

Towards Trustworthy and Explainable Socially Assistive Robots: A Cognitive Architecture for Dietary Guidance

Luigi D’Arco^{†*}, Luca Raggioli[†], Gianni Randazzo[‡], Giovanni De Gasperis[§],
Antonio Chella[‡], Stefania Costantini[§] and Silvia Rossi[†]

[†]Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Naples, Italy
email: {luigi.darco; luca.raggioli; silvia.rossi}@unina.it

[‡]Department of Engineering, University of Palermo, Palermo, Italy
email: gianni.randazzo@community.unipa.it; antonio.chella@unipa.it

[§]Department of Information Engineering, Computer Science and Mathematics, Università degli Studi dell’Aquila, L’Aquila, Italy
email: {giovanni.degasperis; stefania.costantini}@univaq.it

*corresponding author

Abstract—Socially Assistive Robots (SARs) are rising as promising tools for promoting healthy lifestyle habits. To achieve such a goal, it is necessary that they are able to perform trustworthy and legible behaviors. In this work, we propose a cognitive architecture that integrates multimodal perception, symbolic reasoning, memory-enhanced decision-making, and adaptive interaction strategies to create an explainable and engaging dietary assistant. The key idea is to provide the robot with the capability to iteratively interact with a user and adapt the dietary plan based on their current state, preferences, and food restrictions, while conveying explicitly the inner decision and thought process. To achieve this, we employ a graph-enhanced Large Language Model (LLM), which queries contextual, semantic, and episodic acquired knowledge to generate personalized meal recommendations. These must subsequently be refined through a verification process that enforces constraints such as caloric limits and ingredient intolerances, ensuring dietary adherence. To have a transparent decision-verification process, the robot has to progressively verbalize the reasoning process while providing justifications for the recommendations to also enhance the user’s trust. Non-verbal context-relevant movements are also generated to allow the robot to express empathy. We expect our framework to increase user trust, engagement, and adherence to healthy behaviors, allowing SARs to function as credible and effective health assistants.

Index Terms—Socially Assistive Robots, Health Management, Human-Robot Interaction, LLM, Symbolic Reasoning

I. INTRODUCTION

Social relationships play a crucial role in shaping human healthcare and well-being. It has been shown that social support and interaction can have a significant impact on individuals’ physical and mental health, as well as on their ability to adopt and maintain healthy lifestyle habits [1]. Multiple studies have highlighted the importance of adapting and personalizing the robot’s behaviors to meet the specific needs of each individual, as this is a key factor in the

success of strategies aimed at promoting healthy lifestyles [2]. Lifestyle interventions often require people to change their habits, usually resulting in non-acceptance of such indications [3]. Only a very low percentage of people who decide to adopt a healthy lifestyle maintain it for long periods or in their daily routine [4]. This might be due to a lack of motivation, support, and guidance that people need to adhere to these changes. In this context, the use of Socially Assistive Robots (SARs) can be quite beneficial for health and well-being applications [5]. Humans tend to anthropomorphize robots, attributing to them human-like qualities [6] that are fundamental in social scenarios, such as communicative capabilities, social cues, and emotional expressions [7]. This is particularly important in the context of health coaching, where the robot needs to establish a relationship with the user, understand their needs and preferences, and provide personalized support and guidance [8]. Achieving this level of personalization, however, requires the robot to comprehend and predict the user’s behavior and intentions. This can be achieved by endowing the robot with Theory of Mind (ToM) capabilities, which allows them to also make their behavior understandable and predictable to humans, thus fostering trust and engagement [9]. Moreover, when considering the general definition of trust, relying on the dependability of someone or something [10], it becomes evident that a robot’s behavior must not only be transparent but also ethically sound and predictable [11]. It should neither foster improper or unethical situations nor develop unforeseen behaviors that could conflict with the user’s expectations [12]. To meet these requirements, verification methods can be integrated into a robot architecture, including both preemptive system validation against predefined configurations and dynamic verification that adapts based on user interaction [13].

