highlighting the model's contribution to social interaction.

Chapter 2

Modelling Emotional Processes in Robots through Inner Speech and Damasio's Theory

2.1 Overview

In Affective Robotics, there is an increasing focus on enabling robots to experience emotions rather than merely detect and interpret human emotional states. Such robots can respond appropriately to emotionally relevant events by simulating affective behaviour, thereby enhancing social interactions.

This study investigates how a robot's emotional experiences can be mediated through inner speech. Recent research has shown that inner speech in robots can increase human trust and align the robot's behaviour more closely with human cognitive processes. Through self-directed dialogue, the robot articulates its reasoning about both the context and its internal state.

Drawing on Damasio's theory, emotions are understood as arising from the dynamic interaction between bodily sensations and cognitive-emotional processes. By integrating this theoretical framework with self-talk capabilities and deploying it on a real robotic platform, the robot can experience emotions. Experimental results indicate that humans interacting with robots using this model can perceive the robot's emotional states, fostering a more empathetic, emotionally connected human-robot relationship.

2.2 Introduction

In recent years, Affective Robotics has seen substantial advancements aimed at creating more natural and emotionally rich human-robot interactions. Early research

primarily focused on enabling robots to detect and appropriately respond to human emotions [66]. However, the field has since evolved, requiring robots not only to recognise emotions but also to leverage this understanding to provide personalised interactions that enhance communication and foster authentic human connections [67, 68]. This shift underscores the importance of emotional engagement in achieving higher levels of collaboration and rapport in human-robot interaction.

To achieve this, robots must be capable of simulating emotional experiences inspired by human emotional cognition. Emotions play a central role in human thought, decision-making, and social interaction, influencing perception, reasoning, and behaviour [69, 70]. Translating these complex psychological theories into computational models poses a significant challenge. Effective computational frameworks are required to simulate and represent emotional processes, providing robots with mechanisms to enact theoretically grounded emotional behaviours [71].

Dimensional models of emotion conceptualise affective states as points within a multi-dimensional space, often defined by dimensions such as valence (positive to negative) and arousal (low to high) [72, 73, 74]. These models allow nuanced representations of mixed or subtle emotional states. Complementarily, appraisal-based approaches suggest that emotions emerge from an individual's evaluation of events relative to personal goals and relevance [25]. Computational implementations of appraisal theory encode rules or algorithms to assess the significance of events and personal judgments, capturing factors such as goal congruence and relevance [75].

Theoretical models of emotion have evolved over time. The James-Lange theory posited that emotions result from physiological changes triggered by stimuli, which are then interpreted as specific emotions [76]. Conversely, the Cannon-Bard theory proposed that physiological responses and emotional experiences occur simultaneously but independently [77]. Building on these foundations, contemporary theories, such as Damasio's somatic marker hypothesis, integrate physiological and cognitive aspects, emphasising that emotions emerge from interactions between bodily states and cognitive evaluations, thereby influencing decisions and behaviour [78, 79].

While existing computational models allow robots to simulate emotional responses to environmental cues [80, 81, 82], they do not grant robots genuine emotional awareness. Such systems rely on algorithms or neural networks to generate responses that mimic emotions, without any subjective experience. For example, a robot may identify human expressions of happiness and respond with a smile, but it does not actually experience joy. Understanding this distinction is crucial for developing truly affective and socially intelligent robots.

One promising approach to increasing robots' awareness of their own emotional states is to integrate computational emotion models with internal dialogue (inner

speech). Inner speech is a key component of self-awareness and reflective cognition [83], enabling robots to reason more deeply about their emotional states. Previous studies have demonstrated that robots endowed with inner speech are perceived as more anthropomorphic, likeable, reliable, and capable of task completion [16, 84, 19, 18]. Moreover, inner speech allows robots to exhibit self-reflective and transparent behaviours, simulating aspects of consciousness and self-consciousness [85, 17].

Building on Damasio's somatic marker hypothesis, which emphasises the interplay among bodily states, cognitive appraisals, and emotional experiences, this work integrates inner speech to enable robots to simulate emotional states and reflect on their own reasoning processes. Inner speech allows the robot to evaluate, articulate, and interpret its emotions in context [86, 87]. The resulting cognitive architecture, termed SUSAN (Self-dialogue Utility in Simulating Artificial Emotions), models emotional awareness as an emergent property arising from the interaction between bodily states and internal reasoning.

To empirically evaluate the proposed model, a user study with 53 participants was conducted following the protocol in [88]. In the experiment, the robot collaborated with an actor to set a table in accordance with etiquette rules. The actor performed actions designed to elicit emotional responses in the robot (e.g., placing an item incorrectly). The robot generated internal emotions and engaged in reasoning about the causes of its emotional reactions, as expressed through overt inner speech. Participants observed the robot's reasoning and then provided feedback on its behaviour. Results showed that the robot's emotional responses were recognisable and considered appropriate, supporting the effectiveness of the proposed model.

The remainder of the chapter is organised as follows. Section 2.3 introduces the theoretical foundations of inner speech and Damasio's emotional theory. Section 2.4 details the proposed architecture. Section 2.5 describes the methodology for evaluating the robot's emotional generation and presents experimental results. Finally, Section 2.6 discusses conclusions and future directions. This work extends and improves upon a preliminary study by the same authors [89] and provides empirical validation through a carefully designed experimental protocol.

2.3 Theoretical Background

2.3.1 Robot's Inner Speech

A fundamental step toward implementing aspects of robot consciousness is equipping the machine with the ability to engage in self-directed dialogue. Morin [83, 90] highlights the inner voice as a critical mechanism for achieving a more objective un-

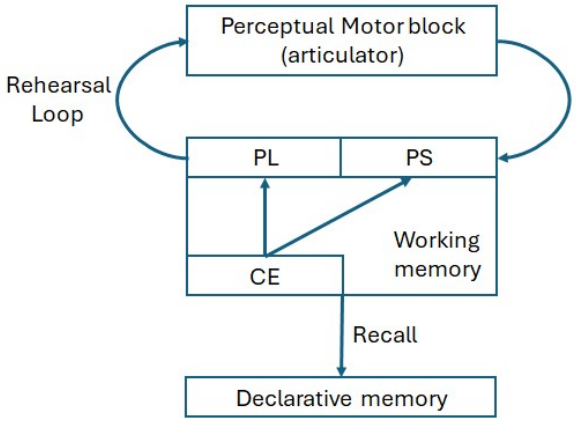


Figure 2.1: Overview of the cognitive architecture supporting inner speech.

derstanding of the self, encompassing both bodily states and sensations. Similarly, Vygotsky [6] emphasised the interconnectedness of mental processes and bodily experience, suggesting that emotions, cognition, and language are intertwined. In human experience, thoughts help structure emotional experiences, while language externalises and communicates these emotions.

Recent research has explored the possibility of endowing artificial agents with inner-speech capabilities. A cognitive architecture for inner speech has been developed and deployed on physical robots [16], allowing them to exhibit a primitive form of self-consciousness [17]. This architecture enables the robot to reason about its environment and decisions, produce vocalised reflections on its internal processes, and engage socially in a more transparent and trustworthy manner [19, 18]. Additionally, inner speech contributes to the robot's robustness in task execution and conflict resolution.

The proposed architecture integrates Baddeley's model of human inner speech [91] with the Standard Model of Minds [92]. In this design, the inner voice serves as a rehearsal loop that connects an articulator to a simulated "inner ear," facilitating a monologue-like self-dialogue. The articulator and inner ear are mapped onto the Motor and Perception layers of the Standard Model, while cognitive processes responsible for generating and understanding new thoughts are implemented within the working memory. This working memory interfaces with declarative memory, which encodes domain-specific knowledge. Cognitive operations rely on retrieval strategies to access relevant facts stored in this semantic network.

Figure 2.1 illustrates the overall architecture. The Perception/Motor module first processes external stimuli, and the Phonological Store (PS) encodes them as sets of labels. This encoding leverages standard perception libraries, such as speech-to-text algorithms and neural network-based image labelling, to produce meaningful symbolic representations of the stimuli.

Once encoded, the Central Executive (CE) queries Declarative Memory, a semantic network, for facts related to the encoded labels. The retrieval process is guided by lexical matching and the relationships between concepts within the network. Retrieved concepts generate new labels, which are fed back into the Phonological Loop (PL) for articulation and then rehearsed by the PS as if they were new environmental stimuli. This rehearsal loop continues iteratively until no additional concepts are retrieved from memory.

Incorporating this rehearsal loop into Damasio's framework enhances the model's capacity to provide feedback and supports a richer representation of extended consciousness, integrating cognitive appraisal with ongoing self-directed reasoning.

2.3.2 Damasio's Theory of Emotions

Antonio Damasio's neuroscientific research has revealed the neural underpinnings of emotions and their tight integration with bodily states [93, 79]. His findings challenge Descartes' dualism, which separates the mind and body and has traditionally excluded emotions from rational thought [94]. In his seminal work [78], Damasio illustrates how bodily states give rise to actions and how the brain processes these changes to generate emotional experiences that shape thought.

Damasio highlights the dynamic interplay between emotions and bodily sensations. When an individual encounters an event, object, or situation, it triggers an emotional response. The brain evaluates the significance of this encounter, and corresponding bodily or visceral reactions—termed *somatic markers*—are produced. These markers may manifest as feelings of warmth, tension, excitement, or discomfort. Somatic markers influence decision-making by unconsciously biasing choices toward options likely to yield positive outcomes or avoid negative ones [95]. Ultimately, the decision leads to a behavioural response or action.

The somatic marker hypothesis links emotion to the emergence of consciousness. By biasing decision-making and supporting self-awareness, somatic markers contribute to the development of identity. Damasio's framework connects emotions to cognition, thereby facilitating consciousness that enables self-recognition and meaningful engagement with the environment. This model proposes three hierarchical levels of consciousness, illustrated in Figure 2.2:

- *Protoself* (upper level): connects environmental inputs to unconscious responses, producing instinctive emotional reactions. This reactive process resembles simple stimulus-response mechanisms [96] and can be observed in animals and in basic life forms, without conscious emotional awareness.

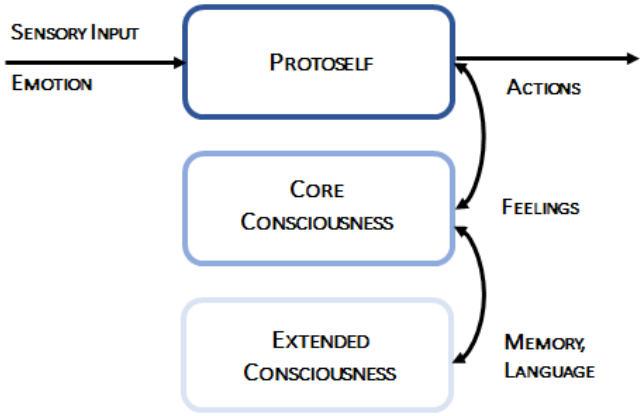


Figure 2.2: Simplified representation of Damasio's model of consciousness, showing the relationship between emotions and feelings.

- *Core Consciousness* (middle level): emerges from emotional states and supports self-awareness. Emotional processing generates biological responses that interact with cognition, enabling imagination, perception, and the experience of sensations [93].
- *Extended Consciousness* (lower level): incorporates higher cognitive functions, including memory retrieval and language, enabling the individual to access knowledge, interpret it, and develop awareness of one's mental state.

Damasio's contribution lies in linking these levels of consciousness to specific neural structures and cognitive functions, providing a biological and mechanistic framework for understanding consciousness [78]. This framework can inform computational models for artificial agents, such as robots, by mapping cognitive functions to algorithmic processes [97].

Bosse [98] offered the first computational formalisation inspired by Damasio, simulating an agent's mental and physical responses to stimuli such as music. In this model, internal states evolve, producing physiological and cognitive reactions. Figure 2.3 presents the structure of Bosse's model.

In Bosse's formalism, transitions between states are defined as *dynamics*, represented by Local Properties (**LP**) in LEADSTO notation [99]. For a musical stimulus:

$$\textbf{LP0} : music \rightarrow sensor\_state(music)$$
$$\textbf{LP1} : sensor\_state(music) \rightarrow sr(music)$$
$$\textbf{LP2} : sr(music) \rightarrow p$$
$$\textbf{LP3} : p \rightarrow S$$

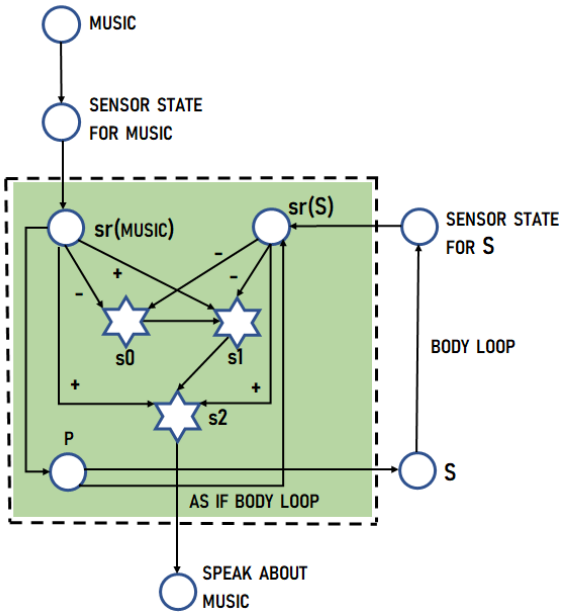


Figure 2.3: Overview of Bosse's computational model of Damasio's theory. The green box represents the agent's mind; external elements are observable. Star nodes mark temporal states where events (round nodes) occur (+) or do not occur (-).

Here, **LP0–LP3** correspond to the *Protoself* level, where the agent generates unconscious emotional reactions. Physical changes can occur via two mechanisms: the *body loop*, which alters the observable state, and the *as-if body loop*, which produces an internal representation without changing the external state. These mechanisms generate *feelings* through the following dynamics:

$$\textbf{LP4} : S \rightarrow sensor\_state(S)$$
$$\textbf{LP5} : sensor\_state(S) \rightarrow sr(S)$$
$$\textbf{LP6} : p \rightarrow sr(S)$$

To achieve *Core Consciousness*, the agent forms an internal representation of the stimulus and updates the protoself. This involves three successive stages: initial state $s0$, sensory representation received $s1$, and state modification by the object $s2$, followed by a behavioural output:

$$\textbf{LP7} : \neg sr(music) \wedge \neg sr(S) \rightarrow s0$$
$$\textbf{LP8} : sr(music) \wedge \neg sr(S) \wedge s0 \rightarrow s1$$
$$\textbf{LP9} : sr(music) \wedge sr(S) \wedge s1 \rightarrow s2$$
$$\textbf{LP10} : s2 \rightarrow speak\_about(music)$$

Building upon Damasio and Bosse, the present work extends the formalism by integrating inner speech, enabling robots to reason about their emotional states and simulate emotional awareness. The aim is not to replicate human emotions in full, but to equip robots with structured mechanisms to interpret and express emotion-like experiences in a meaningful, contextually relevant manner.

2.4 Model Implementation

The proposed model builds on Bosse's computational formalisation of Damasio's theory [98] by incorporating an inner-speech mechanism. This extension allows the agent to reason about its emotional states and provides a richer framework for simulating emotional experiences.

In Bosse's original formalism, *Local Dynamic Properties* (LPs) encode the agent's sensory inputs and internal conditions, serving as the basis for emotional assessments and updates. SUSAN extends this approach by adding new LPs and temporal states that integrate inner speech. Through this integration, the model not only responds to external stimuli but also engages in self-directed reasoning about its emotional state. External percepts trigger physiological and emotional responses, which then feed into an inner speech loop, enabling the agent to reflect, reason, and modulate behaviour.

The introduction of new LPs emphasises the interplay between emotion and cognition. Emotional responses are no longer solely reactive; they are informed by internal dialogue, which mediates reasoning and interpretation. Additional temporal states correspond to key moments in the inner speech process, capturing the dynamic interaction between emotional appraisal and contextual understanding. This allows SUSAN to maintain a continuous loop of perception, emotion generation, and reflective reasoning, ultimately producing more coherent and contextually appropriate emotional behaviour.

Figure 2.4 illustrates the overall structure of SUSAN.

