AFRL-AFOSR-UK-TR-2023-0072

RESPECT: Robot innEr SPEeCH for Trust

Chella, Antonio
UNIVERSITA' DEGLI STUDI DI PALERMO
PIAZZA MARINA 61
PALERMO, , 90133
ITA

09/27/2023
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20230927	Final	START DATE 20190701	END DATE 20230630

4. TITLE AND SUBTITLE
RESPECT: Robot innEr SPEeCH for Trust

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-19-1-7025	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Antonio Chella

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITA' DEGLI STUDI DI PALERMO
PIAZZA MARINA 61
PALERMO 90133
ITA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2023-0072

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The project RESPECT explores the strategic coupling of cognitive robotics modeling and empirical human-robot interaction experiments to analyze the role of inner speech in developing trustworthy interactions between humans and robots. The research is directly inspired by psychological studies of inner speech in human self-consciousness.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
U	U	U	SAR	255

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANDINI IYER	314-235-6161

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

FA9550-19-1-7025

RESPECT
Robot innEr SPEeCh for Trust

Final Performance Report

Antonio Chella
University of Palermo, Italy
antonio.chella@unipa.it

September 8, 2023

Contents

List of Figures

Abstract

This report describes the activities of the project RESPECT for the strategic coupling of cognitive robotics modeling and empirical human-robot interaction experiments to investigate the role of *robot inner speech* in developing trust interactions between humans and robots.

Chapter 1

Summary of the activities

The project RESPECT explores the strategic coupling of cognitive robotics modeling and empirical human-robot interaction experiments to analyze the role of inner speech in developing trustworthy interactions between humans and robots. The research is directly inspired by psychological studies of inner speech in human self-consciousness.

The team of the project RESPECT delivered invited talks and produced scientific articles listed in Chapter 2. The goals and scope of the RESPECT project, and more generally, the use of inner speech in robots, generated media attention. Chapter 3 describes the main media coverage the project RESPECT received.

Detailed research about the state of the art of the investigations in the relations between robot's inner speech and human trust development has been carried out. In particular, the project team analyzed the literature about the close relationships between inner speech and trustworthy interactions in humans and between humans and robots. The team proposed the inner speech as a new critical feature to be considered in human-robot interactions. The description of the theoretical research is reported in Chapter 4, which outlines and discusses the state-of-the-art bibliography. The important result is that no research work exists in the literature. To the authors, the project represents a completely new research field.

It is acknowledged that inner speech is related to awareness and self-awareness. The team thus presented a theoretical inner speech model described by the event calculus based on first-order modal logic. The inner speech reproduces and expands social and physical sources of awareness. The proposed theoretical model is suitable for a robot from the computational point of view. By making a robot able to talk to itself, it is possible

to analyze the role of inner speech in robot awareness and self-awareness, thus opening new interesting research scenarios not yet investigated. By the typical substitution method, the team demonstrated that the formulas of the proposed calculus lead to an inner speech formalization tightly linked to self-awareness. The event calculus model, the results, and the discussions are presented in Chapter 5. A formalization of some processes at the basis of inner speech is then outlined.

Then, a complete cognitive architecture designed for inner speech is presented. The research took inspiration from psychological studies on inner speech and its relations with human self-consciousness. More in detail, the research investigates and implements the psychological model of inner speech proposed by Alain Morin and collaborators. The model analyses many different mechanisms of inner speech, such as the social milieu, the physical environment, the role played during problem-solving, and the role of the information on internal aspects. Morin and collaborators actively contributed to the architecture design, and they validated the proposed cognitive architecture as a realistic computational model of inner speech.

Briefly, the working memory of the architecture includes the phonological loop as a component that manages the exchange of information between the phonological store and the articulatory control system. The inner dialogue is modeled as a loop where the phonological store hears the inner voice produced by the hidden articulator process. A central executive module drives the whole system and contributes to the generation of conscious thoughts by retrieving information from long-term memory. The surface form of thoughts thus emerges by the phonological loop. Once a conscious thought is elicited by inner speech, the perception of a new context takes place and then repeats the cognitive loop. The detailed description and the obtained results are presented in Chapter 6.

During the second year of the RESPECT project, the team members focused on the robot implementation of the theoretical model of inner speech developed during the first year. Moreover, the team refined and extended the previously developed model. The well-known ACT-R framework implements the inner speech architecture to keep analogies to humans' cognitive processes. The inner speech architecture is fitted into the ACT-R framework by identifying a set of correspondences between modules. Then, the ACT-R inner speech model is designed for the cognitive control scenario. Considering that conflict resolution influences daily performances in every task execution, such a scenario is more interesting than others.

Moreover, the psychological literature discusses how inner speech habits greatly influence solving conflict tasks. The proposed ACT-R inner speech

model was validated by fitting the psychological results in the literature. The model was then linked to the robot Pepper using ROS. The extensive implementation of the computational model and the linkage with Pepper are described in detail in Chapter 7.

The RESPECT project also investigated the role of inner speech in robot self-recognition by considering a variant of the mirror test. On the one hand, the mirror is a typical trigger for inner speech, and on the other hand, the ability to recognize oneself is closely related to self-awareness. Thus, a robot that can recognize itself in the mirror would be more able to maintain trustworthy interactions with humans. This research is reported in Chapter 8.

During the third year, the project team extensively carried out empirical experiments on human-robot interaction to investigate the role of the robot's inner speech in establishing and maintaining human participants' trust in robots. The experiments were conducted in the RoboticsLab using the Pepper humanoid robot platform. The empirical studies informed and constrained updating the computational model of inner speech for building trust relationships in autonomous robots. Complete results of the empirical experiments are reported in Chapter 9.

The empirical experiments highlighted how the chosen experimental scenario involving a dining table setup involves low-risk activities. Results of the experimental setup highlighted that robot inner speech did not affect the perception of security, which is important in building trustworthy human-robot interactions. Then, a preliminary investigation was performed concerning a more compromising scenario, when the robot trains a nurse to set up an operating table for surgery operations. Investigations are reported in Chapter 10.

Also, the empirical experiments highlighted that the implemented robot's inner speech is neutral and does not involve any emotional state. Then, during the last year, a model of the link between the robot's inner speech and emotions was designed. By inner speech, the robot appraises the context and infers the parameters of the appraisal variables. Then, the appraisal variables are instantiated, and a robot emotion emerges with a specific intensity. Now, the participant can hear the robot's inner speech and know its emotional state's motivations. The developed model of the robot's emotional inner speech is reported in Chapter 11.