In this research, we aim to investigate how to build trustworthy and legible SARs that can promote healthy lifestyle habits

through social interactions. Specifically, we focus on a dietary assistance scenario where the robot provides personalized meal recommendations based on user preferences, dietary constraints, and emotional state. To this aim, we propose a cognitive architecture combining ToM concepts, both from a cognitive and affective perspective, to manipulate and adapt the level of transparency of both explicit (e.g., verbally) and implicit communication (e.g., non-verbal strategies). Among the various approaches, inner speech techniques have been explored as a means to enhance the robot’s behavioral transparency [14]. By making the thought process more interpretable, these techniques help users form a clearer mental model of the robot, ultimately increasing its perceived reliability, robustness, and trustworthiness [15]. However, how to best incorporate ToM mechanisms in SARs to promote legible intentions for lifestyle intervention applications remains an open question. Our model integrates multimodal perception, symbolic reasoning, and adaptive interaction strategies, enabling the robot to understand and respond to the users in a natural and explainable manner. Importantly, we leverage the computational and reasoning power of Large Language Model (LLM) to generate candidate meal suggestions, assist inner speech mechanisms, and interpret the emotional and social context necessary to generate legible robotic behavior and movement patterns. This powerful tool is complemented by symbolic-based real-time explanations of the recommendations to persuade users of the goodness of the proposed plan. This research contributes to the development of SARs that will foster an effective Human-Robot Interaction (HRI), allowing the users to form accurate mental models of the robot. Thereby, SARs will be not only effective in promoting health and well-being strategies but also trustworthy and transparent in their interactions with users, encouraging long-term engagement.

II. RELATED WORK

Developing robots that exhibit intelligent behaviors is a complex challenge that targets the integration of various cognitive capabilities, including perception, reasoning, learning, and decision-making. Cognitive architectures serve as structured frameworks that model human cognitive processes, enabling robots to interact dynamically with their environments [16]. By simulating human-like cognition, these architectures enhance robotic adaptability, versatility, and efficiency across diverse applications, such as healthcare, education, and service industries [17].

Over the years, several cognitive architectures have been introduced, each specializing in different aspects of cognition [18]. For instance, ACT-R [19] emphasized the interaction between declarative and procedural memory, facilitating learning and problem-solving. Meanwhile, SOAR [20] focused on integrating perception, reasoning, and action to support adaptive decision-making in complex scenarios. Despite their advantages, implementing cognitive architectures in social robots presents significant computational challenges, particularly for real-time applications. To mitigate these constraints, researchers have explored simplified cognitive models that tar-

get specific cognitive functions, such as perception, reasoning, or action, as well as investigating and analyzing frameworks and agent-based architectures in simulated environments [21]. Infantino et al. [22] proposed a cognitive architecture that exploited a perception-reasoning-action loop that integrated the robot’s self-perception, the perception of the environment, and the perception of the user’s mental state. The implication of a long-term memory trained by an external agent using an interactive genetic algorithm allowed the introduction of prior knowledge of action for the robot, while the use of self-organizing maps enabled the creation of a working memory to encode the social context. Tanevska et al. [23] introduced a simplified cognitive architecture designed for practical implementation in robots. Their framework consisted of four key modules: perception, action, adaptation, and management. In particular, the adaptation module regulated the robot’s social needs, ensuring more dynamic interactions. Experimental evaluations revealed that naive users preferred interacting with adaptive robots, as they exhibited greater responsiveness and engagement. Robots as a classical Belief-Desire-Intention model of agency has been proposed in [24]. The proposed architecture, applied in healthcare settings, aimed to orchestrate monitoring, knowledge management, and deliberation modules. More recently, Vinanzi and Cangelosi [25] developed a platform-independent cognitive architecture that integrates both symbolic and data-driven artificial intelligence to enhance robot-agent collaboration. Their framework was designed to infer the user’s possible goals, actions, and mental states, enabling more intuitive and adaptive human-robot interactions. The architecture combined qualitative spatial relations between the agent and objects of interest in the scene with a non-binary tree-based plan library to assess and rank possible courses of action, allowing the robot to generate contextually appropriate responses. Additionally, an ontology-based strategy was implemented to verify relationships between entities and the agent’s actions, ensuring consistency and coherence in decision-making.

Despite these advancements, existing cognitive architectures often rely on predefined behaviors, which can lead to unnatural or repetitive interactions over extended periods. Additionally, many of these systems function as black boxes, making it difficult for users to infer the robot’s mental state, thereby limiting trust and user engagement. Furthermore, current frameworks lack integrated verification mechanisms to assess and adapt behaviors based on user needs, potentially leading to inappropriate or suboptimal interactions.