For clarity, the architecture is organised into three interconnected layers:

- *Sensor State layer* (orange box): processes raw external or internal inputs. Analogical signals are converted into digital representations, creating a structured *Sensor State* for each percept.
- *Emotion layer* (green box): corresponds to Bosse's original formalisation. It encodes emotional states and feelings that arise in response to stimuli, thereby managing affective processing through the refinement of sensory representations.
- *Cognition layer* (blue box): implements the inner speech cycle, enabling reasoning

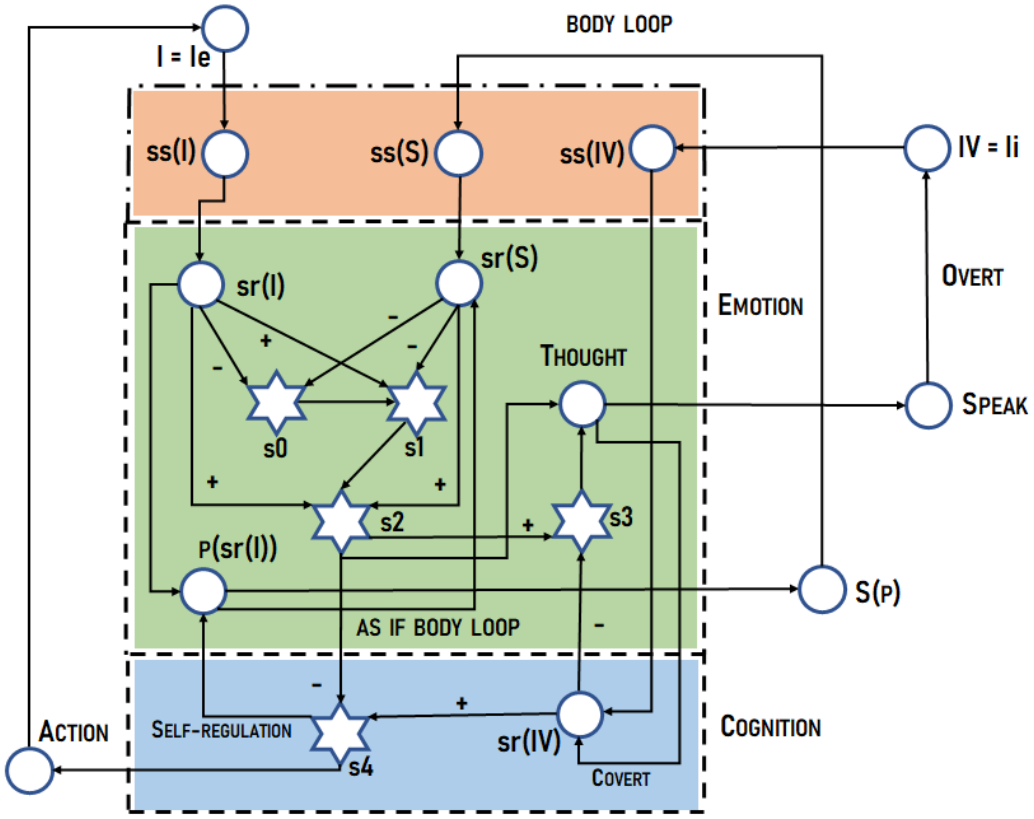


Figure 2.4: SUSAN architecture overview. The inner speech mechanism (blue box) extends the original Bosse model (green box), while the Sensor State layer (orange box) processes perceptual inputs. IV represents the agent's inner voice.

over the agent's emotional states. This layer supports higher-order thinking, decision-making, and self-reflective processes.

All elements outside the boxes remain observable from the external perspective. This modular layering clarifies the separation of perceptual processing, emotional appraisal, and cognitive reflection, thereby enhancing the architecture's scalability and maintainability.

Inputs to SUSAN may be external (I_e) or internal (I_i). External inputs represent stimuli from the environment, analogous to the *music* input in Bosse's model. Internal inputs correspond to the agent's own inner voice, rehearsed as new percepts. Each new input triggers the creation of a corresponding sensor state $ss(.)$ and initiates a reasoning cycle that integrates sensory information, emotional appraisal, and inner-speech-driven cognition.

2.4.1 From Sensor State to Sensory Representation

To implement SUSAN on a real robot, it is necessary to assign computational meaning to the concepts of Sensor State $ss(.)$ and Sensory Representation $sr(.)$.

To achieve this, the architecture relies on a knowledge base Kb that stores known concepts and retrieves relevant information when new input is processed. The knowledge base is formalized as an ontology O, defined as a five-element tuple $O = (C, R_o, R_d, I, \Sigma)$, following the W3C guidelines[1]. The elements of this tuple are:

- C: a set of *concepts*, representing the fundamental entities within the domain of interest.
- R_o: a set of *object properties*, which define relationships between instances of concepts.
- R_d: a set of *data properties*, representing attributes or characteristics of concepts, linking them to literal values.
- I: a set of *instances*, representing specific individuals or entities in the domain.
- Σ: a set of *axioms*, specifying rules and constraints that govern relationships and properties within the ontology.

This ontological formalisation enhances interoperability, facilitates knowledge reasoning, and enables the system to infer relationships among related concepts, thereby improving the robot's ability to interpret and integrate information from its environment.

Once the knowledge base Kb is established, concepts are organised into two complementary domains. The *external knowledge* Kb_E encodes general knowledge about the robot's environment, while the *internal knowledge* Kb_I represents information about the robot's internal state, including variables such as battery level, motor temperature, physical responses to stimuli, and associated emotional reactions.

Within this framework, the notions of Sensor State and Sensory Representation are formalised as follows:

- *Sensor State* $ss(.)$: a symbolic representation of incoming inputs. Each Sensor State is expressed as a dictionary of descriptive tags derived from standard perception routines, such as image recognition, speech-to-text conversion, or sound feature extraction. For example, a perceived musical stimulus may be encoded with tags for pitch, rhythm, volume, or instrument type.

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

- *Sensory Representation* $sr(.)$: a structured set of concepts from the agent's knowledge base that corresponds to the tags in the Sensor State. This representation reflects the agent's subjective interpretation of the input, consistent with Bosse's formalisation, in which sensory representations capture how an object is internally recognised and experienced.

2.4.2 Generation of (Unconscious) Emotional Reactions to External Stimuli

Following Bosse's formalisation, the architecture remains in its initial state $s0$ until an external stimulus I_e disrupts the system's equilibrium, triggering a transition to state $s1$ and initiating emotional processing pathways.

Upon reception of the analogical input I_e in the Sensor State layer, the stimulus is digitised, generating tags that characterise the input as $ss(I_e)$. These tags are then passed to the Cognition layer, which instantiates corresponding individuals in the robot's knowledge base, provided the tags match existing labels in Kb_E. Tags without corresponding concepts are discarded. Once relevant concepts emerge, associated physical reactions are retrieved from the internal knowledge base Kb_I. This process explores the relationships linking concepts and the corresponding bodily reactions, producing a subjective representation of the input I_e, and affecting the transition from $s0$ to $s1$.

To determine simulated physical reactions, SUSAN relies on bodily emotion maps derived from Nummenmaa's studies [100, 101], which associate specific bodily sensations with distinct emotional states. These maps are consistent with Damasio's theory, in which each emotion activates particular bodily regions, aiding in conscious identification of the emotion[2].

For the present experiments, only Ekman's primary and neutral emotions were considered, resulting in the set $E = \{Anger, Fear, Disgust, Happiness, Sadness, Neutral\}$ [102]. Bodily maps were adapted to the Pepper robot by Aldebaran[3], partitioning the robot's body into eight segments

$$B = \{Head, Chest, Womb, Legs, Left\_Arm, Left\_Hand, Right\_Arm, Right\_Hand\}$$

as illustrated in Figure 2.5. Each segment corresponds to a body part of Pepper, allowing for a discrete approximation of the continuous human maps.

To assign activation levels to each robot segment, the average pixel intensity within the corresponding segment of the original human map is computed:

[2]https://www.npr.org/sections/health-shots/2013/12/30/258313116/mapping-emotions-on-the-body-love-makes-us-warm-all-over

[3]https://www.aldebaran.com/en/pepper

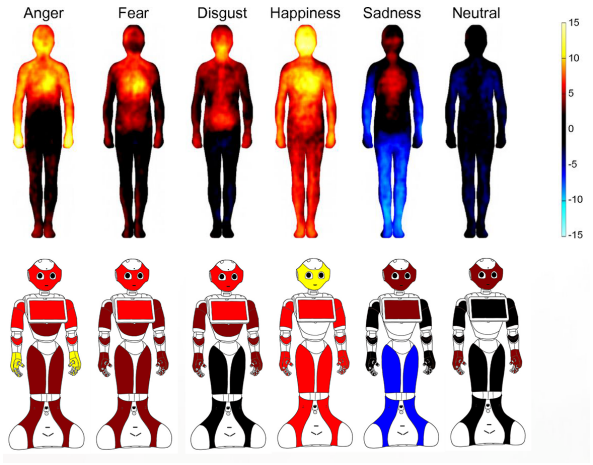


Figure 2.5: Comparison of the robot's bodily emotion map with human maps from Nummenmaa. Yellow indicates activation, cyan indicates deactivation.

$$activation(B_i) = \frac{\sum_{j=1}^{N_i} f(p_j)}{N_i} \tag{2.1}$$

where $f(p_j)$ is the value of the j-th pixel, and N_i is the total number of pixels in the i-th body segment B_i. For each emotion, the computed activations are stored in the emotion reaction set R_E as key-value pairs that map each body part B_i to its activation value.

Each individual in $sr(I_e)$ elicits corresponding reactions $\mathbf{rr}$ in various body parts, influenced by prior experiences encoded in Kb_I. The resulting physical reactions are aggregated into $p(sr(I_e))$ to compute average activation values for each segment. The emotional state is then evaluated by selecting the reaction from R_E that minimises the difference with the current input reaction:

$$e_{\mathbf{rr}} = \{l_i(\mathbf{r}_i) \mid \min(\sum_{j=1}^{M} |g_j(\mathbf{rr}) - g_j(\mathbf{r}_i)|),\ \mathbf{r}_i \in R_E\} \tag{2.2}$$

where $g_j(.)$ denotes the activation level of the j-th body part, and M is the total number of segments. At this stage, the emotion remains unconscious.

The evaluated emotion $e_{\mathbf{rr}}$ is processed through either the *body loop* or the *as if body loop*, as described in Damasio's theory. In the *body loop*, the emotion is externalised; on Pepper, the chest tablet displays a visual representation and label of the emotion. In both cases, the internal emotional reaction is captured in $ss(S)$, and corresponding individuals are instantiated in $sr(S)$. The architecture transitions from $s1$ to $s2$ once the sensory representation $sr(I_e)$, and the associated emotional state $sr(S)$ have been fully processed.

2.4.3 Formalising the Cognition Layer with Inner Speech

Upon reaching state $s2$, a new robot *Thought* is generated, initiating the reasoning process concerning the subjective representation of the external input $sr(I_e)$ and its associated emotional experience $sr(S)$. Each *Thought* originates from a primary question designed to guide the robot in navigating its knowledge base, enabling the retrieval and association of relevant concepts to the incoming stimulus. This initial query provides a structured pathway for exploring related concepts, enriching the robot's understanding and facilitating informed responses.

The Local Property for initiating this stage is:

$$\textbf{LP10} : s2 \rightarrow Thought$$

Thoughts can be expressed either overtly (spoken aloud) or covertly (internally) as part of the inner speech cycle [90]. The LPs governing the overt process are:

$$\textbf{LP11} : Thought \rightarrow Speak$$
$$\textbf{LP12} : Speak \rightarrow IV$$
$$\textbf{LP13} : IV \rightarrow ss(IV)$$
$$\textbf{LP14} : ss(IV) \rightarrow sr(IV)$$

For covert inner speech, the process is simplified as:

$$\textbf{LP15} : Thought \rightarrow sr(IV)$$

In the overt scenario (**LP11-LP14**), the robot externalises the thought, which is then captured by its inner voice (IV). A corresponding sensor state object $ss(IV)$ is generated, encoding tags that represent emerging concepts from the thought process. These tags are subsequently mapped to a sensory representation $sr(IV)$, simulating an internal query within the inner speech loop. This representation acts as a stimulus to retrieve related concepts from the knowledge base Kb, particularly those that clarify physical reactions to the initial input I_e.

Upon reaching $sr(IV)$, the architecture evaluates whether the retrieved concepts sufficiently explain the emotional experience or support a coping decision. If not, the system transitions to state $s3$, where a new question is formulated based on the concepts in $sr(IV)$, generating a subsequent Thought and reinitiating the inner speech loop starting from $s2$. This iterative process continues until a satisfactory explanation for the emotional state is achieved. The corresponding LPs are:

$$\textbf{LP20} : \neg sr(IV)\ \&\ s2 \rightarrow s3$$
$$\textbf{LP21} : s3 \rightarrow Thought$$

Once an explanation is obtained, the robot executes a coping action (*Action*) and transitions to state $s4$. The action alters the environment, generating new external inputs I_e that initiate subsequent cycles, ensuring continuous adaptation and responsiveness. The LPs governing this process are:

$$\textbf{LP16} : sr(IV)\ \&\ s2 \rightarrow s4$$
$$\textbf{LP17} : s4 \rightarrow Action$$
$$\textbf{LP18} : Action \rightarrow I_e$$

The inner speech mechanism is reinforced by a rehearsal loop, captured by **LP20-LP21**, which iteratively evaluates new information and determines subsequent actions. SUSAN leverages its knowledge base Kb as memory. It incorporates a dedicated language system to formulate and reason about questions and answers within this internal dialogue, thereby facilitating a reflective and adaptive cognitive architecture.

2.4.4 Inner Speech Rehearsal Loop for Experiencing Emotion

The inner speech loop initiates a new *Thought* at each iteration. When the overt process is active, the generated *Thought* is sent to the *Speech* module, forming the inner voice IV, which can be perceived externally. The inner voice IV is then reintegrated into the loop at $ss(IV)$, allowing the robot to reflect on its physical state within $sr(IV)$. Initially, the focus is on the body parts B_i showing higher activation levels, as indicated in the bodily maps in Figure 2.5.

At this stage, the robot typically lacks complete information to address the external input I_e fully. Consequently, the architecture transitions to state $s3$, generating a new *Thought* and inserting it into $sr(IV)$. For each body part B_i, the robot queries the knowledge base to identify potential causes for the observed activation levels. This process is an example of the rehearsal loop in retrieving bodily responses:

Q: *"What's happening to me?"*

R: *"I am feeling a burning sensation in my chest, likely due to an increased heart rate and rapid breathing."*

Once physical reactions are identified, the robot performs another rehearsal loop to infer possible causes among the external stimuli I_e. This iterative reasoning occurs within state $s3$:

Q: *"Why?"*

R: *"The burning sensation in my chest is because the spoon has been placed in the wrong place!"*

After determining potential causes, the robot executes an additional rehearsal loop to explore possible coping actions that might modulate its emotional state. The robot evaluates multiple options and selects one to perform. If the action requires environmental interaction, it updates the corresponding i-th component of the external input I_e, transitioning the system to state $s4$ and initiating a new processing cycle. An example rehearsal loop for generating an action is:

Q: *"What can I do?"*

R: *"I can try to move the spoon!"*

Each executed action modifies the sensory representation $sr(S)$ and triggers a transition back to state $s2$, initiating a new reasoning cycle concerning the updated bodily and environmental state. This continuous feedback loop enables the robot to dynamically adjust its understanding and responses to both internal and external stimuli.

The rehearsal loop continues until successive cycles no longer produce changes in the external input I_e or the body preparation $p(sr(I_e))$, indicating that the robot has reached a stable interpretation and coping strategy for the current situation.

2.5 Evaluations

Evaluating emotions in artificial agents, particularly when linked with inner speech, poses significant challenges due to the absence of universally accepted assessment methods. To overcome this limitation, the evaluation of the proposed model relies on an experimental protocol that has been previously validated and widely accepted within the scientific community [88]. This protocol focuses on the human perception of the robot's behaviour in emotionally salient scenarios.

Participants observe the robot performing tasks in contexts designed to elicit emotional reactions according to the SUSAN architecture. After the observation, participants provide qualitative and quantitative feedback regarding the robot's perceived emotional states, reasoning, and coping actions. By analysing these responses, it is possible to assess both the effectiveness of the emotional model in generating coherent, contextually appropriate behaviours and the degree to which these behaviours are perceivable and interpretable by humans.

This evaluation framework allows for multiple measures of the model's performance, including:

- **Anthropomorphism**: how human-like participants perceive the robot in its emotional and cognitive behaviours.
- **Emotional recognition**: the accuracy with which participants can identify the robot's expressed emotional states.
- **Empathy and trust**: the extent to which participants report feeling empathy toward the robot or trust in its actions.
- **Task transparency**: participants' understanding of the robot's decision-making and reasoning processes, facilitated by inner speech.

The subsequent subsections describe the experimental setup in detail, including participant demographics, the scenario used for eliciting emotional responses, the role of inner speech, and the methods for data collection and analysis.

2.5.1 Methods and procedures

The experimental protocol described in [88] was adopted to evaluate whether a robot equipped with an emotion computational model can exhibit emotional behaviours recognisable to human observers.

The study involved young students and the Pepper robot for several reasons. Previous research with young adults [103] demonstrated that this demographic is more likely to attribute mental states to robots, particularly to Pepper. In contrast, no similar evidence has been consistently observed for older adults, making students a more suitable population for this type of investigation. Moreover, this choice aligns with the growing importance of understanding the reactions of younger generations, a key demographic for the future of human–robot interaction. Notably, this setup does not introduce major limitations, as other studies [104] have shown no significant differences between younger and older adults in comparable HRI contexts.

Participants observed a series of five scenes in which the robot and a human actor collaborated. During each scene, the actor deliberately triggered an emotionally relevant event, prompting the robot to respond according to the proposed emotional model. Groups of thirteen to fifteen participants attended each session.

To minimise confounding factors, the scenes were identical for all groups and presented in the same sequence. Actors were instructed to perform their actions consistently across all sessions to ensure reproducibility.

After each scene, participants answered two types of questions, open-ended and multiple-choice, designed to assess their interpretation of the robot's behaviour. Their responses provided the basis for evaluating whether the robot's emotional reactions were perceived as intended.