Eventually, the team integrated the inner speech mechanism with Damasio's theory of emotions, where the central role of emotions in decision-making is revealed, formalizing a computational model for extended consciousness, according to Damasio. In Damasio's theory, emotions came first

in the body, guiding mind reasoning through feeling. In this work, inner speech takes emotion as initial input to its reasoning mechanism, making the robot aware of generated emotion by reasoning aloud. As a result of this work, the system SUSAN (Self-dialogue Utility in Simulating Artificial Emotions) was implemented and tested on a robot. The trial shows how a robot reacts to a musical stimulus by reasoning aloud and becoming aware of the emotion it is experiencing. The model is described in Chapter 12.

The outcomes of the experiments suggest that the human-robot teams' ethical choices must be adequately considered. The research question that arises is *How does the robot's inner speech enhance the ethical behavior of the human-robot team?* In collaboration with Prof. John P. Sullins of Sonoma State University, CA, the research team designed a simulated scenario concerning a nursing home where a patient has dementia. The robot has to set up the spot for that patient and others. Three different robot functioning are possible: (i) without inner speech, (ii) with an inner speech that is used for highlighting differences in facilitating the patient, (iii) with spontaneous inner speech. Preliminary considerations related to this scenario are reported in Chapter 13.

All the empirical studies were conducted following the protocol approved by the University of Palermo Ethics Committee and approved by AFRL HRPO. Particular care was taken to enforce the measures for COVID-19. In addition, a safety protocol was established for conducting interactive sessions between volunteers and robots provided by the project RESPECT: see Chapter 14.

Finally, the PI Antonio Chella wants to greatly thank all the wonderful and committed researchers of the team who contributed to making the RESPECT project a success: Arianna Pipitone, Antonella D'Amico, Valeria Seidita, Alessandro Geraci, Laura Di Domenico, Alain Morin, Famira Racy, John Sullins, Angelo Cangelosi, Sophia Corvaia, Irene Seidita, Giovanna Cataldo, Angelo Sabella, Francesco Lanza. Last but not least, the team wants to thank the PO of the project Nandini Iyer, for her continuous support and advice.

Chapter 2

Outcomes of the activities

The following papers were published under acknowledgment of Air Force Office of Scientific Research award FA9550-19-1-7025:

- Chella, A., Pipitone, A. (2020). A cognitive architecture for inner speech. *Cognitive Systems Research*, **59**, 287-292, `https://doi.org/10.1016/j.cogsys.2019.09.010`.
- Chella, A., Pipitone, A., Morin, A., Racy, F.(2020): Developing Self-Awareness in Robots via Inner Speech, *Frontiers in Robotics and AI*, **7**, 16, `https://doi.org/10.3389/frobt.2020.00016`.
- Chella, A., Pipitone, A.(2020): The inner speech of the IDyOT: Comment on Creativity, information, and consciousness: The information dynamics of thinking by Geraint A. Wiggins *Physics of Life Reviews*, **34**, pp. 42-43, `https://doi.org/10.1016/j.plrev.2019.01.016`.
- Geraci, A., D'Amico, A., Pipitone, A., Seidita, V., Chella, A.(2021): Automation Inner Speech as an Anthropomorphic Feature Affecting Human Trust: Current Issues and Future Directions, *Frontiers in Robotics and AI*, **8**, 620026, `https://doi.org/10.3389/frobt.2021.620026`.
- Pipitone, A., Chella, A. (2021): What robots want? Hearing the inner voice of a robot, *iScience*, **24**, 102371, `https://doi.org/10.1016/j.isci.2021.102371`.
- Pipitone, A., Chella, A. (2021): Robot passes the mirror test by inner speech, *Robotics and Autonomous Systems* 144, 103838, `https://doi.org/10.1016/j.robot.2021.103838`

- Corvaia, S., Pipitone, A., Chella, A. (2022): Human-Robot Cooperation by Robot Inner Speech and Emotions, Poster presented at the *Conference of the International Society for Research on Emotion ISRE 2022*, Los Angeles, CA `https://web.cvent.com/event/958d5907-c2f6-4e2d-a783-7e2ec267fcc0/summary`

- Chella, A., Sullins, J.P. (2022): Building Competent Ethical Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis, in: *Papers from the 2022 AAAI Spring Symposium on Approaches to Ethical Computing Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning*.

- Pipitone, A., Geraci, A., D'Amico, A., Seidita, V., Chella, A. (2023): Robot's Inner Speech Effects on Human Trust and Anthropomorphism, *International Journal of Social Robotics*, `https://doi.org/10.1007/s12369-023-01002-3`.

- Seidita, V., Sabella, A.M.P., Lanza, F., Chella, A. (2023): Agent talks about itself: an implementation using Jason, CArtAgO and Speech Acts, *Intelligenza Artificiale*, **17**, 1, pp. 7-18, `https://doi.org/10.3233/IA-230005`.

- Corvaia, S., Pipitone, A., Cangelosi, A., Chella, A. (2023): Inner Speech and Extended Consciousness: a Model based on Damasio's Theory of Emotions, in: *Proc. of 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Special Track on Affective Robotics, Boston, MA.

- Chella, A., Pipitone, A., Sullins, J.P. (2024): Competent Moral Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis, in: P. Wu, M. Salpukas, H-F. Wu, S. Ellsworth (eds.) *Trolley Crash: Approaching Key Metrics for Ethical AI Practitioners, Researchers, and Policy Makers*. Academic Press (in press).

- Pipitone, A., Corvaia, S., Chella, A. (2023): Robot's Inner Speech and Emotions (*IEEE Transactions on Affective Computing*, in revision).

- Pipitone, A. Cataldo, G., Chella, A. (2023): The Robot Talks to Itself While Training a Nurse (*Cognitive Systems Research*, submitted).

The argument of the Ph.D. thesis of Sophia Corvaia, currently in 2nd year, concerns *Inner Speech and Emotions in Social Robots*, Tutors Prof. Antonio Chella and Prof. Arianna Pipitone.

The following Master's Theses were submitted in partial fulfillment of the Master's Degree in Computer Engineering at the University of Palermo:

- Giovanna Cataldo (2022): *Il robot pensante per applicazioni mediche - Robot infermiere strumentista per la preparazione di un tavolo servitore e un tavolo madre* (The thinking robot for medical applications - Robot nurse instrumentalist for servant table and mother table preparation), Tutor Prof. Antonio Chella, co-tutor Prof. Arianna Pipitone.