III. PROPOSED FRAMEWORK

The proposed framework aims to develop an artificial cognitive architecture that will allow a social robot to understand user needs and work as a recommender system for health coaching. By integrating verbal and non-verbal behavior, the framework aims to guarantee transparency in its decision process, endowing the robot with the capability to communicate its intentions effectively, thus allowing the user to attribute a mental model to the robot. The architecture will

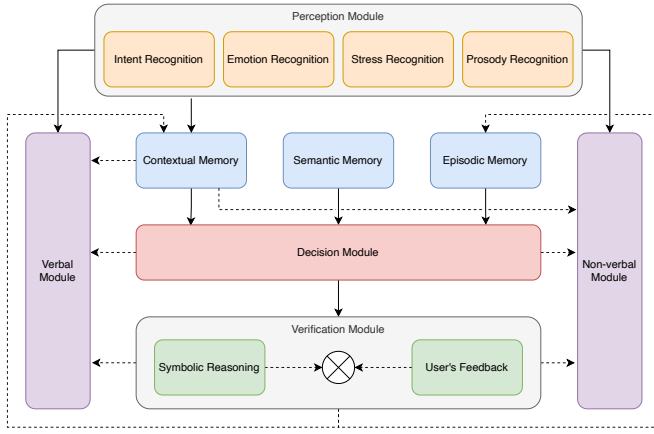


Fig. 1. The proposed Cognitive Architecture consists of five main parts: 1) a Perception Module that estimates the user’s current state; 2) the Memories that store contextual, semantic, and episodic data in a symbolic and structured form; 3) a Decision Module which acts upon the content of the memories and the perceived state of the user; 4) the Verification Module evaluating the Decision Module’s resolutions; 5) the Verbal and Non-Verbal Modules act as actuators of the Architecture, by formalizing the behavior and spoken interaction of the robot.

also incorporate verification mechanisms to assess and adapt behaviors based on user needs and feedback, ensuring that the robot’s actions align with user expectations and preferences. Below, we will discuss the intended scenario for our work, then we will proceed to describe the main components of our architecture in detail. Figure 1 presents the overall cognitive architecture.

A. Use Case Scenario and Architecture Workflow

A use case scenario has been identified for this first iteration, where a social robot has to support post-surgery patients in their daily lives. The primary objective of this approach is to assist patients in maintaining a healthy diet according to a doctor-prescribed plan. Within this envisioned scenario, the patient periodically interacts with the robot to receive tailored dietary recommendations and can request modifications as needed. The robot, starting with a doctor-recommended nutritional framework, dynamically suggests specific meals while considering the user’s emotional state and dietary restrictions (e.g., lactose intolerance) and prior interactions. By analyzing past behaviors and preferences, the system continuously refines its recommendations to enhance user adherence. Beyond meal planning, the robot functions as a persuasive agent, employing both verbal and non-verbal communication techniques to encourage the patient to follow the recommended diet. This persuasive capability ensures that the dietary guidance is not only personalized but also effectively communicated to promote long-term adherence and well-being.

The Proposed Cognitive Architecture is governed by an iterative process: through the Perception Module, the robot produces an estimate of the user’s state, which is passed to the Contextual Memory and the Verbal and Non-Verbal stimuli production Modules. Based on the content of the three

memories, which are queried with an LLM, the Decision Module determines how the interaction and the choice of the next recipe should be affected. By reasoning over these memory structures, the LLM generates a list of candidate meal suggestions. The Verification inspects the plan to refine the short-listed recipes by enforcing soft and hard constraints such as caloric limits and ingredient intolerances; a verification module based on Answer Set Programming (ASP) is employed. The Verification’s evaluation will also be used by the Context Memory in the following iteration. Decision and Verification module estimates are subsequently used by the Verbal and Non-Verbal modules, together with the current context, to produce a response for the user. In particular, the Verbal Module leverages an LLM-driven inner speech mechanism, which progressively verbalizes the system’s internal deliberation, ensuring that users can follow the reasoning behind its suggestions. Additionally, an explainability component provides structured, real-time explanations of the recommendations to persuade users to adhere to the dietary plan. The non-verbal module further enhances interaction by using an LLM to interpret the emotional and social context and generate legible robotic behavior and movement patterns from a predefined set of motion functions.