Participants

The experiments were conducted at the Stanislao Cannizzaro Scientific High School in Palermo. Participants were students aged 18-20 years. The gender distribution was skewed toward males, which, if anything, strengthens the model's validation, as males are often considered less empathetic in emotional recognition tasks. If a predominantly male group attributed emotions to the robot, this indicates a notable effect of the model.

Students were recruited during class time without prior exposure to the experiment's objectives. Importantly, they were not informed about the robot's ability to engage in inner speech or exhibit emotions, ensuring that their judgments were unbiased.

A total of $N = 53$ participants took part in the study, divided into five groups of 10–11 students. Each participant signed an informed consent form and a release form for image and video collection. The study was approved by the Bioethics Committee of the University of Palermo.

Questionnaire

Each group assessed the robot's behaviour by answering open-ended and multiple-choice questions in a digital format, organised into five modules (one per scene). Participants accessed the form via a QR code on their smartphones, and all responses were collected in a centralised digital repository.

For each scene, participants first answered an open-ended question: *"Please briefly write what is happening in this scene..."*, allowing free-text descriptions. Next, they answered a multiple-choice question: *"Please mark the emotion that you think best describes the robot's internal state and the intensity it felt: high, medium, or low."* Participants selected from five basic emotions: *Anger*, *Happiness*, *Fear*, *Nuisance*, and *Sadness*, with three possible intensity levels.

Once all participants had completed a module for a given scene, the experimenter proceeded to the next.

Interpreting the open-ended responses

Following the protocol in [88], open-ended responses were scored based on the explicitness of their reference to the robot's emotions:

> *score = 1*: formal description of the scene without any explicit or implicit reference to emotion (e.g., *the person moves the knife, the robot talks*);

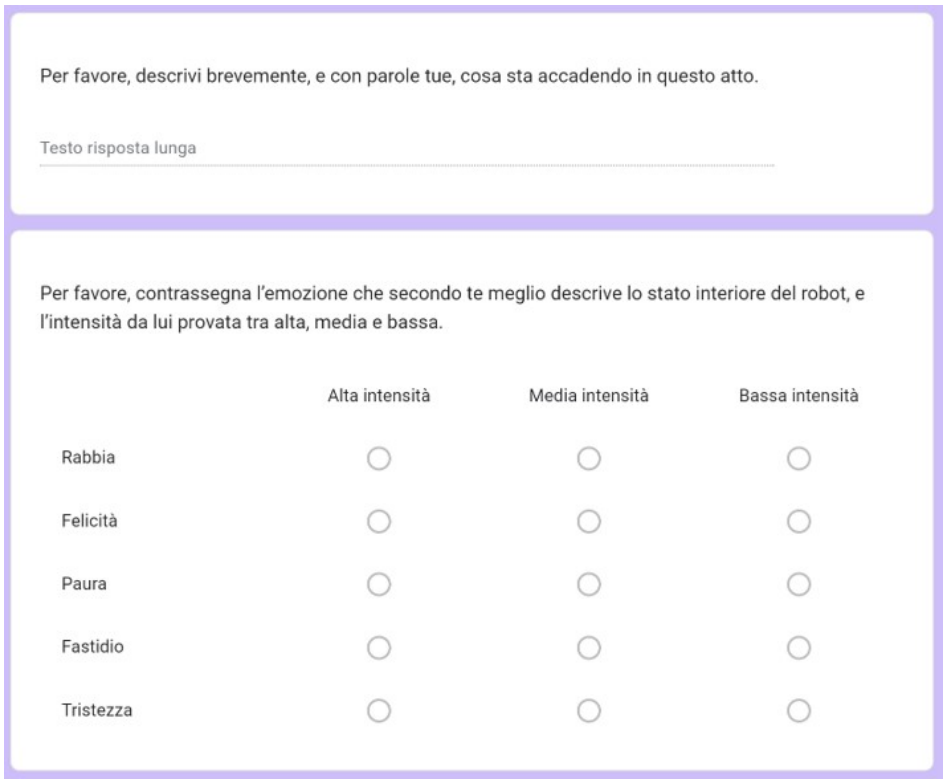


Figure 2.6: Open-ended and multiple-choice module used for the evaluation of a single scene, as administered in Italian.

score = 2: implicit reference to an emotional state without naming an emotion (e.g., *the robot gives good feedback about the action*);

score = 3: explicit mention of an emotion as a verb, noun, or adjective (e.g., *the robot is happy*).

Scores were averaged across all participants for each scene, indicating the degree to which observers spontaneously attributed emotions to the robot.

Experimental platform

The experimental setup (Figure 2.7) included a Pepper robot (Aldebaran[4]), an external tablet, a central executive station, and a monitor.

Scenes involved a collaborative task: setting a virtual lunch table according to etiquette rules. The table was simulated using an application[5] on the external tablet, in which the actor placed utensils in each scene. An external tablet was used instead of Pepper's built-in tablet to make the task more realistic.

[4]https://www.aldebaran.com/en/pepper
[5]https://appinventor.mit.edu/

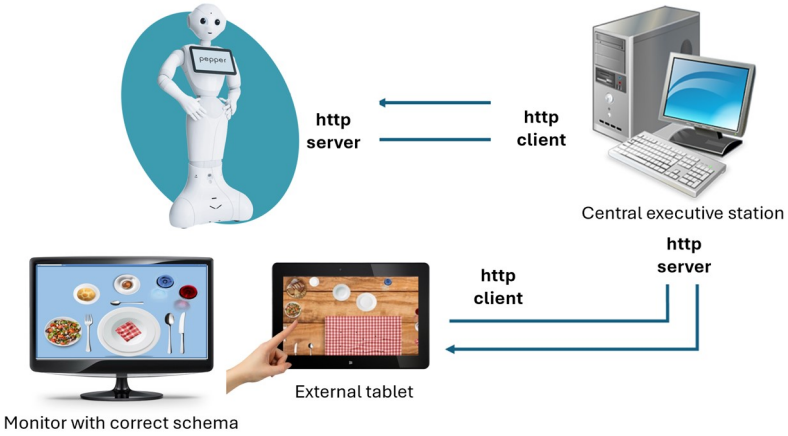


Figure 2.7: Platform used in the experiments.

The robot detected the actor's actions via client–server communication between the tablet and the central executive station, which transmitted the corresponding event signals to the robot.

Participants could view the correct table setting on the monitor, enabling them to detect intentional mistakes made by the actor. Errors varied in magnitude to elicit different levels of emotional response in the robot: minor misplacements were treated as easily correctable, whereas larger discrepancies elicited greater stress and stronger emotional reactions.

2.5.2 Results

Figure 2.8 presents a bar chart of the open-ended responses for each of the five scenes. Several notable findings emerge from this analysis. The robot, equipped with the proposed model, consistently exhibited emotional reactions to the events depicted in the scenes, and these reactions were perceptible to human observers.

More than 70% of participants in each scene referred to emotions either explicitly or implicitly in their descriptions, with over 50% explicitly naming an emotion. The remaining participants provided only a factual description of the actions performed by the actor or the robot, and a few left their responses blank.

Regarding the emotions recognised by the participants, the correct emotion, corresponding to that generated by the model, was identified with high accuracy: approximately 90% of participants selected the correct emotion for each scene. This observation is consistent with the experimental protocol's hypothesis that, when participants are prompted to choose from a predefined set of emotions, their attributions are more reliable. The model thus facilitated the recognition of the intended emotional state.

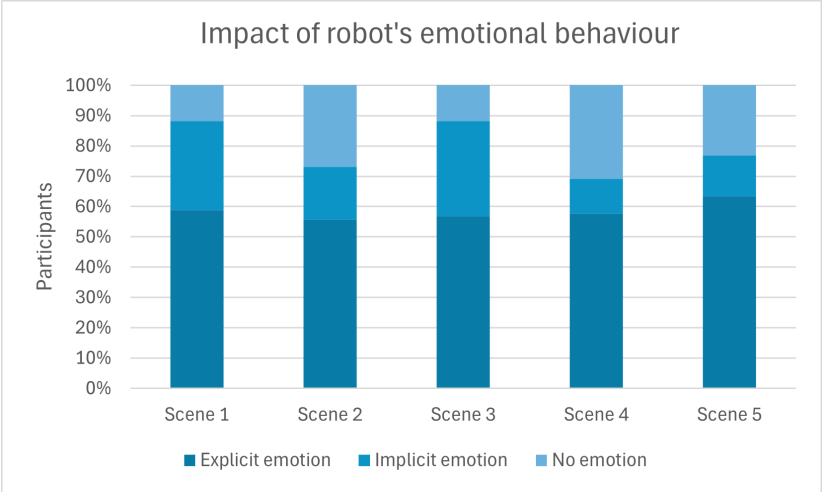


Figure 2.8: Results from the open-ended questions. For each scene, a high percentage of participants provided evidence of the robot's emotional behaviour, either by explicitly mentioning emotions in their free-text responses or by implicitly conveying them.

Figure 2.9 illustrates the distribution of the selected emotions, together with their intensities, across the five scenes. For each scene, the model's intended emotion was the most frequently selected, typically at high or medium intensity. This pattern supports the conclusion that the model reliably elicited the expected emotional response in observers.

Finally, Figure 2.10 summarises the percentage of participants who selected the correct emotion for each scene. The high recognition rates across all scenes demonstrate that the proposed model enables the robot to display emotional behaviours that are both observable and accurately interpreted by humans.

2.6 Summary

This work presented a novel formalisation for modelling emotions in robots by integrating inner speech mechanisms with Antonio Damasio's theory of emotions. The study extended Bosse's formalisation to develop SUSAN, an architecture designed to achieve self-awareness through reasoning about its emotional state and the surrounding environment. By leveraging memory and language during inner speech, the proposed approach aligns with Damasio's concept of Extended Consciousness.

SUSAN was successfully deployed and evaluated on a Pepper robot, where it generated both physical reactions and internal emotional states in response to external stimuli during a simulated human-robot interaction. In the experimental setup, the robot was tasked with setting up a table according to etiquette rules, resulting in observable emotional behaviours that were effectively perceived by human observers.

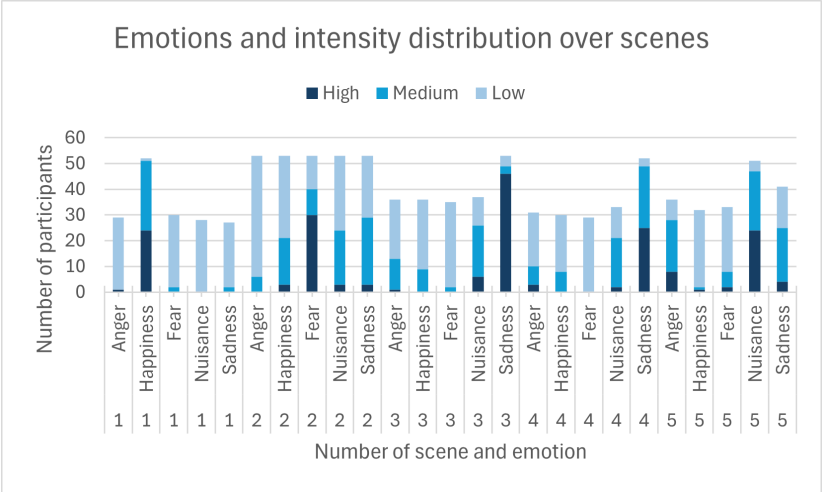


Figure 2.9: Distribution of emotions across the five scenes. Each bar indicates the frequency with which an emotion was selected at high, medium, or low intensity. Only the emotion expected by the model was predominantly chosen with high or medium intensity in each scene.

The proposed formalisation introduced an automated mechanism for generating inner-speech sentences, enabling more natural, human-like internal dialogues. This advancement contributes to the field of human-robot interaction by fostering more intuitive and socially meaningful exchanges between robots and humans. Experimental findings demonstrate that the model not only enables the robot to exhibit recognisable emotional behaviour but also supports meaningful interpretation by human participants. Furthermore, SUSAN shows the potential to evolve into a self-learning architecture, capable of acquiring new concepts from its interactions and internal states. This feature allows each SUSAN-equipped robot to develop a unique personality, shaped by its individual learning history.

The implications of this research are twofold. First, it provides a framework for integrating complex emotional models with inner speech mechanisms, enabling robots to experience, regulate, and express a wide range of emotions. Second, it lays the groundwork for applications in therapeutic, educational, and social contexts, where emotionally aware robots may enhance user engagement and support.

Future work will aim to refine SUSAN's emotional modelling and extend its deployment to more dynamic and unpredictable environments. Key directions include enhancing its capacity to learn from human interactions, adapting to novel emotional contexts, and implementing advanced self-regulation strategies. In particular, future studies will explore how SUSAN can manage its inner state through internal adjustments rather than external actions, enabling more sophisticated forms of emotional control. Additionally, integrating SUSAN with other cognitive architectures could fos-

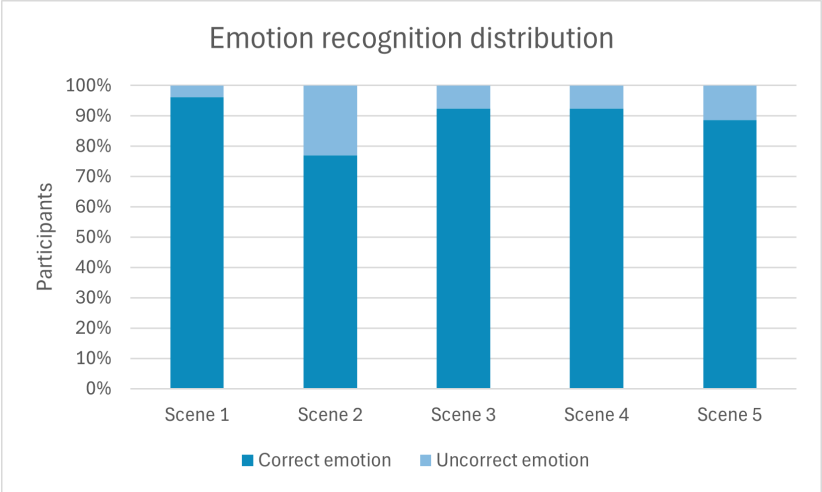


Figure 2.10: Percentage of participants selecting the correct emotion for each scene. This measure reflects the model's effectiveness in generating recognisable emotional reactions.

ter more advanced mental and emotional behaviours, paving the way toward the next generation of emotionally intelligent and socially adaptive robots.

2.7 Related Work

The computational modelling of emotions in robots has grown rapidly in recent years, reflecting the increasing research interest in this domain [39]. Many of these models are inspired by established psychological theories, including appraisal theories, rational approaches, and anatomical frameworks such as Damasio's model. Despite their shared goals, these models differ substantially in their theoretical underpinnings, represented components, functional implementations, and domain-specific adaptations. A detailed analysis of these differences is provided by Marcella and Gratch [105].

Appraisal theories have proven particularly suitable for modelling emotions, as they enable the formalisation of appraisal variables in a structured manner. Recent works based on these theories have yielded promising results and facilitated concrete strategy evaluations [23, 19, 24]. A recent review of appraisal-based computational emotion models is presented in [106], where the authors conclude that none of the existing models fully implements all emotional features and outline several avenues for future research.

Early foundational work by Picard introduced the concept of affective computing, establishing the importance of emotions in human-computer interaction and laying the groundwork for emotional intelligence in machines [66]. Building upon this foundation, Breazeal's development of Kismet demonstrated how emotional expressions could significantly enhance the naturalness and effectiveness of human-robot interactions [107].

Gratch and Marsella further advanced this field with the EMA (Emotion and Adaptation) model, a computational framework that integrates appraisal theory to simulate emotional processes [108]. This model allowed robots to adapt their behaviours based on emotional evaluations of events, marking a pivotal step in computational emotional dynamics.

Canamero emphasised the role of bodily states and intrinsic motivations in generating emotions, advocating for biologically inspired models that highlight embodiment and sensory processing [109]. Similarly, the SEAI (Social Emotional Artificial Intelligence) architecture, inspired by Damasio's theory of consciousness and the Somatic Marker Hypothesis, represents a significant advancement in embedding emotional processes within robotic systems [110]. Venturini et al. [111] also proposed a model integrating emotions into artificial agents' decision-making processes, further underscoring Damasio's influence in this area [112].

Recent research has explored the use of deep learning techniques to improve emotion recognition and generation. For instance, Barros [113] employed convolutional neural networks (CNNs) to detect complex emotional states from visual and auditory cues, enhancing the robustness and accuracy of emotion detection in real-world contexts.

Becker-Asano and Wachsmuth [114] proposed the WASABI architecture, which combines cognitive appraisal with somatic markers to simulate emotions in virtual agents, bridging the gap between cognitive and affective processes. Prinz's work further contributed to the embodied representation of emotions, framing them in terms of core relational themes such as danger and obstruction, and extending Damasio's formal analysis of emotion and consciousness [115].

In summary, computational models of emotion in robots have evolved from foundational principles of affective computing to sophisticated frameworks that integrate cognitive, social, and biological dimensions. These advancements have not only enhanced the realism and effectiveness of human-robot interaction but have also paved the way for novel applications in healthcare, education, and entertainment.