- Irene Seidita (2023): *Misure di competenza del ragionamento etico in un robot dotato di voce interiore* (Measures of ethical reasoning competence in a robot with an inner voice), Tutor Prof. Antonio Chella, co-tutor Prof. Arianna Pipitone.

The PI Antonio Chella was invited to have keynotes or invited talks about the topics of the project at many Italian and International Conferences, under acknowledgment of Air Force Office of Scientific Research award FA9550-19-1-7025:

- Pipitone, A., Chella, A. (2020): Robot Passes the Mirror Test by Inner Speech, Invited talk at 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence (BICA*AI 2020), 10/10/2020 `https://player.vimeo.com/video/473377842`

- Chella, A. (2021): The Inner Speech of a Cognitive Architecture, Invited talk at Virtual International Symposium on Cognitive Architecture (VISCA 2021), 06/08/2021 `https://visca.engin.umich.edu`

- Chella, A. (2021): The Inner Speech of a BICA, Invited talk at BICA 2021: Biologically Inspired Cognitive Architectures 2021, 09/13/2021 `https://summit-2021.is4si.org/schedule/bica-schedule`

- Chella, A. (2021): Robot Inner Speech: A Sign for Robot Sentience? Invited seminar at Computer and Cognitive Science Seminar Series organized By Stevan Harnad at the University of Quebec in Montreal, 09/16/2021 `https://isc.uqam.ca/babillard/16-septembre-seminaire-dic-isc-cria/`

- Chella, A. (2021): The Inner Speech of a Robot, Invited talk at special topic symposium on *Becoming Familiar with Robots* at 14th Conference of the Italian Society for Analytic Philosophy, 09/22/2021 `http://portale.unime.it/sifa2020/?page_id=23`

- Chella, A. (2021): Inner Speech for AGI, Invited talk at *NARS Tutorial and Workshop* at AGI 2021, 10/15/2021 https://cis.temple.edu/tagit/events/workshop2021/index.html

- Chella, A. (2021): Un'Architettura Cognitiva per il Monologo Interiore del Robot (A Cognitive Architecture for the Internal Monologue of the Robot), Invited talk at the National Conference of the Italian Association for Automatic Calculus (AICA), 10/29/2021 https://www.aicanet.it/documents/10776/4095454/Programma+Congresso+2021/1edf9f3c-5820-4023-b91c-4637e4133ded

- Chella, A. (2022): Experiments in Cognitive Robotics, Invited talk at the UK-Italy Robotics and AI Research Collaboration Workshop, 03/01/2022 https://www.turing.ac.uk/events/uk-italy-robotics-and-ai-research-collaboration-workshop

- Chella, A. (2022): Robot Inner Speech: A New Perspective for Social Robotics, Invited talk at the 7th Joint UAE Symposium on Social Robotics, 03/29/2022 https://www.australiaexpo2020.com/whats-on/joint-symposium-social-robotics-unsw-uws

- Chella, A. (2022): Inner speech: a sign for robot self-consciousness? Invited talk at the International Symposium on Conscious and Unconscious Cognition, 04/19/2022 https://iscuc2022.wordpress.com

- Chella, A., Pipitone, A., Seidita, V. (2022): Un Robot Dotato di Saggezza (A Wise Robot), Invited talk at workshop on *Cognitive Robotics 2022* at 2021 Internation Conference on Image Analysis and Processing, 05/24/2022 https://mivia.unisa.it/smartrobot2022/

- Chella, A. (2022): The Heart and the Machine: EI and AI, Joint invited talk with Mark Sparvell at 8th International Congress on Emotional Intelligence, 09/02/2022 https://icei2022-palermo.it/joint-keynotes/

- Chella, A. (2022): Robot Inner Speech: A Sign of Consciousness? KTH Digital Future Distinguished Lecture, 11/10/2022 https://www.digitalfutures.kth.se/event/distinguished-lecture-antonio-chella-university-of-palermo-italy/

- Chella, A. (2023): Ethics Internal to Machines, Invited talk and discussion leader at Air Force Ethics in AI Workshop organized by P.

Friedland, J. Lyons, L. Steckman, at AFOSR BRICC facility center, Arlington, VA, 02/21/2023.

Chapter 3

Media Coverage

The goals and scope of the RESPECT project, and in particular the use of inner speech in the Pepper robot, generated great attention in the scientific media. Particularly, the article: Pipitone, A., Chella, A. (2021): What robots want? Hearing the inner voice of a robot, *iScience*, **24**, 102371, was referenced in more than 60 popular science articles and blogs (Fig. 3.1), and notably:

- The scientific website *Eurekalert* by the American Association for the Advancement of Science (AAAS) `https://www.eurekalert.org/pub_releases/2021-04/cp-ptr041521.php`,
- The popular science journal *New Scientist* `https://www.newscientist.com/article/2275323-robot-taught-table-etiquette-can-explain-why-it-wont-follow-the-rules/`,
- The scientific section of the newspaper *The Guardian* `https://www.theguardian.com/technology/2021/apr/21/study-inner-life-ai-robot-thinks-out-loud`
- The scientific section of the newspaper *Daily Mail* `https://www.dailymail.co.uk/sciencetech/article-9495661/Pepper-robot-think-loud.html`.
- The scientific section of the main Italian news agency *ANSA* `http://www.ansamed.info/ansamed/en/news/sections/science/2021/04/21/first-robot-that-thinks-out-loud-built-in-italy_a6488bfe-89a5-4140-a97d-057f7a97f28b.html`.

- The main Italian Television News (in Italian) *RAI News* https://www.rainews.it/dl/rainews/articoli/robot-pensa-alta-voce-d7449905-146b-4d1d-8f07-cbf0a9423995.html

- The scientific section of the main Italian newspaper *Corriere della Sera* (in Italian) https://www.corriere.it/tecnologia/21_aprile_23/studio-italiano-rivela-cosa-pensa-l-intelligenza-artificiale-mentre-esegue-nostri-ordini-b41b537a-a386-11eb-8b99-a42a4f90039f.shtml.

Notably, the Director of AFOSR, Shery Welsh, mentioned the project in the *AFOSR's Weekly Activity Report* of May 7, 2021 (Fig. 3.2). On May 26, 2021, the social media account of AFOSR published a post on social media describing the project (Fig. 3.3).