B. Perception Module

The Perception Module produces a quantitative representation of the context, user’s state, and intentions. This module is responsible for processing multimodal sensory inputs, including visual and auditory data, through a combination of specialized submodules designed to assess different aspects of the user’s condition:

- 1) Intent Recognition: currently, it is based on explicit user requests. However, future iterations will integrate gaze tracking, voice sentiment analysis, and other behavioral cues to infer user intent implicitly, enhancing the system’s ability to anticipate needs proactively.
- 2) Emotion Recognition: it analyzes facial expressions to determine the user’s emotional state, leveraging the DeepFace library¹ [26] to classify emotions such as happiness, sadness, anger, and fear.
- 3) Stress Recognition: human’s facial expressions can be indicative of stress levels. The stress recognition process unfolds in two steps:
 - Emotions extracted from the emotion recognition submodule are categorized as stressed, if one of anger, fear, or sadness occurred, while other emotions are categorized as not-stressed;
 - An eyebrow movement model², inspired by Gianakakis et al. [27], calculates a stress confidence value based on eyebrow contraction, movement, and the inter-eyebrow distance. The estimation is based on an exponential function, normalized between 1

¹<https://github.com/serengil/deepface>

²<https://github.com/Geek-ubaid/Stress-Detection>

and 100, with a threshold of 60 used to mitigate false positives.

- 4) **Prosody Recognition:** using the My-Voice-Analysis³ library the speech patterns are analyzed, to identify the user’s mood (neutral, calm, or pacy), speech rate, energy, frequency, average speech interval duration, and speaking duration. Additionally, the user’s speech is analyzed with the Azure Cognitive Services Text Analytics libraries⁴ to identify any sentiments conveyed [28], by labeling the speech-to-text as positive, negative, or neutral.

The processed information from these components is stored in the contextual memory in the form of a knowledge graph. This structure allows for continuous expansion and refinement of the user preferences, emotional states, and behavioral patterns, enhancing the robot’s adaptability over time.

C. Memories

The proposed framework adopts a symbolic representation of knowledge to structure the robot’s decision-making process in a manner that aligns with human common sense. This approach leverages both explicit and implicit knowledge formed through interactions with the user. One key advantage of symbolic representation is its ability to facilitate verification and explanation methodologies, which are essential for clarifying the robot’s reasoning and ensuring transparency in decision-making. The memory system has been designed to address three key aspects of knowledge acquisition and preservation. Specifically, it comprises:

- 1) **Contextual Memory:** strongly related to the current interaction with the user. It contains facts that are produced based on the estimations computed by the Perception Module. The formalization of contextual information is fundamental for updating the agent’s beliefs about the users, which might be influenced by their current desires and intentions, as well as their emotional state. In particular, by assessing contextually relevant signals, the model will choose to employ different behavioral strategies and will enforce the dietary constraints with varying willingness to compromise with the user.
- 2) **Episodic Memory:** contains knowledge accumulated during past interactions, which is not only relevant during the interaction in which it occurs but also in future interactions. For instance, let us consider the case in which the user discovers a dislike for a certain ingredient, this information will be relevant for future interactions as any other recipe containing this ingredient might be negatively perceived, thus interfering with the purpose of the application.
- 3) **Semantic Memory:** this part of the memory is populated with persistent knowledge about the recipes, how they are correlated with each other, and how much the user likes them. We formalize the semantic memory with a

knowledge graph, which will contain persistent information (i.e., the ingredients) and variable information that is dependent on the interaction with the user (i.e., emotional reaction to the recipe).

D. Decision Module

The decision module employs a graph-based LLM strategy, wherein an LLM (such as LLAMA [29] with knowledge graph integration) queries the semantic memory, a knowledge graph storing structured representations of recipes, dietary constraints, and user preferences. The reasoning process follows these steps:

- 1) **Knowledge Graph Querying:** The system uses queries to extract a set of recipes aligned with the user’s dietary plan and real-time contextual state.
- 2) **Contextual Filtering:** The retrieved recipes are cross-referenced with perception outputs, to adjust suggestions based on the user’s emotional and stress levels, and previous interactions.
- 3) **Shortlisting via LLM:** The LLM processes structured graph outputs and refines the selection through prompt engineering. The output is a subset of recipes that align with dietary goals and real-time user preferences.

E. Verification Module

The verification module employs declarative logic reasoning to refine the subset of recipes short-listed by the decision module. This ensures that the recommendations align with the user’s dietary restrictions and preferences. The logic programming language chosen is the ASP, which allows for a flexible and expressive representation of complex rules and constraints, enabling the system to reason about possible recipe alternatives while complying with predefined constraints such as calorie limits and ingredient intolerances.