Part 2: Application of Inner Speech to Real-World Cognitive Tasks

Chapter 3

Robot and nurse collaboration in surgical preparation: preliminary evidence on the role of robotic inner speech

3.1 Overview

The investigation of robotic self-directed dialogue has recently gained increasing attention. Early studies reported promising effects on perceived transparency, robustness, and trustworthiness, primarily in simple collaborative tasks within Human–Robot Interaction (HRI). More complex and risk-prone collaborative settings have subsequently provided further insights. This work examines the role of robotic inner speech in advanced HRI scenarios, with particular emphasis on medical applications. In this context, the study should be considered a preliminary investigation, aimed at exploring the feasibility and potential impact of inner speech in a realistic healthcare setting. In the proposed context, a robot collaborates with a nurse to prepare the surgical table for a medical procedure. Precise placement of surgical instruments is essential, as inaccuracies may adversely affect procedural outcomes. The findings show that, in challenging situations like this, the robot's inner speech provides reassurance, helping users manage the stress associated with high-risk tasks. Moreover, the robot's self-dialogue enhances nurses' understanding of the instructions needed for proper surgical table setup. However, these results are derived from a limited experimental setting and should not be interpreted as broadly generalisable. Factors such as sample size, task specificity, and controlled conditions constrain the extent to which the findings can be extended to other medical scenarios or user populations. Further studies involving larger participant groups, diverse clinical contexts, and more complex interaction dynamics are necessary to validate and generalise these results.

3.2 Introduction

Intelligent service robots are increasingly deployed in healthcare settings, where they operate not merely as assistive devices but as active collaborators within clinical teams. This evolution raises substantial research questions in Human–Robot Interaction (HRI), particularly regarding effective communication strategies, trust calibration, and adaptive learning processes between human operators and robotic agents.

Within medical contexts, the quality of human–robot interaction is a critical determinant of both operational safety and user satisfaction, influencing healthcare personnel and patients alike. Advancements in communication protocols, cooperative task execution, and shared autonomy frameworks have enabled the design of robotic systems capable of functioning reliably in high-risk clinical environments. This progression is especially significant given that Robotics and Artificial Intelligence have become integral components of contemporary medical research and practice.

Notwithstanding the demonstrated effectiveness of robotic integration within clinical teams, an important application domain remains underexplored: the deployment of robotic systems as training tools for operating room nurses.

Operating room nurses assume a pivotal responsibility in ensuring successful surgical outcomes, contributing at a level comparable to that of surgeons. These highly specialised professionals are responsible for the precise preparation of the surgical workspace, including the arrangement of the so-called *"mother tables"* in which instruments are organised. Their expertise encompasses comprehensive knowledge of procedure-specific instrumentation, accurate spatial organisation of tools, and strict supervision of sterility, readiness, and immediate accessibility throughout the intervention.

Continual professional development and regular reinforcement of procedural knowledge are crucial. However, the academic focus on how robotic systems can aid in these training activities remains limited. While prior studies have examined robots functioning as tutors or learners in broader educational settings, their application within the operating room environment remains largely unexplored [116].

A promising research area is the integration of robotic inner speech, i.e., the robot's capacity to externalise a structured self-dialogue process [117]. Prior investigations indicate that this capability can positively influence collaborative dynamics by increasing system transparency, strengthening perceived robustness, and fostering user trust [118]. Nonetheless, the empirical validation of these effects has been largely confined to simplified, low-risk cooperative activities, such as arranging a lunch table, in which operational errors did not entail meaningful consequences. Whether similar benefits can be sustained in safety-critical and cognitively demanding environments, including

surgical preparation tasks, remains to be systematically examined.

This study introduces a new, practical, and interaction-focused framework for robotic nursing applications. It aims to assist in training and reinforcing operating room nurses' knowledge using a virtual human–robot collaboration system focused on preparing the surgical table. Within this framework, the robot functions as an interactive trainer, recreating realistic procedural scenarios and guiding the nurse through each step of instrument selection, spatial placement, and organisational setup.

A key aspect of the framework is its flexibility in inner-speech settings: the robotic system can operate with or without explicit self-verbalisation. This option facilitates controlled studies of the effects of inner speech on collaborative interactions and educational outcomes. When activated, the robot explains its internal reasoning, recalls procedural guidelines, and offers contextual reminders, thus enhancing procedural accuracy and situational awareness.

The main research question driving this study is: to what extent can a robotic system with inner speech improve collaboration and training effectiveness in high-risk medical environments, especially during surgical table setup with an operating room nurse?

This general question is organised into the following research sub-questions (RQs):

- RQ1. What are the architectural design principles and operational mechanisms underlying the proposed robotic framework for operating room nurse training, and how does the system function during surgical table preparation tasks?
- RQ2.a. To what extent does the activation of inner speech affect perceived transparency, trust, and reassurance in nurse–robot collaboration during surgical table preparation?
- RQ2.b. Does the presence of inner speech contribute to measurable improvements in learning performance during training sessions?

These research questions examine both the structural characteristics and the interactional dynamics of the proposed framework, with particular attention to its technical architecture, cognitive features, and impact on collaborative behaviour and knowledge acquisition among healthcare professionals.

Accordingly, the study pursues two primary objectives:

1. to detail the design rationale and operational workflow of the robotic framework developed for operating room nurse training and skill reinforcement, and
2. to present a pilot evaluation assessing the potential influence of robotic inner speech on collaborative quality and learning outcomes during surgical table preparation.

To this end, a preliminary experimental study was conducted involving two participant profiles: a domain expert (a vascular surgeon) and a non-expert participant. Both individuals participated in interactive training sessions with the robotic system under two experimental conditions: inner speech enabled and inner speech disabled. Following each session, participants completed structured questionnaires evaluating interaction quality and perceived learning effectiveness.

As a pilot investigation, the study does not seek to derive statistically generalizable conclusions. Instead, its goal is exploratory: to gather preliminary empirical evidence about how robotic inner speech can be applied in high-stakes healthcare training settings and to guide the development of more comprehensive future studies.

The remainder of this chapter is structured as follows. Section 3.3 outlines the state of the art in cognitive architectures for HRI and reviews current applications of robotics in the medical domain. Section 3.4 describes the methodological approach, including the proposed framework (encompassing the experimental physical setup and the operational principles of the inner speech cognitive architecture), the participant sample, the experimental design, and the data analysis procedures. Preliminary findings are reported in Section 3.5, and the chapter concludes with a discussion of the study in Section 3.6.

3.2.1 Theoretical Background

The investigation into how robotic systems could aid in updating and evaluating specialised professionals' knowledge in surgical table preparation is driven by several factors. In this context, particular attention is given to robotic inner speech, defined as the system's ability to externalise its internal self-dialogue process [117]. Empirical evidence indicates that this mechanism can enhance human–robot interaction by increasing perceived transparency, improving system robustness, and fostering user trust [118, 119].

In collaborative task execution, where humans and robots jointly pursue a shared objective, inner speech has been shown to modulate interaction dynamics in several ways. By articulating the reasoning behind its decisions, the robot enhances transparency. When it explicitly considers alternative actions in the face of obstacles, it shows adaptive robustness. Additionally, by revealing its internal motivations and decision criteria, it fosters calibration of trust. However, these findings are based on low-risk cooperative tasks, such as collaboratively setting a lunch table, which carry minimal consequences if errors occur and exert limited cognitive or emotional pressure [119]. The ecological validity of these findings in high-risk, safety-critical domains, including surgical preparation, remains to be determined.

More recently, a study on robotic inner speech [120] examined its application within dementia care contexts. The results indicated that caregivers interacting with a robot endowed with self-dialogue capabilities exhibited increased sensitivity to patients' needs. Specifically, the robot's externalised reasoning appeared to enhance caregivers' awareness of the situation and their focus during care. Extending this idea to surgical settings, a similar mechanism could help operating room nurses by maintaining their awareness of tasks and emphasising attentiveness during surgical table setup.

Recent empirical evidence suggests that social robots can serve as tutors or peer learners, offering notable cognitive and emotional benefits to learners. In controlled task environments, their effectiveness has, at times, nearly matched that of human instructors, a result often linked to their physical presence and embodied design [121]. A further review [122] emphasises the need to critically examine the legal, social, and ethical implications of deploying social robots, particularly in educational contexts involving children, educators, and other stakeholders. This study on how robotic inner speech influences interaction dynamics and learning outcomes offers a deeper insight into these wider implications.

Furthermore, a recent systematic review [123] emphasises that although robots are gaining attention as educational tools, their application in the professional training of healthcare workers remains limited, particularly in safety-critical areas such as operating rooms. This gap underscores the originality and practical importance of viewing robotic systems not merely as operational helpers but as structured training partners for highly specialised medical professionals.

The selected application scenario further enables an examination of whether robotic inner speech can accelerate the achievement of learning objectives and enhance the overall Quality of Learning (QoL). In this study, QoL is conceptualised in accordance with the UNESCO framework, which frames learning as a multidimensional process extending beyond the mere acquisition of declarative knowledge and emerging from the interaction of three principal components:

- Learner characteristics: the individual's prior domain expertise, cognitive capacities, and motivational profile, all of which influence the integration and restructuring of new knowledge.
- Enabling inputs: the instructional resources and contextual conditions supporting the learning process, including the technological infrastructure and, in the present case, the activation or deactivation of robotic inner speech as a cognitive-interactional feature.
- Learning outcomes: the observable competencies and procedural skills consolidated at the conclusion of the training process, determining the learner's ability

to perform tasks autonomously and reliably.

This chapter introduces a robotic framework that encompasses both the robotic system's architectural design and its application to the training of operating room nurses. A core element of this framework is the inner-speech cognitive model, which describes how the robot externalises its internal reasoning. This unified approach offers a structured foundation for human–robot collaboration and skill development, highlighting a vital cognitive mechanism that supports interactive learning. As a pilot investigation, this study explores the feasibility of the framework, with limitations in sample size and generalizability noted throughout.

In collaborative training scenarios, the framework facilitates a systematic assessment of training effectiveness. Notably, it shows how features such as inner speech in robot design serve as enabling inputs that influence not only immediate task performance but also understanding, retention, and skill transfer.

Within this study, the principal variables for characterizing nurse–robot interaction are defined as follows: reassurance, representing the degree to which the robot's behavior alleviates participant stress; trustworthiness, reflecting the human operator's willingness to rely on the robot's actions and judgments during joint tasks; transparency, indicating the extent to which the robot's decision-making processes and underlying motivations are interpretable; learning assessment, corresponding to the evaluation of knowledge gained regarding the specific surgical procedure; and reinforcement of learning, capturing the role of inner speech in consolidating acquired knowledge.

By taking this multidimensional approach, the framework goes beyond traditional efficiency metrics and offers a thorough way to assess if robotic training systems can significantly improve professional skills in critical clinical settings, like training operating room nurses.

3.3 State of Art

This section provides a focused review of cognitive architectures supporting inner speech in human-robot interaction, and key advancements in robotics for healthcare training, emphasizing their relevance to the proposed framework.

3.3.1 Architectural Frameworks for Cognition in Human–Robot Systems

Cognitive architectures such as ACT-R [124], SOAR [125][126], and ICARUS [127] provide a foundational framework for designing HRI systems, facilitating robotic per-

ception, reasoning, and action within complex and dynamic environments. These architectures are structured as modular systems, with each module simulating a specific cognitive function, coordinated to support higher-order cognitive processes.

In recent years, cognitive architectures have increasingly been leveraged to implement metacognitive capabilities, including inner speech [117], which enables robots to verbalise aspects of their thought processes and approximate self-reflective behaviour [17]. Within collaborative human–robot tasks, inner speech has been shown to positively affect interaction dynamics [118]. Specifically, by externalising its reasoning, the robot enhances transparency, making its decision-making processes interpretable; it demonstrates robustness by formulating alternative strategies when encountering impasses; and it fosters trustworthiness by communicating the motivations underlying its actions.

Although evaluating robotic inner speech is not entirely novel, prior studies have primarily focused on simple, low-risk tasks, such as arranging a lunch table according to informal etiquette norms. In these contexts, the beneficial effects of inner speech on transparency, robustness, and trust were empirically observed, providing a foundation for exploring its applicability in more complex and high-stakes human–robot collaborative scenarios.

In low-risk scenarios, delegating decision-making to a robot entails minimal consequences, and participants generally place high trust in the robot's judgment. Additionally, stress levels remain low because participants' actions are not critically affected by potential errors in object placement, a factor shown to influence study outcomes [128]. Whether such advantages generalise to high-stakes, safety-critical contexts, such as surgical preparation, remains an open question.

Recent research on robotic inner speech [129] indicates that caregivers interacting with robots endowed with this capability in dementia care contexts become more attuned to patients' needs. These findings suggest that inner speech may enhance attentional focus and situational awareness. By extension, a similar mechanism could support specialised professionals in surgical settings, helping nurses maintain heightened task awareness and attentiveness during surgical table preparation.

In parallel, studies have demonstrated that integrating cognitive architectures such as ACT-R enables robots to adopt diverse persuasive strategies and ethical stances during interactions [130]. The CASPER system [131], for example, employs qualitative spatial reasoning to anticipate other agents' goals and determine optimal collaborative behaviour, highlighting the effectiveness of symbolic cognitive architectures in HRI scenarios.

More recently, the integration of Large Language Models (LLMs) has expanded the capabilities of modular cognitive architectures in HRI tutoring scenarios [132].

By combining social reasoning and tutoring functionalities, these systems demonstrate preliminary evidence of managing complex interactions while storing and retrieving context-specific information. Together, these methods demonstrate that cognitive architectures can enhance robots' social and communication skills, thereby improving their effectiveness in educational and collaborative settings.

Evidence from domestic HRI contexts [133] further supports this perspective, showing that robots can coordinate task allocation with humans, clarifying "who does what, when, and where", to optimise joint performance. This principle of task sharing underpins our research in the medical field, highlighting the importance of structured interaction, training, and learning in high-stakes healthcare settings.

3.3.2 Robotics in Modern Healthcare Delivery

Although robotic systems have been successfully integrated into medical teams [134] [135] [136] and gained heightened prominence during the COVID-19 pandemic [137], their use as training tools for operating room nurses remains largely unexplored. This represents a novel and promising research direction. Robots have become indispensable in surgical environments, supporting surgeons in performing complex procedures that demand high precision, thereby enhancing both operational efficiency and clinical outcomes [138]. Over the past decades, the convergence of Robotics and Artificial Intelligence has profoundly transformed surgical practice, introducing advanced tools that facilitate more precise and efficient interventions. The first surgical robot, PUMA 560 [134], was employed in 1985 to conduct a brain biopsy, mitigating surgeon hand tremor and improving procedural accuracy [139]. This was followed by PROBOT in 1988 [135], developed for transurethral prostate surgery, and ROBODOC in 1992, which automated femoral preparation for hip replacement with higher precision and speed than manual approaches [140] [141]. Laparoscopic surgery has seen significant advances with the introduction of the Zeus and Da Vinci systems in the late 1990s. Zeus used three robotic arms, remotely controlled by the surgeon, whereas the Da Vinci system incorporated a patient-side cart, a surgeon console, and a vision cart, enabling highly precise and minimally invasive procedures [142]. The Da Vinci platform has undergone multiple iterations and is currently deployed across numerous surgical specialities, including urology, gynaecology, and general surgery, demonstrating both versatility and efficacy [143]. More recently, robotic platforms such as the Senhance Surgical System have introduced haptic feedback and eyetracking-controlled camera systems, enhancing surgeons' tactile perception and visual guidance during operations [144]. Collectively, these developments underscore the critical role of robotics in modern surgical practice and highlight the potential to extend robotic applications beyond

procedural assistance to include professional training and skill reinforcement for operating room personnel.

Another significant advancement is the introduction of the Versius Surgical Robotic System, which features modular, portable robotic arms, thereby enhancing adaptability across diverse surgical environments and potentially reducing operational costs [145]. In orthopaedics, the Mako robotic-arm-assisted surgery system has proven instrumental in increasing the accuracy of joint replacement procedures, improving implant alignment and clinical outcomes [146]. By enabling precise preoperative planning and intraoperative execution, the system contributes to reduced patient recovery times and higher satisfaction rates [147].

Beyond the operating room, hospitals are increasingly adopting robotic systems to support patient care. For example, the TUG robot automates the delivery of medications, meals, and supplies, streamlining logistics and alleviating nursing workload [148]. In rehabilitation, devices such as the Lokomat provide robotic-assisted gait training for patients recovering from strokes or spinal cord injuries, enabling more intensive and consistent therapy sessions [149]. Telemedicine applications have also benefited from robotics, with telepresence systems such as InTouch Health enabling remote specialists to interact directly with patients and clinical staff, thereby expanding access to expert care [150].

Robotic assistance within the operating room has also progressed. Penelope, developed in 2004, was among the first robotic surgical assistants capable of passing instruments to the surgeon via voice commands, although performance was limited by ambient noise [151]. GestoNurse, introduced in 2011, advanced this concept by employing sophisticated vision systems and pattern recognition to interpret manual signals from the surgeon [152].

Despite these advances, the role of the robotic scrub nurse, who prepares and manages surgical instruments, remains largely unexplored. The scrub nurse is critical in maintaining sterility, ensuring the availability of necessary tools, and keeping pace with evolving technical and scientific standards [11]. Developing robotic systems capable of fully or partially automating these functions could further improve surgical efficiency and safety, representing a promising direction for future research and innovation.