Also, the *Communications of the ACM* dedicated an article on the project in the *ACM NEWS* written by the popular science author Samuel Greengard https://cacm.acm.org/news/253654-channeling-the-inner-voice-of-robots/fulltext.

Recently, the *New York Times* on Jan 6, 2023, interviewed Antonio Chella on the topics of the robot inner speech project https://www.nytimes.com/2023/01/06/science/robots-artificial-intelligence-consciousness.html.

Italian Television *Sky Tg 24* on May 11, 2023, broadcast a TV report on our experiments on robot inner speech https://tg24.sky.it/tecnologia/now/2023/05/11/palermo-robot-pepper-empatico.

Italian Television *Mediaset Focus* on July 21, 2023, broadcast a long TV report on AI, and a part of the report was devoted to the project on robot inner speech https://www.tgcom24.mediaset.it/televisione/focus-speciale-intelligenza-artificiale_67305124-202302k.shtml.

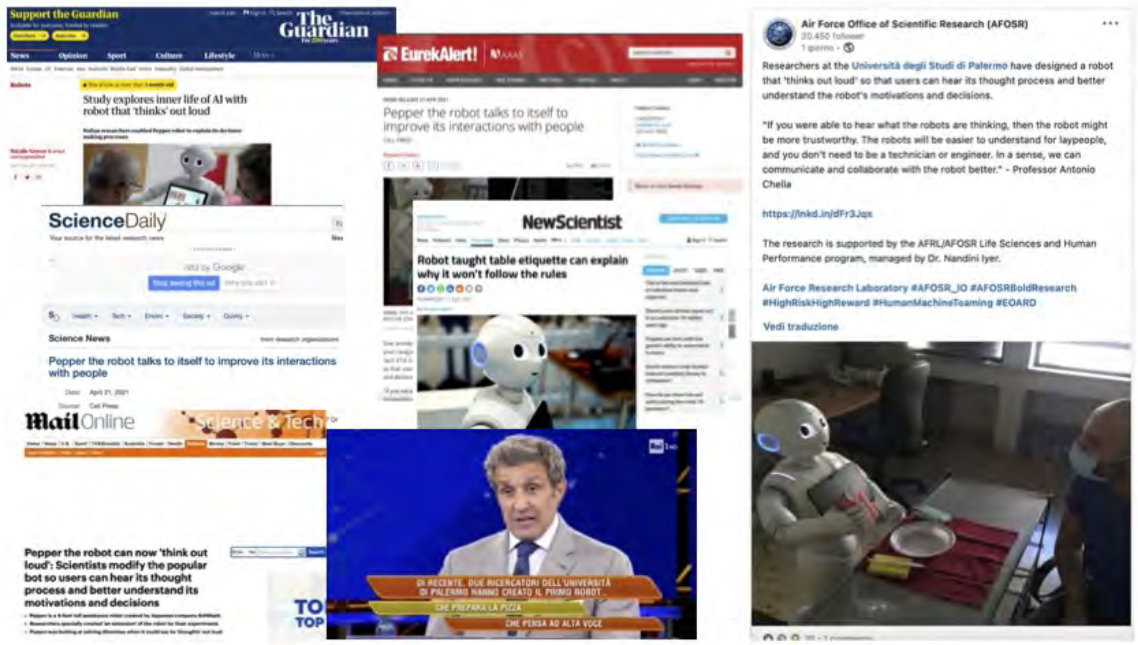


Figure 3.1: Screenshots of some websites covering our project.

BRILLIANT RISK-TAKING

Robot's inner speech improve human-machine partnerships. Humans engage in inner dialogue that often scaffolds high-level cognition, such as planning, focus and reasoning, and critically impacts transparency in human-human communication. AFOSR-funded researcher Prof. **Antonio Chella** of the Univ of Palermo, Italy designed and implemented a cognitive architecture for inner speech for robots and tested the same in a simple cooperative task with humans. The results satisfied international (ISO/TS:2016) functionality and transparency requirements when inner speech, implemented using a cognitive architecture, accompanies human-machine interactions. The proposed architecture could also be applied to other contexts such as robot learning, or switching attention across multiple tasks. (POC: Dr. Nandini Iyer Nandini.Iyer.2@us.af.mil) (Distro A)

Figure 3.2: Screenshots of a portion of AFOSR's Weekly Activity Report of May 26, 2021.

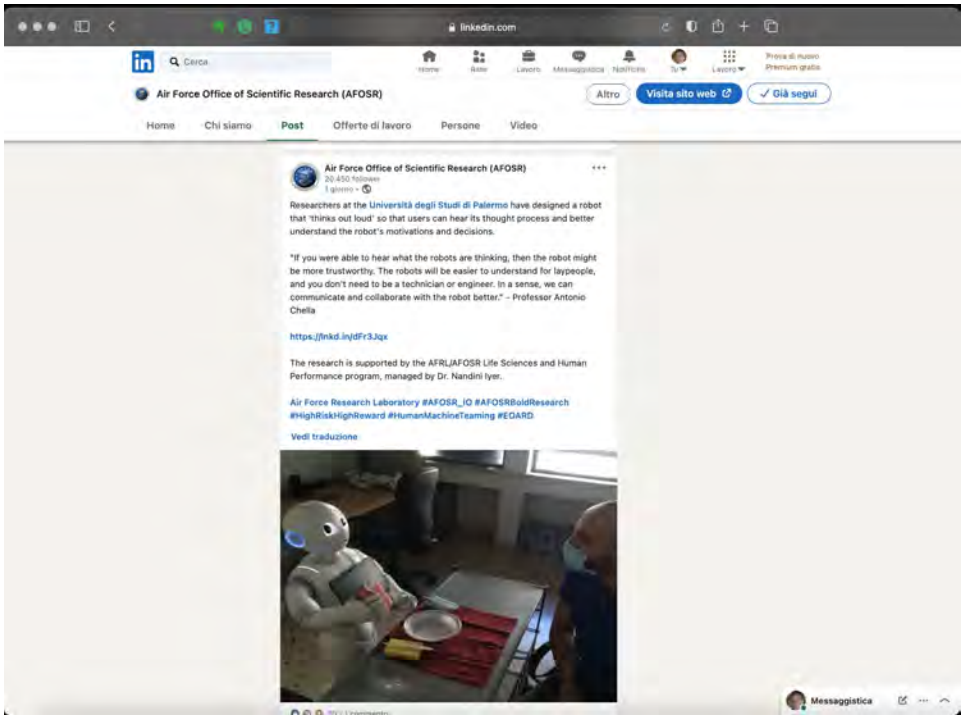


Figure 3.3: Screenshot of the AFOSR social account of LinkedIn taken on 05.27.2021.