The module evaluates the shortlisted recipes, producing one of three possible outcomes:

- i. an empty set, indicating that all options violate the constraints. In this case, a persuasion strategy employing adaptive communication techniques is used to encourage the user to adhere to the prescribed diet.
- ii. a singleton set, where only one valid alternative remains. It is presented alongside an explanatory rationale to enhance transparency and user trust.
- iii. a collection of feasible options. An interactive selection process is initiated, allowing the user to provide feedback and refine the final choice.

Regardless of the outcome, an explanatory step is included at the end to ensure clarity and reinforce the reasoning behind the system’s recommendation. This structured verification approach not only enhances decision reliability but also fosters a more engaging and user-centric interaction. The verification procedure employed is shown in the Algorithm 1.

F. Verbal Module

The verbal module is responsible for facilitating communication between the robot and the user, providing explanations

³<https://github.com/Shahabks/my-voice-analysis>

⁴<https://azure.microsoft.com/>

Algorithm 1 Verification Mechanism Using ASP

```
% --- Extracted recipes from the decision layer ---
recipe(1, "grilled-chicken-salad", 350).
recipe(2, "beef-steak", 700).
...
recipe(10, "gluten-free-pasta-with-tomato-sauce", 500).

recipe_ingredient(1, butter).
recipe_ingredient(1, chicken).
...
recipe_ingredient(10, tomato).

% --- User constraints ---
max_calories(500).
intolerant_to(lactose).
intolerant_to(gluten).

% --- Ingredients that contain allergens ---
contains_allergen(butter, lactose).
...
contains_allergen(pasta, gluten).

% --- Rules to filter recipes ---
valid_calories(R) :- recipe(R, _, C), max_calories(Max), C
    =< Max.
valid_ingredients(R) :- recipe(R, _, _), not
    contains_restricted_ingredient(R).
contains_restricted_ingredient(R) :- recipe_ingredient(R, I
    ), contains_allergen(I, A), intolerant_to(A).
valid_recipe(R, Name) :- recipe(R, Name, _), valid_calories
    (R), valid_ingredients(R).

% --- Output ---
#show valid_recipe/2.
```

and clarifications to make the robot’s decision-making process more transparent and trustworthy. This module is designed to embody two critical components: explainability and inner speech. This module relies on an LLM (i.e., LLAMA), to produce an answer based on a prompt obtained from the user’s state and the Decision module’s reasoning.

1) *Explainability*: The explainability component of the verbal module is designed to offer real-time, accessible explanations of the robot’s decisions and rationale. It accomplishes this by providing a comprehensive analysis of the decision-making process, highlighting the pivotal factors that led to a particular action or recommendation. This helps the robot improve the human’s trust by demonstrating its transparency.

Example of Explainability Prompt/Response

Prompt: You are a robot assisting a user in maintaining a suitable diet based on their needs and preferences. At this stage, we have already filtered out dishes that do not fit the user’s dietary style. Now, we have generated multiple meal combinations that satisfy the remaining calorie and macronutrient constraints but the chosen one has been the gluten-free pasta with tomato sauce. Your task is to explain to the user why this meal has been selected. The user’s intolerances are lactose and gluten, and the maximum number of calories per meal is 500.

Response: I have selected the Gluten-free Pasta with Tomato sauce because it adheres to your calorie limit of 500 and does not contain any lactose or gluten to which you are intolerant. This meal provides a balanced portion of carbs, aligning with your dietary goals.

2) *Inner Speech*: The inner-speech component’s primary function is to externalize the system’s internal deliberative process by progressively revealing the robot’s thought process before the execution of an action or the formation of a

decision. This helps the user to understand the motivations and constraints that lead to a specific response.

A Progressive Multistage Explanation (PME) method was developed, in which the system generates an initial explanation of the processed data and iteratively refines it when new information is received. This incremental processing cycle facilitates the provision of partial and progressive explanations during HRI, circumventing the protracted wait times associated with comprehensive data processing. Consequently, this enhances the overall interaction experience.

Example of Inner Speech Prompt/Response

Prompt: You are a robot assisting a user in maintaining a suitable diet based on their needs and preferences. At this stage, we are selecting dishes based on the user’s emotional state and according to the defined requirements.