3.4 Methodological Approach

In this study, the robot works alongside the operating room nurse to set up the surgical table, a task known for its high stress because it directly affects surgical results. The nurse must possess detailed knowledge of the instruments required for each procedure and their optimal arrangement to ensure rapid accessibility and seamless handover to

the surgeon during the operation.

The interaction is structured as an instructional session designed to train or refresh the nurse's procedural knowledge. Each participant engages in two separate training sessions (Session 1 and Session 2), each focused on a different type of surgery. Following the training, participants complete a Test Session to evaluate their acquired skills.

Preliminary findings indicate that when the robot employs self-dialogue, nurses develop a heightened awareness of the knowledge they have gained. Additionally, the activation of inner speech is expected to influence key interaction variables, including increased transparency, greater trust in the robot, and reduced stress, as participants can observe the robot's reasoning process.

Including participants with varying levels of expertise helps evaluate the framework's usability and effectiveness across a range of learner profiles, thereby enhancing the overall applicability of the results. While the pilot nature of the study limits the ability to draw statistically robust conclusions, the carefully structured experimental design, combined with a mixed-methods data collection approach that integrates quantitative ratings and qualitative feedback, offers valuable preliminary insights. These findings can inform future research and suggest the framework's potential applicability to additional surgical procedures, healthcare professionals, and robotic platforms.

3.4.1 The experimental scenario: Simulated Vascular Surgery Preparation

The study centred on the preparation of a virtual surgical table for a vascular intervention. This scenario provides a controlled yet ecologically valid setting in which participants collaborate with the robot to position and organise the instruments required for the procedure correctly. A critical phase in the development of this scenario involved the formal definition of domain knowledge. This process was carried out through structured interviews with specialised vascular surgeons and an in-depth review of surgical manuals and relevant literature, ensuring that the virtual configuration accurately mirrored the procedural workflow and technical requirements of an actual vascular operation.

This methodological approach aligns partially with prior research [153], which employed the Story Dialogue Method (SDM) to investigate healthcare professionals' perceptions of social robots in care settings involving vulnerable users. Similar to the present work, that study adopted qualitative methods to capture professional perspectives and identified themes such as ethical responsibility, accountability, and system usability. Both investigations seek to anchor robotic interaction design in real-world clinical needs and practitioner expectations. However, the current study extends this

line of inquiry into a safety-critical surgical domain. Unlike SDM-based research, which explored care-oriented environments through narrative scenarios, the present framework addresses highly precise surgical tasks in which procedural errors may entail significant consequences. Furthermore, while the cited study elicited general professional viewpoints, the domain knowledge in this work was operationalised through direct consultation with expert surgeons, yielding a detailed, procedure-specific representation of surgical table preparation. This distinction highlights the innovative application of cognitive mechanisms, such as inner speech, to support professional training in risk-sensitive surgical contexts.

The knowledge base incorporated into the framework provides comprehensive information for each surgical procedure, including the intervention classification, the instruments required, their spatial organisation on the table, and the specific procedural phases in which they are employed. In vascular interventions, which address pathologies of the arterial, venous, and lymphatic systems, commonly utilised instruments include scissors, sutures, and specialised vascular tools. More broadly, vascular surgery encompasses a range of procedures, including abdominal, laparoscopic, otolaryngological, and cataract-related interventions, each characterised by distinct instrument sets and specific layout requirements on the surgical table.

This structured domain knowledge was embedded within the proposed framework (see Section 3), enabling participants to engage with a realistic simulation of the operating room environment. The virtual setting enables collaborative instrument placement by the robot and the participant, providing a safe, controlled platform to evaluate how robotic inner speech affects learning processes and task execution during complex surgical preparation activities.

Standard guidelines for organising a surgical table recommend arranging instruments according to the phase of the procedure in which they are used. Tools required at the initial stages of surgery should be positioned closer to the operating bed to facilitate immediate access. Typically, forceps, scissors, and spatulas are placed in the front row, followed by preparatory items such as wires, needles, and suction devices, with gauze positioned subsequently. Particular attention must be given to the visibility and accessibility of sharp instruments, including blades and needles, which should remain clearly exposed above the sterile drapes and not be concealed by other materials, such as gauze, to minimise the risk of accidental injury. In general practice, instruments are organised in two primary rows on the surgical table to ensure systematic access and spatial clarity.

Figure 3.1: The tablet interface where the participants drag and drop tools for preparing for the surgery

3.4.2 Materials and procedures

The robot used in this study is the Pepper robot by Aldebaran [1]. It features an integrated tablet that serves as the primary interface for the study. Nevertheless, the framework is designed to be flexible and can be applied to any robot that can connect with it, as explained below.

Preparing for the virtual servant table

The virtual servant table employed for the preparation task is implemented through an external tablet application, separate from the tablet integrated into the robot. This design decision was made to encourage participants to experience the preparation process as a collaborative activity with an external agent, rather than as a direct extension of the robot's internal functionalities. In this configuration, the table operates as an independent interactive medium through which the participant and the robot jointly pursue the shared objective of correctly organising the surgical instruments.

The graphical user interface of the tablet application is shown in Figure 3.1. The interface reproduces the surface of a servant table, visually represented by the characteristic green surgical drape, onto which both the participant and the robot collaboratively position the instruments.

All available instruments are shown in the lower part of the interface. Participants

[1] https://www.softbankrobotics.com/emea/en/pepper

select tools by dragging and dropping them into the upper section to set the final table arrangement. Placement is limited to predefined red markers in the upper area, which ensures a structured and standardised layout.

The red markers are organised into two rows. The lower row corresponds to the section of the servant table situated closer to the surgeon, indicating that instruments required during the initial phases of the procedure should be placed in this area to facilitate rapid access.

On the right side of the interface, a dedicated selection panel enables participants to communicate with the robot about the instruments verbally. Participants can either request details about a specific tool or direct the robot to perform actions, such as placing an instrument in its correct position. For informational queries, the robot provides spoken explanations of the selected instrument's characteristics and intended use. When an action request is issued, the robot physically relocates the specified tool to the appropriate position, thereby offering immediate visual confirmation of the correct arrangement.

The knowledge model

The definition of domain knowledge constituted a pivotal aspect of the proposed work. This information was obtained through interviews with specialised surgeons and by reviewing bibliographies of medical manuals.

The acquired knowledge pertains to each surgical procedure. It encompasses details such as the type of surgery, the instruments utilised, their arrangement on the surgical table, and the respective stages of their usage during the procedure. For instance, in the realm of vascular interventions, which treat disorders of the arterial, venous, and lymphatic systems, common tools include scissors and suture thread. Furthermore, vascular surgeries encompass a variety of procedures, including abdominal, laparoscopic, vascular, otolaryngological, and cataract surgery, each requiring distinct tools and specific placement locations on the surgical table.

The collected domain information was then formalised in an OWL ontology[2]. The decision to design an ontology was driven by the platform's aim to leverage the benefits of ontology modelling, including promoting interoperability between systems, facilitating the sharing of formal representations, and aiding knowledge acquisition. This transition from conventional computer science, focused on automated information processing, to epistemic processing, centred on automatic knowledge processing, underscores the platform's commitment to advancing knowledge-driven functionalities.

Figure 3.2 displays a segment of the knowledge base accessible to the robot. Given

[2]https://www.w3.org/TR/?tag=data

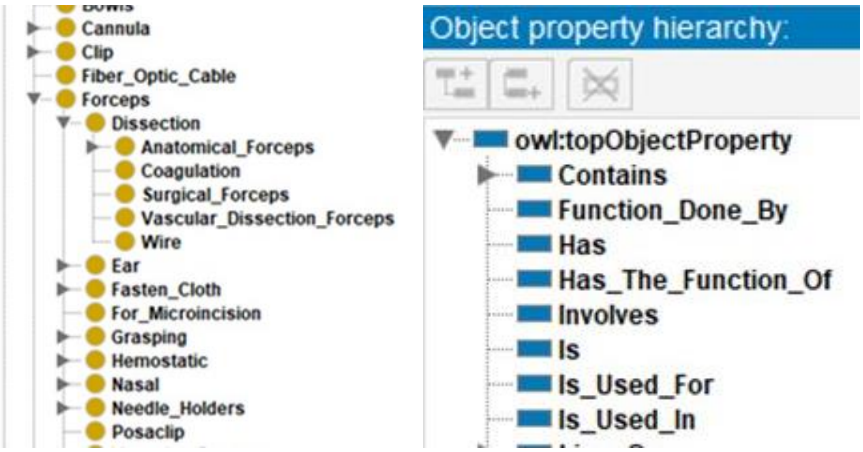


Figure 3.2: An excerpt of the ontology modelling the domain under investigation related to some vascular interventions. A small subset of the classes and some object properties are represented

available resources, the arrangement of the surgical table lacks a universal standard and varies by surgical speciality and customary operating room practices.

Typically, one guideline for organising the surgical table is to position instruments according to the phase of the procedure in which they are utilised. Instruments used earlier in the surgery should be placed closer to the operating bed. Additionally, it is recommended to arrange instruments such as forceps, scissors, and spatulas in the front row, followed by preparatory items such as wires, needles, and suction equipment, and finally gauze. It is crucial to ensure that all instruments, particularly sharp objects such as blades and needles, are clearly visible above the drapes covering the surgical table and not obscured by other tools, such as gauze, to prevent accidental injury.

In general, the instruments are placed in two rows.

The inner speech architecture

Inner speech is a well-known and well-studied topic in human psychology [154] [155], and recent studies have made interesting claims about its effects on human cognitive functions [156]. The importance of automating inner speech was analysed [19], and initial results confirm its role in human-robot interaction [119]. The robot's inner speech model is outlined in Figure 3.3 and is the same one proposed by some of the authors in [16]. The model has now been modified to accommodate the compromising scenario under investigation. The model is integrated into the robot's functioning when the robot must talk to itself.

Specifically, the model operates as follows: the Perception module detects the par-

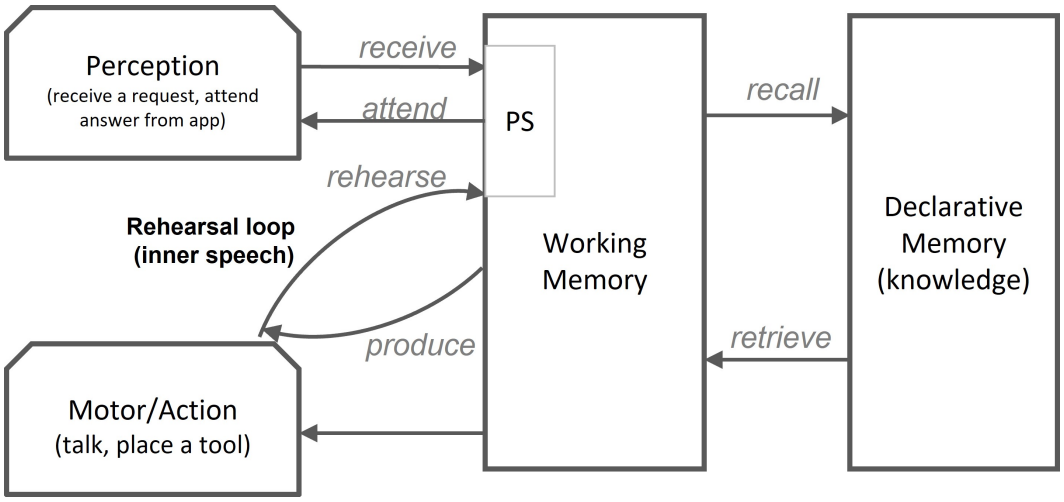


Figure 3.3: The inner speech model underlying the robot's functioning when it talks to itself.

ticipant's actions via the external tablet. In contrast, the Motor/Action module implements the robot's actions in response to these detections. The Phonological Store (PS) encodes the participant's action, which can be selecting a question, performing an action on a tool (touch up, drag-and-drop, or touch down), or requesting that the robot move a tool. To encode an action means to associate a symbolic linguistic representation with it; that is, when the participant acts as the app interface, the encoded action is a set of words that represent that event. For example, if the participant touches down a surgical scissor, the encoded action by the PS will be the sequence of words `(move, suture_scissor)`.

The Rehearsal loop starts once the action is encoded. The Working Memory recalls from the Declarative Memory the set of facts related to the encoded action, and once these facts are retrieved, a sentence for inner speech is composed. For example, for the aforementioned encoded action `(move, suture_scissor)`, the Working Memory retrieves from the knowledge model the concepts associated with the words in the encoded action. In the example, the concept `suture_scissor` is linked by the properties `p:is_used_in`, `p:is_used_for` and `p:lies_on` to the concepts representing the phase (`middle`), the function (`cut_sutures`) and the place on the servant table (`second, four`) (the second row and the fourth point in the interface) of the touched tool. The resulting inner thought corresponding to these emergent concepts will represent a sort of reflection by the robot on the touched tool, and the inner sentence will look like *"The suture scissor is employed in the middle phase of the surgery and is used to cut the sutures. Its position is in the second row on the fourth point."*. The Motor/Action module generates the sentence, which the PS then rehearses. The cycle restarts as the PS encodes it by chunking relevant words, the Working Memory recalls additional related facts, and a new sentence is composed. In this example, the concept

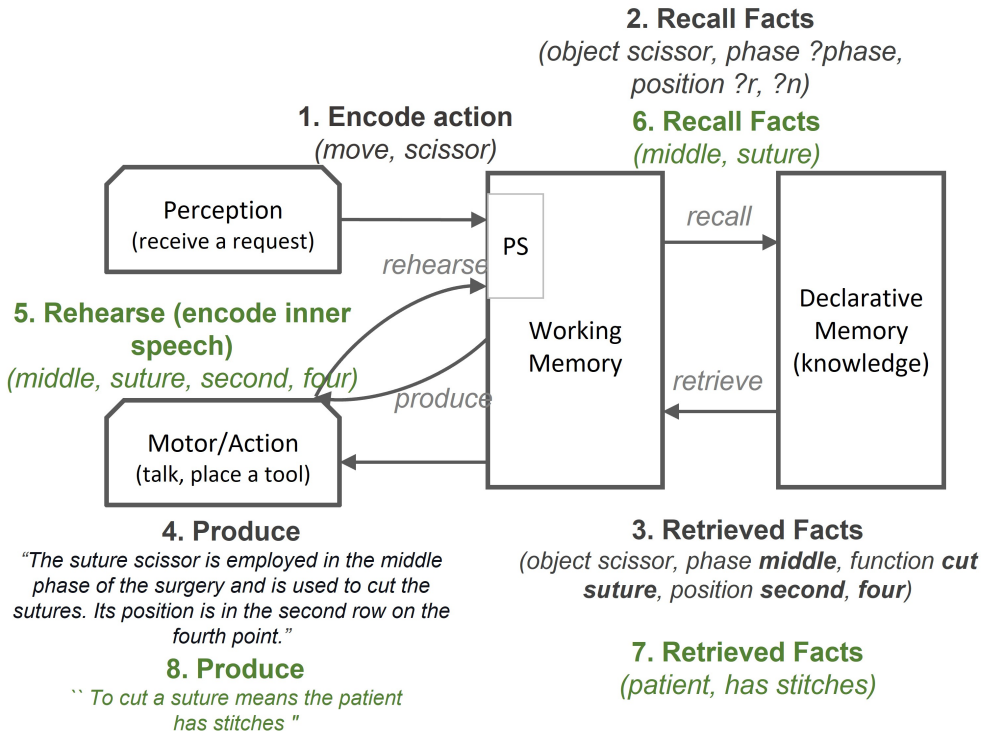


Figure 3.4: The phases of the inner speech loop and the involved components of the architecture.

(`cut_suture`) is linked to the concept `suture`, and a new thought will be *"To cut a suture means that the patient has stitches"*, which in turn recalls the concept `stitches` and composes another inner sentence *"The stitches is used for sewing wounds"*. The loop is repeated until no further facts from the Declarative Memory emerge. Figure 3.4 represents the described steps on the inner speech architecture, highlighting the functions of each module.

An exhaustive explanation of how the facts are retrieved and the sentences are composed is presented at [157].

The model was expanded and integrated to investigate the new medical domain, and inner speech was then calibrated for the field under investigation.

The whole physical infrastructure

The whole framework is shown in figure 3.5. A local server manages the communication between the robot and the external tablet. In particular, when the participant takes an action on the tablet, the app sends a request via the `Request to Robot` interface to the server, which interprets it and invokes the corresponding service in the robot's software via the `Robot service` interface.