Part I

Theoretical Investigations

Chapter 4

Literature Review

4.1 The Research Study

Trust research span many disciplines (i.e. psychology, philosophy, sociology, human-automation interaction) expanding the scientific literature on trust over the years, but also many definitions and operationalization have spread leading to different theories and models (Paliszkiewicz, 2011). Trust is multidimensional concept with no universal definition, which generally refers to an underlying psychological state affected by both cognitive and emotional processes (Chowdhury, 2005; Cummings Bromiley, 1996; Johnson Grayson, 2005; Kramer, 1999; Lewis Wigert, 1985; McAllister, 1995; Paliszkiewicz, 2011). Cognitive trust refers to an individual's conscious decision to trust based upon one's beliefs and knowledge about partner's reliability and competence (McAllister, 1995; Paliszkiewicz, 2011). On contrary, affective trust stems from interpersonal and emotional bonds, mostly based on the feelings of security, care and mutual concern (Johnson Grayson, 2005; McAllister, 1995). From a functional perspective, trust serves as a psychological mechanism for the reduction of social complexity through the formation of expectations and beliefs about other's intentions and behaviors (Luhmann, 1979; Lewis Wigert, 1985; Rompf, 2014). Lewis and Wigert (1985) states that rational prediction requires time and mental resources for collecting and processing information in order to determine highly probable outcomes, and thus trust may be an efficient alternative. Indeed, "by extrapolating past experiences into the future, individuals save the cognitive resources which would be otherwise needed for the search of information and its deliberate processing " (Rompf, 2014, p. 98).

Within the psychological literature, trust definitions highlights two key

elements: on one side, it involves positive attitudes, expectations, or confidence in the trustee (Corritore, Kracher, Wiedenbeck, 2003; Lee See, 2004; Rotter, 1967), on the other it implies a willing to put oneself at risk or in a vulnerability state (Kramer, 1999; Lee See, 2004; Mayer, Davis, Schoorman, 1995). According to Rotter's (1967) definition, interpersonal trust is as "an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon " (p. 651). Mayer, David and Schoorman's (1995) definition is one the most widely cited (Lee and See, 2004) and their model has become the dominant approach to the construct (Hamm, Smidt, Mayer, 2019). They define trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party " (Mayer et al., 1995, p. 712). In their model, they identify three perceived characteristics and action of the trustee (i.e. trustworthiness) that may lead to trust: ability, benevolence, and integrity. Ability describes trustor's subjective evaluation of trustee's skills and competences within the domain of interest. Benevolence accounts for the emotional dimension of trust; in fact, the trustor believes that the trustee cares about trustor's well-being beyond egocentric motivation. Lastly, trustor perceives trustee's integrity when he believes trustee to follows a set of consistent principles and values that are considered acceptable.

Muir (1987) states that trust is generally defined as an expectation of, or confidence in another, and it always has a specific referent. Indeed, it involves a relationship between "a trustor A that trusts (judges the trustworthiness of) a trustee B with regard to some behavior X in context Y at a time T " (Bauer Freitag, 2017, p. 2). Moreover, trust is dynamic in nature, because it develops and changes over time, but it is not a linear process: it may evolves as well as it may deteriorates through a process of loss and repair, in response to individual, social and environmental factors (Fulmer Gelfand, 2013; Paliszkiewicz, 2011). Trust research has identified three main: trust formation, dissolution and restoration (Fulmer Gelfand, 2013; Kim, Dirks, Cooper, 2009; Rousseau, Sitkin, Burt, Camerer, 1998). Trust formation starts when trustor choose to trust the trustee based on the perceived trustworthiness (i.e. ability, benevolence, integrity; Mayer et al. 1995). If trust is repeatedly violated, trustor decreases trust levels in trustee, entering in the dissolution phase. The restoration phase happens when trustor deliberately adopt repair strategies that allow trust levels in trustee to increase again, eventually stabilizing. These three phases are not necessarily linear and straightforward, since trust, at some point in time,

may be the result of ongoing violations and repairs.

According to Muir (1987), the same elements that serve as a basis for trust between individuals may affect people's trust for automation. However, in the psychological literature, both cognitive and emotional processes affect human-human trust because trustor needs to see trustee as competent but also careful and concerned about trustor's interests and well-being (Lewis Weigert, 1985; Mayer, Davis, Schoorman, 1995). On contrary, cognitive processes affect predominantly human- automation trust since the machine is expected to operate consistently with respect to certain performance standards (Lewis, Scyara, Walker, 2018; Merritt Ilgen, 2008; Muir, 1994). In human-automation interaction, similarly to Mayer et al. (1995) model (i.e. ability, benevolence, integrity), three factors as basis of trust have been identified: performance, process and purpose (Lee Moray, 1992; Lee See, 2004). Performance refers to the automation's capabilities and competences to achieve the operator's goals. Process focuses on the algorithms and operations that govern the conduct of the automation. Purpose concerns designer's intent behind the automation development. All of them taken together, these factors addresses the operator's perception and knowledge of what the automation does, how it functions and why it was developed.

Still, Merritt and Ilgen (2008) suggest four machine characteristics that may affect human trust: competence (i.e. automation's abilities to perform well), responsibility (i.e. automation's functioning information availability to the user), predictability (i.e. automation's behavior consistency), and dependability (i.e. automation's behavior consistency over time). In the past years, automation's development and implementation have increased exponentially in every context, leading to growing interactions with humans (Merritt Ilgen, 2008). Robots are now used in different contexts such as military, security, medical, domestic, and entertainment (Li, Rau, Li, 2010). Robots, compared to automation, are designed to be self-governed to some extent, in order to respond to situations that were not pre-arranged (Lewis, Scyara, Walker, 2018). Therefore, the greater the complexity of robots the higher the importance of trust in human-robot interaction (Lee See, 2004), as their collaboration increases over time (Schaefer, Chen, Szalma, Hancock, 2016). HRI (Human-Robot Interaction) is different from HCI (Human-Computer Interaction) and HMI (Human-Machine Interaction) because it concerns complex, dynamic and autonomous systems that operate in changing real-world environments (Scholtz, 2003). In the humans and automation interaction, the operator often must make a decision whether to use the automated system, which is a precursor to the effectiveness of interaction strategies (Merritt Ilgen, 2008).