- Emotional state: Happy
- Requirements: no lactose, no gluten
- Possible dishes: chicken nuggets, winter salad

At this point in the process, we have already excluded dishes that do not fit the user’s eating style, and in this step, we have generated several dish combinations that meet the remaining calorie and macronutrient constraints. Given the list of possible dishes, your task is to explain to the user how they were chosen.

Response: You have a lactose and gluten intolerance. I have filtered out any recipes that contain dairy ingredients. After reviewing the options, I can suggest chicken nuggets or a winter salad, as it meets all of your preferences and dietary restrictions.”.

G. Non-Verbal Module

The non-verbal module enables the robot to communicate through human-like movements, based on the context and emotional intent that the robot wishes to express, and the estimates of the Decision and the Verification modules. An LLM identifies which action to perform among a predefined list of movements and animations. The predefined motion functions, which can range from shifting body posture to specific gestures like head tilts or hand movements, according to the type of robot employed, are selected based on these inputs. This ensures that the robot’s non-verbal behavior is contextually relevant, emotionally appropriate, and aligned with the user’s needs, even when the robot lacks traditional human-like expressive capabilities such as facial expressions or arm gestures. The system operates within the framework of Belief-Desire-Intention (BDI) logic, which allows the robot to infer the user’s psychological and cognitive state by analyzing their behavior, emotional cues, and situational context. Furthermore, such an approach allows the integration of a degree of non-determinism in the identification of the best movement to perform, hence showing a more natural behavior in long-term interactions fostering better HRI.

IV. CONCLUSIONS

In this work, we presented a cognitive architecture for SARs that integrates memory-augmented decision-making, logical-based verification, and multimodal communication to foster trustworthy and legible interactions in the domain of healthy lifestyle promotion. The framework enables SARs to generate and validate personalized dietary recommendations, ensuring

adherence to user dietary constraints, such as caloric intake and ingredient intolerances, while enhancing transparency and explainability through inner speech. Furthermore, we devise verbal and non-verbal communication strategies to enhance user engagement, combining progressive explanations with context-aware robotic movements to improve persuasion and adherence.

Future work will focus on expanding the ToM model to allow the robot to infer user mental states more accurately, improving its ability to anticipate user needs and adapt its recommendations dynamically. Additionally, refining dietary constraints beyond simple caloric limits, such as macronutrient distribution (carbohydrates, proteins, fats), will enhance the precision of recommendations. A feasibility analysis will be conducted to investigate the performance, response times, and efficiency of the proposed solution. Another area of investigation involves the integration of multimodal strategies, combining natural language, graphical visualizations, and haptic feedback to make explanations more intuitive and effective.

Finally, to validate these advancements, extensive user studies will be conducted in-home assistance scenarios and dietary coaching to rigorously assess the effectiveness, usability, and psychological impact of the proposed system. To this aim, we are currently collaborating with practitioners to create user profiles to reflect their dietary habits and medical conditions. By integrating these improvements, SARs will become more reliable, persuasive, and socially aware companions, advancing their adoption in healthcare and lifestyle management applications.

ACKNOWLEDGMENT

We would like to thank Professor Valeria Seidita (University of Palermo) for her valuable guidance and support in this research project.

REFERENCES

- [1] D. Umberson and J. Karas Montez, "Social relationships and health: A flashpoint for health policy," *Journal of health and social behavior*, vol. 51, no. 1_suppl, pp. S54–S66, 2010.
- [2] C. Di Napoli, G. Ercolano, and S. Rossi, "Personalized home-care support for the elderly: a field experience with a social robot at home," *User Modeling and User-Adapted Interaction*, vol. 33, no. 2, pp. 405–440, 2023.
- [3] E. Neuvonen, J. Lehtisalo, A. Solomon, R. Antikainen, S. Havulinna, T. Hänninen, T. Laatikainen, J. Lindström, N. Rautio, H. Soininen, et al., "Psychosocial determinants for adherence to a healthy lifestyle and intervention participation in the finger trial: an exploratory analysis of a randomised clinical trial," *Aging Clinical and Experimental Research*, vol. 34, no. 8, pp. 1793–1805, 2022.
- [4] K. R. Middleton, S. D. Anton, and M. G. Perri, "Long-term adherence to health behavior change," *American journal of lifestyle medicine*, vol. 7, no. 6, pp. 395–404, 2013.
- [5] N. L. Robinson, T. V. Cottier, and D. J. Kavanagh, "Psychosocial health interventions by social robots: systematic review of randomized controlled trials," *Journal of medical Internet research*, vol. 21, no. 5, p. e13203, 2019.
- [6] S. Lemaignan, J. Fink, and P. Dillenbourg, "The dynamics of anthropomorphism in robotics," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 226–227, 2014.