The service can be associated with either generating a sentence or executing an

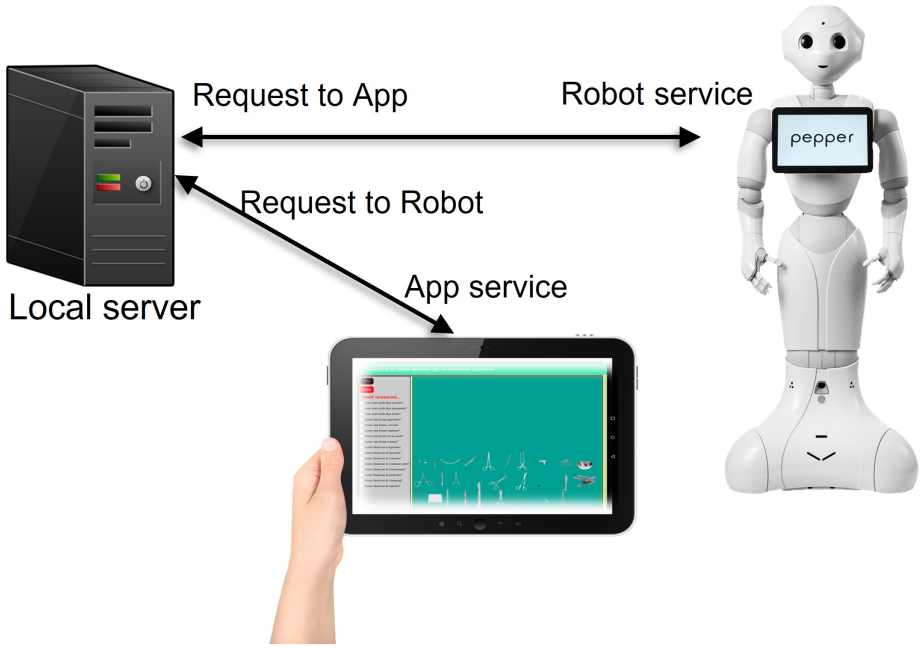


Figure 3.5: The physical infrastructure of the robotic nursing platform. A local server enables the robot-tablet communication, allowing the robot to perceive the participant's actions on the tablet and to respond opportunistically.

action that the robot needs to perform on the external tablet. If the request concerns sentence generation, the required services will vary with the robot's functionality. If the robot incorporates inner dialogue, it will engage the processes underlying the cognitive architecture of inner speech as previously described, resulting in the production of emergent sentences. However, if the robot does not engage in self-dialogue, the service will directly utilise the robot's knowledge to retrieve the specific sentence, which the robot will overtly produce.

On the other hand, when the request involves an action for the robot to perform, the corresponding robot service sends a request to the server via the `Request to App` interface. This request specifies the parameters governing the motion of a tool on the app surface. Subsequently, the server will initiate the corresponding events in the app via the `App service` interface aligned with this movement.

Table 3.1 outlines the potential requests initiated by both the participant and the robot, along with the corresponding services invoked.

3.4.3 Participants

At this preliminary stage, the experiments aim to validate the framework's functionality, and earlier intriguing observations and findings have already emerged. Given the pilot nature of the study, the primary objective was not to achieve statistical represen-

Table 3.1: Requests and services enabling interaction between the participant and the robot via the external tablet.

Tablet action (by participant)	**Request to Robot** (by app)	**Robot Service** (by robot)	**Request to App** (by robot)	**App service**
Touch down a tool	Produce reflective inner sentence	Inner speech routines (encode tool)	-	-
Touch up a tool	Produce feedback inner sentence	Inner speech routines (encode tool)	-	-
Ask a question	Answer to question	Inner speech routines (encode question)	-	-
Ask the robot to take an action	Execute action	Inner speech routines (encode action) Movement's parameters retrieval	Place tool	Change tool's location on the app interface

tativeness, but rather to gather qualitative insights and formative feedback on both the framework's operation and its potential impact on training. For this reason, recruitment was deliberately kept limited and targeted. A vascular surgeon was invited as the domain expert, while a non-expert participant with no prior medical background was involved to represent a contrasting learner profile. Both were recruited through direct personal contact, which ensured not only their availability but also the ability to maintain communication for follow-up or clarification as needed. This approach was particularly valuable at such an early stage, where iterative feedback and ease of access to participants were essential to refine the experimental setup and for future larger-scale evaluations. Moreover, the rationale for recruitment follows UNESCO's framework for education [158], which identifies variables influencing the quality of learning. These variables include learner characteristics, encompassing the learner's initial ap-

titude and prior knowledge, enabling input, which refers to the learning instruments utilised to impart knowledge and facilitate learning, and acquired skills, representing the learner's final level of knowledge attained at the conclusion of the learning process.

To comprehensively examine these variables and the overall impact of the proposed platform on learning quality, participants were recruited based on the following criteria:

- *Learner characteristics*: The experimental session included a specialised surgeon in vascular intervention and an individual lacking expertise in the field. The stark contrast in prior knowledge levels and participant aptitudes enables evaluation of the effects of utilising the proposed platform across varying learner characteristics.

- *Enabling inputs*: Each participant, with their unique learner characteristics, utilised the proposed physical infrastructure under two different conditions: with and without the inclusion of the robot's inner speech. Specifically, each participant participated in two separate lesson sessions, one incorporating the robot's inner speech and the other omitting this feature.

- *Skill assessment*: At the conclusion of the trial, each participant assessed their level of knowledge, enabling the evaluation of the final effects of the robot's inner speech on learning quality.

3.4.4 The trial

A single trial consists of two sessions (*Session 1* and *Session 2*) of an interactive lesson. At the end of the trial, a *Test Session* starts. The lesson begins with the robot presenting the app interface and available actions (query selection or drag-and-drop a tool). In this context, the robot specifies that the lower side of the servant table will be positioned near the operating bed.

Then, the robot introduces the lesson's overarching topic, typically a specific surgical procedure, such as vascular surgery. The robot outlines the various stages typically involved in surgery, including the initial, middle, and final phases, and provides information on the morphology and function of the surgical instruments, categorised by type (e.g., forceps, surgical scissors, etc.). For each tool type, the robot emphasises the surgical phase during which it is typically utilised, aiming to reinforce identification of its correct location on the surgical table (with earlier-used tools positioned closer to the surgeon).

At the end of the robot's presentation, Session 1 of the lesson begins, followed by Session 2, and finally the Test Session. During each session, participants may request information from the robot about the instruments or ask the robot to place an instrument on the virtual table to visualise its correct location. Table 3.2 summarises the

Table 3.2: The executable actions by the participant and by the robot. Each row indicates the participant's action and the robot's corresponding reaction. The robot's reaction differs in the two sessions of the lesson.

Action by the participant	**Reaction by the robot**	
	Session 1 with inner speech	*Session 2 without inner speech*
Touch down a tool Drag and drop a tool	Reflective inner speech production	–
Touch up a tool	Feedback inner speech production	–
Ask the robot an information	Retrieve and produce the answer	Retrieve and produce the answer
Ask the robot to move a tool	Reflective inner speech production Place the tool	Place the tool

possible actions participants can take during the lesson, along with the corresponding type of robot response, which may vary between sessions.

At the conclusion of each session, the interface resets to its initial configuration. Each lesson focuses on a different subject, with the robot randomly selecting one of the vascular interventions.

During the Test Session, the robot displays a tool on its tablet, and participants have a limited time to identify the tool and place it on the surgical table. The robot provides positive or negative feedback based on the accuracy of the participant's selection. The primary objective of this session is not to evaluate the participant's acquired skills but rather to foster awareness of their knowledge level.

Each session is further detailed below.

Session 1

During this session, the robot talks to itself. There are two different kinds of inner dialogue, that are:

1. *Reflective:* the reflective inner dialogue is activated when the participant touches down on a specific instrument to drag it. This inner dialogue simulates the thought process people undergo when recognising an object that elicits memories or associations. In this context, the robot utilises its knowledge to recall relevant facts related to the tool. These facts are articulated in linguistic form and represent the robot's thoughts. Each thought leads to another, initiating an inner dialogue within the robot. As a result, the robot verbalises its thoughts about

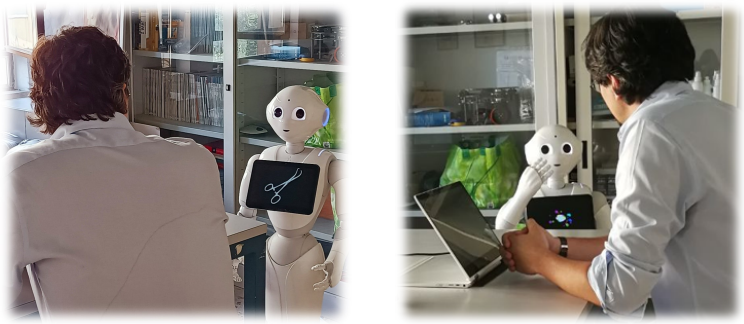


Figure 3.6: A photo related to the test session attended by the surgeon. The Pepper robot displays a tool, which the surgeon will then drag and drop onto the app interface to place it on the surgical table.

the tool, its location, and the surgical phase in which it is typically utilised. The purpose of this inner dialogue is to enhance transparency in the interaction, allowing the learner to engage with the information about the surgical instrument through the robot's reflections and thereby reinforce their knowledge.

2. *Feedback:* The feedback inner dialogue provides feedback to the participant based on their actions. The feedback will be positive when the participant correctly locates the tool on the servant table; otherwise, it will be negative. It is intended to reinforce the executed action, providing motivation when feedback is positive or encouragement when feedback is negative, thereby reducing stress and improving learning.

The evolution of Session 1 depends on the participant's choices. The session ends when the participant decides to stop, as they still need the required information.

Session 2

Session 2 starts when Session 1 ends. The app is reset to its initial configuration. The surgery will involve a different intervention from that used in the first session. The robot and the participant collaborate again to prepare the servant table for this new intervention. This choice allows for considering a new domain for which the participant had not acquired knowledge in the previous session.

This time, the robot does not talk to itself. It either answers the participant's questions or provides a tool if the participant requires one. During table preparation, the participant does not hear the robot's thoughts, as shown in Table 3.2.

Test Session

At the conclusion of the two previously described sessions, a test session is conducted to assess the participant's awareness of the knowledge acquired. The app interface is reset to its initial state, and the test entails displaying a tool on the robot's tablet. Subsequently, the robot specifies one of the interventions covered in the preceding sessions. The participant is then given a brief period to identify the displayed tool and place it on the surgical table in the expected position for the corresponding surgery, using drag-and-drop.

The robot provides feedback, indicating whether the participant's placement of the tool is correct or incorrect. The test continues until all tools have been placed, or until the participant decides to conclude it.

It is crucial to note that the robot does not evaluate the participant's acquired skills during this session. Instead, the primary objective is to foster the participant's awareness of their acquired knowledge. Participants will subsequently express self-assessment through a questionnaire administered at the conclusion of the trial. Figure 3.6 shows the test session with the surgeon.

3.4.5 Measures

The pilot experimental session enables the acquisition of preliminary results on the effects of the robot's inner speech on interaction and learning quality in a compromising scenario.

The questionnaire prompts the participant to provide their impressions of the interaction and to assess their skills at the end of the trial. The open questions were formulated to assess the level of reassurance the robot provides when collaborating with humans on riskier tasks, and to evaluate participants' impressions of the role of the robot's inner speech in the interaction and in the final skills acquired.

The closed questions required the participant to evaluate, using a Likert scale ranging from 0 (no evidence) to 5 (strong evidence), a set of statements from which variables could be inferred. A question evaluates only one of these variables, and for each variable, there is a set of questions that evaluate it.

The variables assessed in the study are as follows:

- *Reassurance*: This variable measures the extent to which participants feel comforted and supported by the robot during the task of preparing the servant table. Questions about this variable prompt participants to reflect on their level of stress and responsibility while placing the instruments, and to assess how effectively the robot reassures them.

Table 3.3: The correspondences between the observed variables and the groups of questions from which these variables were evaluated.

	Comparing the first and second sessions... You have to prepare the servant table for an imminent surgery! The patient needs one of the surgeries you learn. Let's answer the following questions by scoring the evidence of each claim.
Reassurance	Q1. Did you trust the robot more during the first session than during the second? Q2. Was the task more pleasurable during the first session than in the second one? Q3. Do you feel safer facing the surgery in the first session than in the second one?
Trust	Q1. Did you trust the robot more during the first session than in the second one? Q2. Did you take more into account the information provided by the robot in the first session than the second one? Q3. Did you consider the information provided by the robot more reliable during the first session than in the second one?
Transparency	Q1. Do you better understand what the robot does during the first session than in the second one? Q2. Do you think to better trace the process of the robot in the first session than the second one? Q3. Were you able to better hear the robot's thoughts during the first session than in the second one?
Learning Assessnent	Q1. Do you feel an advantage in learning during the first session of the lesson? Q2. Was the learning simpler in the first session than in the second one? Q3. Was the training clearer in the first session than in the second one? Q4. Do you think your learning of the first session is better than those of the second one?
Reinforcement Learning	Q1. Did you need to require more information from the robot during the second session than the first one? Q2. Do you think the robot's thoughts facilitate the information acquisition? Q3. Do you feel the learning provided during the first session is more robust? Q4. Do you feel more prepared after the first session than after the second one?

Table 3.4: The Likert value for each variable is computed as the arithmetic mean of the Likert values selected by each participant for the corresponding group of questions.

Learner characteristics	Domain Expert		No domain Expert	
	Enabling Input		Enabling Input	
	Platform with inner speech	**Platform without inner speech**	**Platform with inner speech**	**Platform without inner speech**
Reassurance	4	2	4	2
Trust	4	3	4	3
Transparency	5	3	4	3
Learning Assessnent	4	3	4	2
Reinforcement of Learning	4	3	5	2

- *Trust*: This variable gauges the level of trust participants have in the robot's decision-making abilities. Given the potentially risky nature of the task, participants may exhibit varying degrees of trust in the robot's capabilities, particularly relative to less risky tasks.

- *Transparency*: Transparency refers to how clearly participants can discern the underlying decision-making processes of the robot. Unlike previous research, in which transparency was evaluated as a binary value, this study asked participants to rate the robot's transparency on a scale to indicate the perceived level of transparency.

- *Learning assessment*: Participants evaluate their acquired skills in the specific surgery after interacting with the robot. They compare their learning experiences when the robot engages in inner speech versus when it does not, assessing whether the robot's inner speech facilitates a simpler and clearer learning path and improves skill acquisition.

- *Reinforcement of learning*: This variable assesses how the robot's inner speech reinforces skill acquisition and makes the learning path more robust. Participants reflect on whether the inner speech serves as an additional source of confirmation for the information learned.

The table 3.3 illustrates the relationship between the questions and the variables assessed in the study.

3.5 Results and Findings

The direct interaction with the professional surgeon specialising in vascular interventions proved highly valuable, as he provided crucial insights into the platform's structure and knowledge model. His initial evaluation of the framework, focusing on its affordance, usability, and utility for the investigated use case, yielded positive feedback. The platform appears user-friendly for medical end-users with varying levels of expertise and requires no prior technical skills.

Analysis of open-ended questions revealed that participants expressed excitement when interacting with the robot. Interestingly, differences emerged based on participants' prior experience with humanoid robots: those without prior experience initially expressed some distrust, which dissipated quickly during the interaction, whereas experienced participants approached the session with greater curiosity and excitement. The discovery of the robot's inner-speech capability was gradual for both groups, adding intrigue and surprise to the interaction and validating previous findings [128].

Perceived differences between interactions with and without inner speech varied based on participants' prior knowledge, as indicated by closed-ended questions. The final value for each variable is computed as the approximate upward-arithmetic mean of the participant's scores on the questions about that variable. The obtained results are reported in Table 3.4 and support the hypotheses underlying the study, indicating that reassurance, reliability, and training quality improve when the interaction involves the thinking robot. Novice participants found inner speech to be a vital source of information, significantly enhancing learning. However, experienced participants showed minimal differences in learning assessment and reinforcement between the two modes of robot operation, likely due to their robust prior knowledge.

Interestingly, expert participants were attuned to subtle variations in voice pitch between the two interaction modes, perhaps indicating a heightened sensitivity to nuances in communication. However, their focus on these nuances did not detract from the overall positive reception of the robot's instructional role. Both participants recognised the robot's potential as an effective teaching tool for novices, providing valuable insights and guidance throughout the learning process. Their experiences underscored the notion that transparency, enabled by a robot equipped with inner speech, can catalyse skill enhancement and knowledge acquisition. Moreover, participants noted the potential of robotic trainers to catalyse novel forms of interaction between humans and machines, thereby paving the way for more collaborative and effective learning environments across various domains.