In HAI literature trust is generally defined as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability " (Lee See, 2004, p. 54) and it is considered one of the main factors linked to automation use (Lee See, 2004; Lewis, Scyara, Walker, 2018; Merritt Ilgen, 2008; Parasuraman Riley, 1997). In human-human interaction, trust allows people to rely on each other, to accept the uncertainty and vulnerability deriving from someone else's action with the expectation of positive outcomes (Mayer, Davis, Schoorman, 1995). Accordingly, trust could play a key role in reliance between human and robot, allowing the latter to take on the role of a full collaborative partner (Hoff Bashir, 2015; Lee, Know, Baumann, Breazeal, DeSteno, 2013). HAI studies found that people have a tendency to rely on automation they trust and reject those who they do not trust (Bonnie Moray, 1996; Lee See, 2004; Lewandowsky, Mundy, Tan, 2000). Misuse occur when humans over-trust a robot relying excessively on its abilities compared to what it can actually execute, whereas disuse refers to the lack of trust on robot's capabilities so that the human choose simply not to use it resulting in a worse outcome (Lee See, 2004; Parasuraman Riley, 1997). Indeed, especially for high-risk situation, both misuse and disuse may have catastrophic consequences, like plane crash (Lee See, 2004; Lyons Stokes, 2012). If trust is appropriately calibrated, and that is when human's trust properly matches the true capabilities of the automation (Lee See, 2004), misuse and disuse may be avoided enabling an adequate, optimal, and safe human-robot interaction (Hoff Bashir, 2015; Lewis, Scyara, Walker, 2018). Nevertheless, there is some evidence that people rely on automation due to a bias that they make fewer mistakes than humans do, which in turn lead people to reduce reliance on automation as they perceive and remember more automation error and omission than human ones (Dzindolet, Pierce, Beck, Dawe, 2002; Madhavan Wiegmann, 2004). However, it is still unclear the extent to which an automation error produces changes in human trust level: for instance, trust levels may drop rapidly in response to first automation errors (Sauer, Chavaillaz, Wastell, 2016), but it may also decreases when automation fails at humans' easily perceived tasks (i.e. easy-error hypothesis; Madhavan, Wiegmann, Lacson, 2006), so that the operator infers that most likely the automation will not be able to perform difficult and complex task either.

In the psychological literature, research have identified several psychological, social and organizational factors that may enable the development of types of trust (e.g. dispositional trust, history-based trust, category-based trust, role-based trust, rule-based trust; Kramer, 1999). Dispositional trust is linked to other human traits, and it is shaped on early human experi-

ences where beliefs about other people are made. Based on their very early trust-related experiences, people builds up general beliefs that will be used to form expectations about others' behaviors, slowly ending up being a stable personality feature (Rotter, 1980). Conversely, history-based trust is a dynamic concept since it is founded on cumulative and continuous interactions between two or more agents. Therefore, an individual decides to trust another based on information he collected in their repeated interactions (Kramer, 1999; Merritt Ilgen, 2008). In a study, Merritt and Ilgen (2008) showed that history-based trust are likely to be more influenced by automation characteristics and less by propensity to trust, as the rates of interaction increases. Moreover the found that, when paired with not-functioning automations, participants with higher propensity trust suffered from a greater loss of trust levels.

In HAI and HRI literature, most researchers agree that trust dynamically emerges from the exchange of the distinct features of the operator, the machine and the specific environment where the interaction takes place (Hancock et al., 2011; Hoff Bashir, 2015; Kessler, Stowers, Brill, Hancock, 2017; Lewis, Scyara, Walker, 2018; Schaefer et al., 2016). Two extensive meta-analyses carried out by Hancock et al. (2011) and Schaefer et al. (2016) supported a three-factor model, which identify three main factors affecting trust in human-automation/robot interaction: human-related (e.g. cognitive and emotional factors, age, gender personality etc.), automation/robot-related (e.g. performance, errors, level of automation, anthropomorphism, etc.), and environment-related factors (e.g. role interdependence, tasks complexity and type, culture, physical environment, etc.). These studies addressed issues related to HRI and to the broader HAI, showing similarities but also some differences (Kessler et al., 2017). Hancock et al. (2011) found that robot-related factors had the largest influence on trust compared to the small effect of environment-related and human-related factors. Within the robot characteristics, performance-based elements, like behavior and failures, presented the highest correlation with trust. The authors argued that these results suffer from a lack of sufficient number of studies mostly related to human-related and environment-related factors. Moreover, they suggest incorporating physiological indicators and objective measures of trust since most of the studies rely on subjective responses, a topic that is still debated in the psychological literature (Baumeister, Vohs, Funder, 2007; Huffman, Van Der Werff, Henning, Watrous-Rodriguez, 2014; Lewandowski Jr Strohmetz, 2009; Muckler Seven, 1992). The meta-analysis carried out by Schaefer et al. (2016) extended the investigation to every kind of machines, including robots, because "understanding human trust development with

all of automation promises to render a much more complete picture of the general issue " (p. 2). Indeed, they corroborated previous results about the importance of automation-related factors among all the factors, specifically the automation's capabilities (e.g. behaviors, failures, etc.). By contrast, they found that human-related factors have an impact on trust development, showing a moderate to large effect of emotive factors followed by age and cognitive factors. Moreover, compared to Hancock et al. (2001) results, no environment-related factors influence was detected due to the lack of studies. Schaefer et al. (2016) suggest expanding the area of automation features, like mode of communication and appearance/anthropomorphism, since it is a relatively recent area of research in HAI literature.

4.2 Main References Analyzed

Bauer, P. C., & Freitag M. (2017). Measuring trust. In E. M. Uslaner (Eds.), The oxford handbook of social and political trust (pp. 1-30). Oxford: Oxford University Press.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? Perspectives on Psychological Science, 2(4), 396-403.

Chowdhury, S. (2005). The role of affect- and cognitions-based trust in complex knowledge sharing. Journal of Managerial Issues, 17(3), 310–326.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. International Journal of Human-Computer Studies, 59(6), 737-758.

Cummings, L. L., & Bromiley, P. (1996). The Organizational trust inventory (oti): Development and validation. In R. M. Kramer, & T. R. Tyler (Eds.), Trust in organizations: Frontiers of theory and research (pp. 68-89). Thousand Oaks, CA: Sage.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1), 79-94.