- [7] L. D'Arco, A. Rossi, and S. Rossi, "Assessing emotion mitigation through robot facial expressions for human-robot interaction," in *Workshop on Advanced AI Methods and Interfaces for Human-Centered Assistive and Rehabilitation Robotics - AIXIA*, 2024.
- [8] M. M. De Graaf, S. B. Allouch, and T. Klamer, "Sharing a life with harvey: Exploring the acceptance of and relationship-building with a social robot," *Computers in human behavior*, vol. 43, pp. 1–14, 2015.
- [9] C. Castelfranchi, A. Chella, R. Falcone, F. Lanza, and V. Seidita, "Endowing robots with self-modeling abilities for trustful human-robot interactions," *framework*, vol. 2, p. 1, 2019.
- [10] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, no. 4, pp. 297–309, 2020.
- [11] S. Costantini, "Ensuring trustworthy and ethical behaviour in intelligent logical agents," *Journal of Logic and Computation*, vol. 32, no. 2, pp. 443–478, 2022.
- [12] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility: Autonomy trust and responsibility," in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*, pp. 159–168, Springer, 2016.
- [13] R. H. Bordini, M. Fisher, W. Visser, and M. Wooldridge, "Verifying multi-agent programs by model checking," *Autonomous agents and multi-agent systems*, vol. 12, pp. 239–256, 2006.
- [14] A. Chella and A. Pipitone, "A cognitive architecture for inner speech," *Cognitive Systems Research*, vol. 59, pp. 287–292, 2020.
- [15] A. Pipitone and A. Chella, "What robots want? hearing the inner voice of a robot," *Iscience*, vol. 24, no. 4, 2021.
- [16] P. Baxter and J. G. Trafton, "Cognitive architectures for human-robot interaction," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, HRI '14*, (New York, NY, USA), p. 504–505, Association for Computing Machinery, 2014.
- [17] M. Ogunsina, C. P. Efunniyi, O. S. Osundare, S. O. Folorunsho, and L. A. Akwawa, "Cognitive architectures for autonomous robots: Towards human-level autonomy and beyond," *International Journal of Frontline Research in Engineering and Technology*, 2024.
- [18] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: core cognitive abilities and practical applications," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 17–94, 2020.
- [19] J. R. Anderson, *The architecture of cognition*. Psychology Press, 2013.
- [20] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artificial intelligence*, vol. 33, pp. 1–64, 1987.
- [21] V. Seidita and A. Chella, "Enhancing robotic systems in healthcare: A preliminary analysis of agent-based paradigms and simulation environments," in *Proceedings of the CEUR Workshop Proceedings, CEUR-Ws, Bard, Italy*, pp. 8–10, 2024.
- [22] I. Infantino, A. Augello, U. Maniscalco, G. Pilato, and F. Vella, "A cognitive architecture for social robots," in *2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI)*, pp. 1–5, IEEE, 2018.
- [23] A. Tanevska, F. Rea, G. Sandini, L. Cañamero, and A. Sciutti, "A cognitive architecture for socially adaptable robots," in *2019 joint IEEE 9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob)*, pp. 195–200, IEEE, 2019.
- [24] F. Lanza, V. Seidita, and A. Chella, "Agents and robots for collaborating and supporting physicians in healthcare scenarios," *Journal of Biomedical Informatics*, vol. 108, p. 103483, 2020.
- [25] S. Vinanzi and A. Cangelosi, "Casper: Cognitive architecture for social perception and engagement in robots," *International Journal of Social Robotics*, pp. 1–19, 2024.
- [26] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–4, IEEE, 2021.
- [27] G. Giannakakis, M. Padiaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," *Biomedical Signal Processing and Control*, vol. 31, pp. 89–101, 2017.
- [28] M. Atzeni and D. Reforgiato Recupero, "Deep learning and sentiment analysis for human-robot interaction," in *The Semantic Web: ESWC 2018 Satellite Events* (A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, and M. Alam, eds.), (Cham), pp. 14–18, Springer International Publishing, 2018.
- [29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.