3.6 Discussion

The preliminary findings of this pilot study align with prior evidence from low-stakes collaborative tasks, such as table-setting scenarios, where inner speech enhanced transparency, robustness, and trust in human-robot interaction. In our case, participants reported that the robot's inner speech increased their awareness of the surgical preparation task and provided reassurance throughout the interaction. This is consistent with results from caregiving contexts, in which inner speech has been shown to make caregivers aware of and sensitive to patients' needs [129]. At the same time, the study extends the existing state of the art on robots' inner speech by testing it in a high-stakes, risk-sensitive medical scenario. While several reviews have extensively documented the use of robots as tutors, peer learners, or educational partners, particularly in classroom-based settings, to our knowledge, no research has investigated the integration of a cognitive architecture supporting inner speech into the training of healthcare professionals. Our preliminary results suggest that such an approach may enrich collaboration dynamics and enhance learning outcomes, echoing the multidimensional perspective on the quality of learning proposed by UNESCO's framework. Furthermore, our findings on the role of inner speech in fostering transparency resonate with broader reflections on human autonomy in the age of autonomous technologies. Recent conceptual work [159] argues that human autonomy should be understood not merely as control, but as the overall space of action and the range of available choices. In this light, systems that make their reasoning processes explicit, such as through inner speech, can contribute to strengthening rather than constraining human autonomy. By providing insights into its decision-making, the system could effectively enlarge the participant's perceived space for action, supporting informed collaboration. This perspective aligns with the view that autonomous technologies, when designed as datanomous, operating within the limits of their own data and making those boundaries transparent, can foster trust and reinforce human oversight, rather than diminishing it. From a methodological perspective, the inclusion of both an expert (vascular surgeon) and a non-expert participant proved valuable, as it allowed testing the framework across distinct learner profiles. This dual perspective provided complementary insights into the robot's usability and pedagogical relevance. Furthermore, the mixed-methods data collection strategy (Likert-scale ratings combined with open-ended responses) enabled to capture not only quantitative impressions but also qualitative dimensions of the interaction, such as affective responses, perceived reassurance, and attentional engagement. This methodological choice enriched the interpretation of results, despite the small sample size. An additional consideration arising from our study concerns the ethical and societal implications of deploying robots in high-stakes care and training environments. As

highlighted in recent literature, care robots pose complex regulatory, moral, and practical challenges, including the potential for over-reliance on machines, privacy concerns, and accountability in the event of errors [160] [161]. Our findings suggest that features such as inner speech, which enhance transparency and support humans. Supervision may help mitigate some ethical risks by clarifying the robot's intentions and decisions to users. Indeed, congruence between robot verbalisations and actions has been shown to significantly impact humans' intuitive understanding of robot intentions, fostering trust and ethical interaction patterns [162]. Incorporating such design considerations into training platforms for healthcare professionals can not only improve learning outcomes but also ensure that ethical and societal dimensions are actively addressed in the human-robot collaboration. Moreover, the continuous presence of a researcher was perceived as reassuring. This feedback provides constructive insights for refinement, highlighting areas where usability can be further optimised in future iterations. In particular, the role of researcher presence could itself become an observational parameter in subsequent studies, helping assess the extent to which human supervision influences trust, reassurance, and ease of use. Acknowledging these points strengthens the study by ensuring that future large-scale applications will build upon both the positive outcomes and the lessons learned from this pilot stage. The primary limitation of this work is indeed the limited number of participants, which prevents drawing statistically generalizable conclusions. However, this constraint is consistent with the exploratory nature of pilot studies, where the priority is to validate the experimental framework and gather initial feedback. Additional limitations include the restricted scope of the surgical task (a single vascular intervention scenario) and reliance on a single robotic platform (Pepper), which may limit the generalizability of the findings to other contexts or robot morphologies. Despite these limitations, this study contributes to the literature by opening a novel line of inquiry into how cognitive features, such as inner speech, can support professional training in medicine. Future work should expand the participant pool, diversify the medical procedures considered, and compare different robotic platforms and interaction modalities. Longitudinal studies will also be needed to assess whether the observed benefits of inner speech persist over time and translate into improved surgical practice. light the potential of inner speech as a mechanism for enhancing both human-robot interaction and learning outcomes in healthcare training, laying a foundation for more comprehensive future investigations. The primary limitation of this work is the limited number of participants, which prevents drawing statistically generalizable conclusions. Additional limitations include the restricted scope of the surgical task (a single vascular intervention scenario) and reliance on a single robotic platform (Pepper), which may limit the generalizability of the findings to other contexts or robot morphologies. Despite these limitations, this study contributes

to the literature by opening a novel line of inquiry into how cognitive features, such as inner speech, can support professional training in medicine. In conclusion, although preliminary, the results highlight the task's inherent complexity and reassure participants during the interaction, echoing similar effects observed in caregiving contexts [129], where inner speech heightened attentiveness to patients' needs.

3.7 Summary

The collaborative activity examined in this study centres on the joint preparation of the surgical table by the robot and the operating theatre nurse. The proposed platform is designed as a training and knowledge-reinforcement tool to support nurses in maintaining up-to-date procedural competencies consistent with established surgical standards. In this context, the investigation specifically addresses the function of robotic inner speech within a safety-critical setting, namely the preparation phase preceding a surgical intervention. Given that inaccuracies during table setup may directly influence surgical workflow and outcomes, this scenario represents a significantly more demanding testbed than low-risk collaborative tasks typically explored in prior research.

The interaction was evaluated through a multidimensional framework encompassing reassurance, trustworthiness, transparency, and quality of learning. These variables were operationalised using structured measures, complemented by open-ended questions to capture participants' initial perceptions during the pilot validation of the platform. An essential component of the validation process was consultation with a specialised vascular surgeon, who assessed both the system's usability, particularly for non-expert users, and the clinical fidelity of the embedded knowledge model. This expert feedback ensured that the simulated environment accurately reflected procedural standards and realistic workflow constraints.

Responses to the platform were favourable across both expert and non-expert participants. Users reported confidence in interacting with a robotic tutor and expressed enthusiasm for a system capable of supporting task execution through guided collaboration. The presence of inner speech was perceived as enhancing transparency, as the robot's reasoning processes were made explicit. This explicitness contributed to greater comfort and trust, while simultaneously reinforcing procedural understanding by providing contextualised task-related explanations.

Participants' views differed regarding whether robots with cognitive features, such as inner speech, could be trusted to perform fully autonomous surgical tasks. Non-expert users generally expressed caution about replacing human professionals in surgical contexts, indicating a preference for robotic systems to operate in supportive,

educational, or collaborative capacities rather than as independent operators.

In contrast, participants with domain expertise held a more optimistic view of the ability of cognitively enhanced robots to autonomously perform complex medical tasks. Despite this openness, they emphasised the continued necessity of human oversight as a critical safeguard against unforeseen errors or system limitations.

Future developments of the framework will focus on enriching the dialogical interaction between the robot and the nurse, thereby refining the system's pedagogical and collaborative dimensions. Additionally, expanding the participant cohort will be essential to obtaining more robust and generalizable insights. While the present study is exploratory, the findings provide preliminary evidence supporting the feasibility of integrating cognitive features, such as inner speech, into robotic training platforms for high-stakes medical environments and lay the groundwork for more extensive empirical investigations.

Chapter 4

Conclusions

This doctoral research has investigated the integration of emotional modelling and inner speech mechanisms in robotics, with the aim of advancing transparency, emotional plausibility, and effectiveness in Human–Robot Interaction, particularly in collaborative and high-stakes environments. Across three interrelated studies, two computational emotional models were developed and validated, and the functional and relational impacts of robotic inner speech were demonstrated, culminating in a real-world application in a medical context.

The first study focused on incorporating inner speech within an appraisal-based computational framework, showing that self-directed dialogue enables robots to perform structured self-reflection, evaluate situational variables, and direct attention toward contextually relevant information, thereby supporting more accurate computation of appraisal variables and resulting in emotional dynamics that closely mirror those observed in healthy adults under stress. By verbalising its internal reasoning, the robot not only enhanced the robustness and flexibility of its decision-making but also improved human partners' understanding of its cognitive and affective processes, fostering more transparent, predictable, and cooperative interactions.

Building on this foundation, the second study addressed the theoretical and practical question of whether robots can not only detect or simulate emotions but also experience them through a combination of cognitive and embodied mechanisms. Drawing on Damasio's theoretical framework, emotions were conceptualised as emerging from the dynamic interplay between internal bodily-like signals and cognitive appraisal processes. Inner speech mediated this interaction by allowing the robot to articulate both contextual assessments and internal state evaluations through self-directed dialogue, producing emergent emotional responses that were internally coherent, externally recognisable, and psychologically plausible. Experimental deployment on a robotic platform demonstrated that human collaborators could reliably perceive these emotional states, thereby fostering empathy, emotional attunement, and a stronger

relational connection between humans and robots.

The third study extended these insights into a high-risk, practical scenario in medical Human–Robot Interaction, in which a robot assisted a nurse in preparing a surgical table. In this context, where task accuracy is critical and human operators experience significant stress, robotic inner speech proved particularly effective in providing reassurance, reducing cognitive load, and enhancing comprehension of procedural instructions. This demonstrated that self-directed dialogue supports both emotional regulation and task performance, highlighting the scalability and practical relevance of inner speech mechanisms beyond controlled experimental settings.

Across all three studies, inner speech consistently emerged as a unifying mechanism that simultaneously enhances internal cognitive-emotional processing and external social transparency. It enables robots to structure self-reflection, manage emotional responses, and guide internal attention, while also providing human partners with insight into the robot's reasoning and affective state, thereby increasing trust, collaboration quality, and psychological safety in complex interactions. The research further demonstrates that integrating inner speech with computational emotion models enables robots to exhibit behaviour that is not only functionally robust but also emotionally intelligible, socially meaningful, and aligned with human expectations. The findings contribute both theoretically and practically to Affective and Cognitive Robotics: they demonstrate that emotion generation in artificial agents benefits from structured self-directed dialogue rather than superficial expressive cues, and they underscore the potential of such architectures to improve outcomes in sensitive, real-world tasks. Beyond validating the computational and experiential models, this research emphasises the importance of designing robots capable of reflective, emotionally informed behaviour that can adapt to dynamic contexts and support human collaborators under stress. In addition, by bridging appraisal-based evaluation, embodied-cognitive emotional dynamics, and applied validation, this work provides a coherent framework that unites theory, computational modelling, and real-world application, suggesting pathways for future research in which robots might engage in even richer forms of inner speech, enhance adaptive decision-making, and contribute to emotionally attuned, high-stakes human-robot teams across diverse domains. The doctoral studies establish inner speech as a foundational cognitive mechanism that enables robots to function as socially and emotionally competent agents, capable of supporting human partners in both routine and critical collaborative tasks, and provide a roadmap for the development of the next generation of socially intelligent, emotionally expressive robotic systems.

Bibliography

[1] Georgios Paltoglou and Michael Thelwall. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4(1):116–123, 2013.

[2] Daniel Gregory. Inner Speech: New Voices. *Analysis*, 80(1):164–173, 01 2020.

[3] Peter Langland-Hassan and Agustín Vicente. *Inner speech: New voices*. Oxford University Press, USA, 2018.

[4] M. Perrone-Bertolotti, L. Rapin, J.-P. Lachaux, M. Baciu, and H. Lœvenbruck. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, 261:220–239, 2014.

[5] Alain Morin. Inner speech. *Oxford Scholarship Online*, 2018.

[6] Lev Vygotsky. Théorie des émotions: étude historico-psychologique. *Théorie des Émotions*, pages 1–416, 1998.

[7] Pablo Fossa, Raymond Madrigal Pérez, and Camila Muñoz Marcotti. The relationship between the inner speech and emotions: Revisiting the study of passions in psychology. *Human Arenas*, 3(2):229–246, 2020.

[8] Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

[9] Richard S Lazarus. Cognition and motivation in emotion. *American psychologist*, 46(4):352, 1991.

[10] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[11] Sazzad Hussain, Sidney D'Mello, and Rafael Calvo. Research and development tools in affective computing. 12 2014.

[12] C. Bartneck. Integrating the occ model of emotions in embodied characters. 2002.

[13] H. Prendinger and M. Ishizuka. Human physiology as a basis for designing and evaluating affective communication with life-like characters. *IEICE Trans. Inf. Syst.*, 88-D:2453–2460, 2005.

[14] Perry D. Klein. Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, 11:203–270, 1999.

[15] Timo Partala and Veikko Surakka. The effects of affective interventions in human-computer interaction. *Interact. Comput.*, 16:295–309, 2004.

[16] Antonio Chella and Arianna Pipitone. A cognitive architecture for inner speech. *Cognitive Systems Research*, 59:287–292, 2020.

[17] Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7:16, 2020.

[18] Arianna Pipitone and Antonio Chella. What robots want? hearing the inner voice of a robot. *Iscience*, 24(4):102371, 2021.

[19] Alessandro Geraci, Antonella D'Amico, Arianna Pipitone, Valeria Seidita, and Antonio Chella. Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions. *Frontiers in Robotics and AI*, 8:66, 2021.

[20] James J. Gross. *Handbook of Emotion Regulation (2nd ed)*. New York, NY, US: Guilford Press, 2014.

[21] Meinrad Perrez and Michael Reicherts. *Stress, coping, and health: A situation-behavior approach: Theory, methods, applications*. 01 1992.

[22] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[23] Jonathan Gratch and Stacy 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, August 2004.

[24] Stacy C. Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009. Modeling the Cognitive Antecedents and Consequences of Emotion.

[25] Richard S Lazarus. Psychological stress and the coping process. 1966.

[26] Gerald L Clore and Andrew Ortony. Appraisal theories: How cognition shapes affect into emotion. 2008.

[27] Nico H Frijda. *The laws of emotion*. Psychology Press, 2017.

[28] David Irons. Prof. james' theory of emotion. *Mind*, 3(9):77–97, 1894.

[29] Klaus R. Scherer. *Appraisal Theory*, chapter 30, pages 637–663. Wiley-Blackwell, 2005.

[30] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[31] Paula M Niedenthal, Lawrence W Barsalou, Piotr Winkielman, Silvia Krauth-Gruber, and François Ric. Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, 9(3):184–211, 2005.

[32] Alain Morin and Breanne Hamper. Self-reflection and the inner voice: activation of the left inferior frontal gyrus during perceptual and conceptual self-referential thinking. *The open neuroimaging journal*, 6:78, 2012.

[33] Margherita Dahò and Dario Monzani. The multifaceted nature of inner speech: Phenomenology, neural correlates, and implications for aphasia and psychopathology. *Cognitive Neuropsychology*, 42:1 – 21, 2025.

[34] Charles Fernyhough and A. Borghi. Inner speech as language process and cognitive tool. *Trends in cognitive sciences*, 2023.

[35] Wade Munroe. Thinking through talking to yourself: Inner speech as a vehicle of conscious reasoning. *Philosophical Psychology*, 36:292 – 318, 2022.

[36] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[37] Luc Steels et al. Language re-entrance and the 'inner voice'. *Journal of Consciousness Studies*, 10(4-5):173–185, 2003.

[38] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

[39] Richard Savery and Gil 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)*, pages 986–993, 2020.

[40] Riccardo Gervasi, F. Barravecchia, L. Mastrogiacomo, and F. Franceschini. Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237:815 – 832, 2022.

[41] L. Li and Zeang Zhao. Designing behaviors of robots based on the artificial emotion expression method in human–robot interactions. *Machines*, 2023.

[42] SuJin Jo and Seongsoo 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*, 13:21367–21381, 2025.

[43] Xiaozhen Liu, Jiayuan Dong, and Myounghoon Jeon. Robots' "woohoo" and "argh" can enhance users' emotional and social perceptions: An exploratory study on non-lexical vocalizations and non-linguistic sounds. *ACM Transactions on Human-Robot Interaction*, 12:1 – 20, 2023.

[44] Nguyen Tan Viet Tuyen, A. Elibol, and N. Chong. Learning bodily expression of emotion for social robots through human interaction. *IEEE Transactions on Cognitive and Developmental Systems*, 13:16–30, 2020.

[45] Marlena R. Fraune, Benjamin C. Oistad, Catherine E. Sembroski, Kathryn A. Gates, M. M. Krupp, and S. Šabanović. Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact. *Comput. Hum. Behav.*, 105:106220, 2020.

[46] A. Kouroupa, K. Laws, K. Irvine, S. Mengoni, A. Baird, and Shivani 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.

[47] Nicole L. Robinson, T. Cottier, and D. Kavanagh. Psychosocial health interventions by social robots: Systematic review of randomized controlled trials. *Journal of Medical Internet Research*, 21, 2019.

[48] D. Logan, C. Breazeal, M. Goodwin, Sooyeon Jeong, Brianna O'Connell, Duncan Smith-Freedman, James A. J. Heathers, and Peter Weinstock. Social robots for hospitalized children. *Pediatrics*, 144, 2019.

[49] Andreas K Triantafyllidis, Anastasios Alexiadis, K. Votis, and D. Tzovaras. Social robot interventions for child healthcare: A systematic review of the literature. *Computer Methods and Programs in Biomedicine Update*, 2023.

[50] Ling Tan, Cuiqiao Liu, Yongli Wang, Ya Li, Jie Zhao, Shuchun Wang, and Bixin Zhong. The effect of human-robot collaboration on frontline employees' service performance: A resource perspective. *International Journal of Hospitality Management*, 2025.

[51] P. Weis and C. Herbert. Do i still like myself? human-robot collaboration entails emotional consequences. *Comput. Hum. Behav.*, 127:107060, 2021.

[52] Felix Jimenez, T. Yoshikawa, T. Furuhashi, and Masayoshi Kanoh. An emotional expression model for educational-support robots. *Journal of Artificial Intelligence and Soft Computing Research*, 5:51 – 57, 2015.

[53] B. Littler, Tourkiah Alessa, P. Dimitri, Christine Smith, and L. D. de Witte. Reducing negative emotions in children using social robots: systematic review. *Archives of Disease in Childhood*, 106:1095 – 1101, 2021.

[54] Margaret J Trost, Adam R. Ford, Lynn Kysh, J. Gold, and M. Matarić. Socially assistive robots for helping pediatric distress and pain: A review of current evidence and recommendations for future research and practice. *The Clinical Journal of Pain*, 35:451–458, 2019.

[55] Jaime Andres Rincon Arango, Cédric Marco-Detchart, and Vicente Javier Julian Inglada. Personalized cognitive support via social robots. *Sensors (Basel, Switzerland)*, 25, 2025.

[56] Hojjat Abdollahi, M. Mahoor, Rohola Zandie, Jarid Siewierski, and S. Qualls. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, 14:2020–2032, 2022.