Fulmer, C. A., & Gelfand, M. J. (2013). How do I trust thee? Dynamic

trust patterns and their individual and social contextual determinants. In K. Sycara, M. Gelfand, & A. Abbe (Eds.), Advances in group decision and negotiation: Vol. 6. Models for intercultural collaboration and negotiation (p. 97–131). Heidelberg: Springer. Hamm, J. A., Smidt, C., & Mayer, R. C. (2019). Understanding the psychological nature and mechanisms of political trust. PLOS ONE, 14(5), 1-20.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517-527. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Huffman, A. H., Van Der Werff, B. R., Henning, J. B., & Watrous-Rodriguez, K. (2014). When do recycling attitudes predict recycling? An investigation of self-reported versus observed behavior. Journal of Environmental Psychology, 38, 262-270.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. Journal of Business Research, 50, 500-507.

Kessler, T., Stowers, K., Brill, J. C., & Hancock, P. A. (2017). Comparisons of human-human trust with other forms of human-technology trust. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61(1), 1303–1307.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bi-lateral perspective and multi-level conceptualization. Academy of Management Review, 34, 401-422

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. Annual Review of Psychology, 50, 569-598.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35, 1243-1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Lee, J. J., Know, B., Baumann, J., Breazeal, C., & DeSteno, D. (2013).

Computationally modeling interpersonal trust. Frontiers in Psychology, 4, 893.

Lewandowski Jr, G. W., & Strohmetz, D. B. (2009), Actions can speak as loud as words: Measuring behavior in psychological science. Social and Personality Psychology Compass, 3, 992-1002.

Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. Journal of Experimental Psychology: Applied, 6(2), 104-123.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. Social Forces, 63, 967-985.

Lewis, M., Scyara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), Foundations of trusted autonomy (pp. 135-160). Cham, Switzerland: Springer.

Li, D., Rau, P. P., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. International Journal of Social Robotics, 2(2), 175-186.

Luhmann, N. (1979). Trust and power. New York: Wiley.

Lyons, J. B., & Stokes, C. K. (2012). Human-human reliance in the context of automation. Human Factors, 54(1), 112-121.

Madhavan, P., & Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: Is trust in human comparable to trust in machines? Proceeding of the Human Factors and Ergonomics Society 48th Annual Meeting.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. Human Factors, 48(2), 241-256.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20, 709-734.

McAllister, D. J. (1995). Affect- and cognitive-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal, 38(1), 24-59.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. Human Factors, 50(2), 194-210.

Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures: "Objective" versus "subjective" measurement. Human Factors, 34(4), 441–455.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies, 27, 527-539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics, 37(11), 1905–1922.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics, 39(3), 429-460.

Paliszkiewicz, J. O. (2011). Trust management: Literature review. Management, 6(4), 315-331.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253.

Rousseau, D., Sitkin, S., Burt, R., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. Academy of Management Review, 23(3), 393-404

Rompf, S. A. (2014). Trust and rationality: An integrative framework for trust research. New York: Springer.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. Journal of Personality, 35, 651-665.

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. American Psychologist, 35(1), 1-7.

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. Ergonomics, 59(6), 767-780.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors, 20(10), 1-24.

Scholtz, J. (2003). Theory and evaluation of human robot interactions. Proceedings of the 36th Hawaii International Conference on System Sciences.

Chapter 5

A Calculus for Robot Inner Speech and Self-Awareness

5.1 Introduction

The dialogue with the self plays a fundamental role in human's consciousness [45] [146] [153], and is linked to self-awareness: by talking to himself, someone accesses to self-information, or extends existing self-knowledge [161].

Morin [165] stated three main causal directions between inner speech and self-awareness: (i) inner speech always precedes (causes) self-awareness, (ii) inner speech accompanies a state of awareness, and (iii) inner speech is triggered by self-focus, that is the attention on the self.

When analyzing the robot awareness and self-awareness, it seems desirable to provide the robot with a kind of self dialogue. Few works investigate this scenario. The authors proposed a cognitive architecture [189] for inner speech, that is modeled as the rehearsing process between a working-memory and a motor module wich produces language. Same authors suggested to integrate their architecture into the IDyOT system [38]. Other works [217] demonstrated that the language re-entrance, that is a form of inner speech, allows to refine the emergent grammar shared by a population of agents.

We attempt to automate the Morin's causal directions by defining a calculus couched in first-order modal logic. The **Deontic Cognitive Event Calculus** ($\mathcal{DCEC}$) [84] underlies the proposed one which includes some of the $\mathcal{DCEC}$'s elements while adding new propositions and terms for formalizing inner speech. $\mathcal{DCEC}$ subsumes in turn the Event Calculus (EC) [209], and it was proposed to formalize thorny situations, such as *akrasia* [22] and the *Doctrine of Double Effect* ($\mathcal{DDE}$) [150].

We show that our formalization may affect the execution by robot of a simple task when a self-perception stimuli occurs and triggers inner speech. The idea we propose lays the groundwork for new investigations related to the robot's inner speech ability, and for testing its effects on robot awareness and self-awareness.

The Chapter is organized as follow: a brief overview about $\mathcal{DCEC}$ is presented at section 5.2. Then, how we think to encode inner speech and self-awareness by EC is explained at section 5.3. The syntax of our calculus (with sorts, functions and axioms definitions) is detailed at section 5.4. Inference schemata and formal conditions, that allow to make deduction and to reason on the syntax, are discussed at section 5.5. The section 5.6 shows a simple simulation for reasoning by the proposed calculus. Future works and implications are discussed at 11.7.

5.2 The Deontic Cognitive Event Calculus ($\mathcal{DCEC}$)

The common Event Calculus (EC) [209] is a sorted logic analogous to a typed programming language. It allows to formalize intuitive reasoning about actions and the changes which occur in the universe after doing these actions. The notion of 'action affecting' arises, meaning the **causal** influences of actions on the states of the universe.

$\mathcal{DCEC}$ [84] is a calculus that subsumes EC, by adding new operators and functions to enable **intensional** reasoning processes. The intensional property is necessary when modeling the typical artificial agent's states (such as knowledge, belief and intention), which are intensional (i.e. it is not possible to declare and to know all the values of these states). The models based on extensional property, which attempt to list all values, was demonstrated [23] generate inconsistencies for these states.

The $\mathcal{DCEC}$'s intensional processes was successfully used to automate the false-belief task [7], the *akrasia* situation [22] (which represents the temptation to violate moral principle) and the $\mathcal{DDE}$ situation [150] (which arises when a dilemma with positive or negative effects has to be solved by an autonomous agent). The intensional property of the model is achieved by the inclusion in the formalization of intensional modal operators. For example, the intensional modal operator for modeling the knowledge of a fact has the typical form $\mathbf{K}(a, t, \phi)$, which means that agent a knows the proposition ϕ at time t. An instance of this operator allows to model an intentional knowledge state of the agent.