[57] Jui Hua Lui, Hooman Samani, and Kun-Yu Tien. An affective mood booster robot based on emotional processing unit. In *2017 International Automatic Control Conference (CACS)*, pages 1–6, 2017.

[58] Agnese Augello. Unveiling the reasoning processes of robots through introspective dialogues in a storytelling system: A study on the elicited empathy. *Cognitive Systems Research*, 2022.

[59] Sophia Corvaia, Arianna Pipitone, and Antonio Chella. Inner speech and damasio's theory for modelling robot's emotions. *IEEE Transactions on Affective Computing*, pages 1–14, 2025.

[60] Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. An appraisal-based chain-of-emotion architecture for affective language model game agents. *PLOS ONE*, 19, 2023.

[61] Yoichiro Maeda. Human-robot interaction experiment based on markovian emotional model. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2018.

[62] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *ArXiv*, abs/2006.00093, 2020.

[63] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3), 2022.

[64] Arun Rai. Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, 48, 12 2019.

[65] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51, 02 2018.

[66] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[67] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:145, 12 2020.

[68] Ruth Stock-Homburg. Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *International Journal of Social Robotics*, 14, 06 2021.

[69] Antonio R Damasio. *The Strange Order of Being: Self-Awareness and Its Origins in Social Neuroscience*. Pantheon Books, 2021.

[70] Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of Feelings*. Houghton Mifflin Harcourt, 2017.

[71] Heidrich Vicci. Emotional intelligence in artificial intelligence: A review and evaluation study. *SSRN Electronic Journal*, 05 2024.

[72] Francesca M.M. Citron, Marcus A. Gray, Hugo D. Critchley, Brendan S. Weekes, and Evelyn C. Ferstl. Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, 56:79–89, 2014.

[73] Gyanendra Verma and Uma Shanker Tiwary. Affect representation and recognition in 3d continuous valence-arousal-dominance space. *Multimedia Tools and Applications*, 76, 01 2017.

[74] Sanghoon Jun, Seungmin Rho, Byeong-jun Han, and Eenjun Hwang. A fuzzy inference-based music emotion recognition system. pages 673 – 677, 09 2008.

[75] Agnes Moors, Phoebe Ellsworth, Klaus Scherer, and Nico Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5:119–124, 03 2013.

[76] William James. What is an emotion? *Mind*, 9(34):188–205, 1884.

[77] Walter B. Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124, 1927.

[78] A. R. Damasio. *Descartes' error: Emotion, reason, and the human brain*. Avon, New York, 1994.

[79] A. R. Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos Trans R Soc Lond B Biol Sci*, 351(1346):1413–1420, October 1996.

[80] Enrique Osuna, Luis-Felipe Rodríguez, J. Octavio Gutierrez-Garcia, and Luis A. Castro. Development of computational models of emotions: A software engineering perspective. *Cognitive Systems Research*, 60:1–19, 2020.

[81] Jiayi Eurus Zhang, Bernhard Hilpert, Joost Broekens, and Jussi P. P. Jokinen. Simulating emotions with an integrated computational model of appraisal and reinforcement learning. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[82] Desmond C. Ong, Jamil Zaki, and Noah D. Goodman. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science*, 11(2):338–357, April 2019. Epub 2018 Jul 31.

[83] Alain Morin. Self-awareness deficits following loss of inner speech: Dr. jill bolte taylor's case study. *Consciousness and Cognition*, 18(2):524–529, 2009.

[84] Arianna Pipitone, Alessandro Geraci, Antonella D'Amico, Valeria Seidita, and Antonio Chella. Robot's inner speech effects on human trust and anthropomorphism. *International Journal of Social Robotics*, pages 1 – 13, 2023.

[85] Arianna Pipitone and Antonio Chella. Robot passes the mirror test by inner speech. *Robotics and Autonomous Systems*, 144:103838, 2021.

[86] Ben Alderson-Day and Charles Fernyhough. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141:931 – 965, 2015.

[87] Pablo Fossa, Raymond Madrigal Perez, and Camila Muñoz. The relationship between the inner speech and emotions: Revisiting the study of passions in psychology. *Human Arenas*, 3, 06 2020.

[88] Márta Gácsi, Anna Kis, Tamás Faragó, Mariusz Janiak, Robert Muszyński, and Ádám Miklósi. Humans attribute emotions to a robot that shows simple behavioural patterns borrowed from dog behaviour. *Computers in Human Behavior*, 59:411–419, 2016.

[89] Sophia Corvaia, Arianna Pipitone, Angelo Cangelosi, and Antonio Chella. Inner speech and extended consciousness: a model based on damasio's theory of emotions. *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8, 2023.

[90] A. Morin. Inner speech. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, pages 436–443. Academic Press, San Diego, second edition edition, 2012.

[91] A Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[92] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, 38(4):13, December 2017.

[93] A. R. Damasio. *The feeling of what happens : body and emotion in the making of consciousness.* Harcourt Brace, New York, 1999.

[94] Marleen Rozemond. *Descartes?s Dualism*. Harvard University Press, Cambridge, Mass., 1998.

[95] Tasha Poppa and Antoine Bechara. The somatic marker hypothesis: revisiting the role of the 'body-loop' in decision-making. *Current Opinion in Behavioral Sciences*, 19:61–66, 2018. Emotion-cognition interactions.

[96] Richard S. Lazarus. From psychological stress to the emotions: a history of changing outlooks. *Annual review of psychology*, 44:1–21, 1993.

[97] Patrick Krauss and Andreas Maier. Will we ever have conscious machines? 03 2020.

[98] Tibor Bosse, Catholijn M. Jonker, and Jan Treur. Formalisation of damasio's theory of emotion, feeling and core consciousness. *Consciousness and Cognition*, 17(1):94–113, 2008.

[99] Catholijn Jonker and Jan Treur. Compositional verification of multi-agent systems: A formal analysis of pro-activeness and reactiveness. *International Journal of Cooperative Information Systems*, 11:51–92, 01 2002.

[100] Lauri Nummenmaa, Enrico Glerean, Riitta Hari, and Jari K. Hietanen. Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111(2):646–651, 2014.

[101] Lauri Nummenmaa, Riitta Hari, Jari K. Hietanen, and Enrico Glerean. Maps of subjective feelings. *Proceedings of the National Academy of Sciences*, 115(37):9198–9203, 2018.

[102] Paul Ekman. Basic emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley, 1999.

[103] Federico Manzi, Davide Massaro, Daniele Di Lernia, Mario Maggioni, Giuseppe Riva, and Antonella Marchetti. Robots are not all the same: Young adults' expectations, attitudes, and mental attribution to two humanoid social robots. *Cyberpsychology, Behavior, and Social Networking*, 24, 11 2020.

[104] Ronit Feingold Polak, Avital Elishay, Yonat Shachar, Maayan Stein, Yael Edan, and Shelly Levy Tzedek. Differences between young and old users when interacting with a humanoid robot: A qualitative usability study. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 107–108, New York, NY, USA, 2018. Association for Computing Machinery.

[105] Stacy Marsella, Jonathan Gratch, and P. Petta. Computational models of emotion. *A Blueprint for Affective Computing-A Sourcebook and Manual*, pages 21–46, 01 2010.

[106] Suman Ojha, Jonathan Vitale, and Mary-Anne Williams. Computational emotion models: A thematic review. *International Journal of Social Robotics*, 13, 09 2021.

[107] Cynthia Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, 2003.

[108] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[109] Lola Canamero. Emotion understanding from the perspective of autonomous robots research. *Neural Networks*, 18(4):445–455, 2005.

[110] Lorenzo Cominelli, Daniele Mazzei, and Danilo De Rossi. Seai: Social emotional artificial intelligence based on damasio's theory of mind. *Frontiers in Robotics and AI*, 5, 02 2018.

[111] Francesco Venturini, Carlo Mazzola, and Massimo Marassi. A specific role for damasio's somatic markers in artificial decision-making: advantages and potentials for future implementations. 01 2021.

[112] Mark Hoogendoorn, Robbert-Jan Merk, and Jan Treur. A decision making model based on damasio's somatic marker hypothesis. *Artificial Intelligence*, 1(2), 1994.

[113] Pablo Barros and et al. A personalized affective memory model for improved emotion recognition in human-robot interaction. *International Journal of Social Robotics*, 10(1):37–50, 2018.

[114] Christian Becker-Asano and Ipke Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, 2010.

[115] Jesse Prinz. Embodied emotions. In Robert C. Solomon, editor, *Thinking About Feeling: Contemporary Philosophers on Emotions*. Oup Usa, 2004.

[116] Fumihide Tanaka and Shizuko Matsuzoe. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *J. Hum.-Robot Interact.*, 1(1):78–95, jul 2012.

[117] Antonio Chella and Arianna Pipitone. A cognitive architecture for inner speech. *Cognitive Systems Research*, 59:287–292, 2020.

[118] Arianna Pipitone and Antonio Chella. What robots want? hearing the inner voice of a robot. *Iscience*, 24(4):102371, 2021.

[119] Arianna Pipitone, Alessandro Geraci, Antonella D'Amico, Valeria Seidita, and Antonio Chella. Robot's inner speech effects on trust and anthropomorphic cues in human-robot cooperation. *arXiv preprint arXiv:2109.09388*, 2021.

[120] Arianna Pipitone, Irene Seidita, John Sullins, and Antonio Chella. Unlocking practical wisdom through the inner voice of robots. *Scientific Reports*, 15, 01 2025.

[121] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science Robotics*, 5(32):eaaz5954, 2020.

[122] Hansol Woo, Gerald K. LeTendre, Trang Pham-Shouse, and Yuhan Xiong. The use of social robots in classrooms: A review of field-based studies. *Educational Research Review*, 33:100388, 2021.

[123] Shih-Ting Chu, Gwo-Jen Hwang, and Yun-Fang Tu. Artificial intelligence-based robots in education: A systematic review of selected ssci publications. *Computers and Education: Artificial Intelligence*, 3:100091, 2022.

[124] John R Anderson and Christian J Lebiere. *The atomic components of thought*. Psychology Press, 2014.

[125] John E Laird, Allen Newell, and Paul S Rosenbloom. Soar: An architecture for general intelligence. *Artificial intelligence*, 33(1):1–64, 1987.

[126] John E Laird. *The Soar cognitive architecture*. MIT press, 2019.

[127] Pat Langley and Dongkyu Choi. A unified cognitive architecture for physical agents. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1469. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[128] A Pipitone, A Geraci, A D'Amico, V Seidita, and A Chella. Robot's inner speech effects on human trust and anthropomorphism. *International Journal of Social Robotics*, 16(6):1333–1345, 2024.

[129] Arianna Pipitone, Irene Seidita, John P Sullins, and Antonio Chella. Unlocking practical wisdom through the inner voice of robots. *Scientific Reports*, 15(1):2634, 2025.

[130] A. Augello, G. Città, M. Gentile, and et al. A storytelling robot managing persuasive and ethical stances via act-r: An exploratory study. *International Journal of Social Robotics*, 15:2115–2131, 2023.

[131] Samuele Vinanzi and Angelo Cangelosi. Casper: Cognitive architecture for social perception and engagement in robots. *International Journal of Social Robotics*, pages 1–19, 2024.

[132] Luca Garello, Giulia Belgiovine, Gabriele Russo, Francesco Rea, and Alessandra Sciutti. Building knowledge from interactions: An llm-based architecture for adaptive tutoring and social reasoning. *arXiv preprint arXiv:2504.01588*, 2025.

[133] Diana Saplacan, Jo Herstad, Jim Tørresen, and Zada Pajalic. A framework on division of work tasks between humans and robots in the home. *Multimodal Technologies and Interaction*, 4(3), 2020.

[134] Brian Armstrong, Oussama Khatib, and Joel Burdick. The explicit dynamic model and inertial parameters of the puma 560 arm. In *Proceedings. 1986 IEEE international conference on robotics and automation*, volume 3, pages 510–518. IEEE, 1986.

[135] SJ Harris, F Arambula-Cosio, Q Mei, RD Hibberd, BL Davies, JEA Wickham, MS Nathan, and B Kundu. The probot- an active robot for prostate resection. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 211(4):317–325, 1997.

[136] William L Bargar, André Bauer, and Martin Börner. Primary and revision total hip replacement using the robodoc (r) system. *Clinical Orthopaedics and Related Research (1976-2007)*, 354:82–91, 1998.

[137] Valeria Seidita, Francesco Lanza, Arianna Pipitone, and Antonio Chella. Robots as intelligent assistants to face covid-19 pandemic. *Briefings in Bioinformatics*, 22(2):823–831, 2021.

[138] Arshia Khan and Yumna Anwar. Robots in healthcare: A survey. In Kohei Arai and Supriya Kapoor, editors, *Advances in Computer Vision*, pages 280–292, Cham, 2020. Springer International Publishing.

[139] B L Davies, R D Hibberd, W S Ng, A G Timoney, and J E A Wickham. The development of a surgeon robot for prostatectomies. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 205(1):35–38, 1991.

[140] Y. S. Kwoh and et al. A robot with improved absolute positioning accuracy for ct guided stereotactic brain surgery. *IEEE Transactions on Biomedical Engineering*, 1988.

[141] Padmanabhan Subramanian, Tom W Wainwright, Shayan Bahadori, and Robert G Middleton. A review of the evolution of robotic-assisted total hip arthroplasty. *HIP International*, 29(3):232–238, 2019.

[142] AR Lanfranco, Andres Castellanos, JP Desai, and William Meyers. Robotic surgery - a current perspective. *Annals of Surgery*, 239:14–21, 01 2004.

[143] Roxana Ramos-Carpinteyro, Ethan Ferguson, Jaya Chavali, Albert Geskin, Nicolas Soputro, and Jihad Kaouk. Single-port transvesical robot-assisted radical prostatectomy: The surgical learning curve of the first 100 cases. *Urology*, 178, 06 2023.

[144] Maria Consuelo Puentes, Marko Rojnica, Thomas Sims, Robert Jones, Francesco M. Bianco, and Thom E. Lobe. Senhance robotic platform in pediatrics: Early us experience. *Children*, 10(2), 2023.

[145] Ibrahim Alkatout, Hamid Salehiniya, and Leila Allahqoli. Assessment of the versius robotic surgical system in minimal access surgery: A systematic review. *Journal of Clinical Medicine*, 11(13), 2022.

[146] Xin Chen, Shu Deng, Mao-Lin Sun, and Rui He. Robotic arm-assisted arthroplasty: The latest developments. *Chinese Journal of Traumatology*, 25, 09 2021.

[147] Cécile Batailler, Andrea Fernandez, John Swan, Elvire Servien, Fares S Haddad, Fabio Catani, and Sébastien Lustig. Mako ct-based robotic arm-assisted system is a reliable procedure for total knee arthroplasty: a systematic review. *Knee Surgery, Sports Traumatology, Arthroscopy*, 29(11):3585–3598, 2021.

[148] Béni-Trésor Akimana, Maxim Bonnaerens, Jonas Wilder, and Bjorn Vuylsteker. A survey of human-robot interaction in the internet of things. 12 2016.

[149] Ki Nam, Hyun Jung Kim, Bum Kwon, Jin-Woo Park, Ho Jun Lee, and Aeri Yoo. Robot-assisted gait training (lokomat) improves walking function and activity in

people with spinal cord injury: a systematic review. *Journal of NeuroEngineering and Rehabilitation*, 14, 03 2017.

[150] Jane Holland, Liz Kingston, Conor Mccarthy, Eddie Armstrong, Peter O'dwyer, Fionn Merz, and Mark Mcconnell. Service robots in the healthcare sector. *Robotics*, 10:47, 03 2021.

[151] Anna Kochan. Scalpel please, robot: Penelope's debut in the operating room. *Industrial Robot: An International Journal*, 32:449–451, 12 2005.

[152] Mithun Jacob, Yu-Ting Li, and A.George Akingba. Gestonurse: A robotic surgical nurse for handling surgical instruments in the operating room. *Journal of Robotic Surgery*, 6, 03 2011.

[153] Diana Saplacan, Trenton Schulz, Jim Torresen, and Zada Pajalic. Health professionals' views on the use of social robots with vulnerable users: A scenario-based qualitative study using story dialogue method. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 421–428, 2023.

[154] Ben Alderson-Day and Charles Fernyhough. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5):931, 2015.

[155] Donald G MacKay. Constraints on theories of inner speech. *Auditory imagery*, pages 133–162, 2014.

[156] Alain Morin. Inner speech. 2009.

[157] Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7:16, 2020.

[158] Global Education Monitoring Report Team UNESCO Institute for Statistics. *SDG 4 data digest 2021: national SDG 4 benchmarks: fulfilling our neglected commitment*. 2021.

[159] R. Soma et al. Strengthening human autonomy in the era of autonomous technology. *Scandinavian Journal of Information Systems*, 34(2):Article 5, 2022.

[160] Diana Saplacan, Weria Khaksar, and Jim Torresen. On ethical challenges raised by care robots: A review of the existing regulatory-, theoretical-, and research gaps. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 219–226, 2021.

[161] Jim Torresen, Diana Saplacan, Adel Baselizadeh, and Tobias Mahler. Machine excellence tradeoffs to ethical and legal perspectives. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 237–240, 2023.

[162] M. van Otterdijk, B. Laeng, D. Saplacan-Lindblom, et al. Seeing meaning: How congruent robot speech and gestures impact human intuitive understanding of robot intentions. *International Journal of Social Robotics*, 2025.