Involving same kinds of states, the calculus for inner speech and self-

awareness is an intensional model, and we inspired to $\mathcal{DCEC}$ for our work.

5.3 Encoding Inner Speech and Self-Awareness

As shown, Morin's theory [161] claims the causal directions between inner speech and self-awareness. Because of this causality property, EC is suitable to automate such directions.

Since a fluent represents a state (or fact) of the world, we can define *inner* and *external* fluents. An inner fluent represents an internal agent state, eg: an emotion, its intensity value, the status of an its physical component (on, off, functioning,...). An external fluent is related to the external context, and represents a fact of the environment, that is the typical definition of EC's fluent. Thus, how a typical fluent is initiated or terminated by an action which changes the environmental set (named *reactive action*), similarly an internal fluent is initiated or terminated by an internal action (named *inner action*, i.e. a self-regulation action, the perception of an entity, the rehearse of inner voice,...) which changes the internal set.

We think that *awareness and self-awareness of an artificial agent could be computationally modeled as all the fluents (both inner and external) the agent knows to be active at a time.* A fluent is active at time t when it holds at t. The active fluents that are not in the knowledge of the agent are not in the set modeling awareness and self-awareness: the agent does not know them, and it is not aware of them.

Now, let suppose that the object x at the instant t_1 is at the location l_1. The native EC function

$$holds(location(x, l_1), t_1)$$

models the state of the world related to that fact, that is the active state of the fluent $location(x, l_1)$ at the instant t_1.

The action $a = move\ x\ to\ l_2$ at the instant t_2 with $t_1 < t_2$, causes the end of the previous fluent

$$terminates(a, location(x, l_1), t_2)$$

and the beginning of the new fluent $location(a, l_2)$ by

$$initiates(a, location(a, l_2), t_2)$$

and for the inferential logic of EC, the creation of a new function *holds* related to such a new fluent.

When such an action has to be performed by an autonomous agent, it may involve inner fluents and statements beyond the previously seen ones. The agent will use inner speech as a cognitive tool [163] to accomplish the task $a = move\ x\ to\ l_2$. For example, it has to evaluate its abilities to solve the problem, the position of the object in respect to its position, the form of the object, the feasibility of the action, the state of its physical components for making the action. To answer to the questions it makes to itself, the robot has to retrieve useful self-information and information from the environment.

Kendall et. al [118] proposed four categories of self-questions emerging during the problem-solving process: (1) the questions for a clear formulation of the problem ('What's the problem?', 'What's I'm supposing to do?'), (2) the questions for proposing a possible solution ('I have to find a strategy'), (3) the questions for focusing on relevant aspects for the solution ('It's important', 'It's not important, I discard it'), (4) the statements for praising oneself when the solution is reached ('Good! I find the solution!) or for readjusting the approach when someone fails ('Oh, no! It's no ok').

All these evaluations will modify the set of active fluents, and hence the state of awareness and self-awareness of the agent.

The calculus we propose attempt to model such evaluations. It could be extensible by adding fluents representing further internal or external facts.

5.4 The Proposed Calculus

Commonly used functions, sorts and relation symbols from EC and some ones from $\mathcal{DCEC}$ are included in our formalization. The resulted calculus has a unique syntax and a proof calculus for reasoning and theorem proving. In particular, the *natural deduction* by Gentzen [80] is used as prover, as shown at section 5.6.

5.4.1 Modal Fragment

The modal fragment specifies the modal operators of the calculus.

While $\mathcal{DCEC}$ includes the standard modal operators for *common-sense knowledge* $\mathbf{C}$ and *domain knowledge* $\mathbf{K}_d$, in our formalization we include other two types of knowledge by adding two new modal operators, representing:

- the *self-knowledge* $\mathbf{K}_{self}$, that is the knowledge the robot owns about itself, (i.e. *inner world*). Then, $\mathbf{K}_{self}(a, t, \phi)$ means that the agent a knows the proposition ϕ representing an inner fact at time t;

- the *contextual-knowledge* $\mathbf{K}_{cx}$, that is the knowledge about the physical environment in which the robot is plunged, (i.e. *external world*). The contextual-knowledge we considered is not equal to the general one $\mathbf{K}$, because it may include some new objects the robot does not know (it has never seen them). Moreover, the contextual knowledge may not include some concepts of the general one because these concepts could not be in the environment at a time. $\mathbf{K}_{cx}(a, t, \phi)$ means that agent a knows the proposition ϕ representing an external fact at time t.

The *general knowledge* we refer to becomes

$$\mathbf{K} = \mathbf{K}_{self} \vee \mathbf{K}_{d} \vee \mathbf{K}_{cx}$$

The standard intensional operators for belief $\mathbf{B}$, desire $\mathbf{D}$, intention $\mathbf{I}$ are included too, and have the same semantics from $\mathcal{DCEC}$.

Other $\mathcal{DCEC}$ modals operators we include are $\mathbf{P}$ for perceiving a state, and $\mathbf{S}$ for agent-to-agent communication or public announcement. But the $\mathbf{S}$ operator we define crucially differs from those of $\mathcal{DCEC}$ because we consider the possibility to have the same agent (the robot) in its agent-to-agent arguments, leading to the inner speech formalization. The single argument means public announcement as for $\mathcal{DCEC}$. Formally:

- $\mathbf{S}(a, b, t, \phi)$ means that the agent a sends the message related to proposition ϕ to the agent b at time t;
- $\mathbf{S}(a, a, t, \phi)$ means that the agent a rehearses the message related to proposition ϕ it sends to itself at time t;
- $\mathbf{S}(a, t, \phi)$ means that the agent a makes a public announcement related to proposition ϕ at time t.

Finally, we add the modal operator $\mathbf{M}$ for producing a message related to a specific proposition . Then, $\mathbf{M}(a, t, \phi)$ means that the agent a produces the message related to the proposition ϕ at time t. We have to specify that $\mathbf{M}$ regards the action "to produce a message" which can be in turn sent to another agent or rehearsed according to $\mathbf{S}$.

5.4.2 Sorts Specification

We define new sorts upon the native EC and $\mathcal{DCEC}$ ones. Moreover, we changed the typical `Agent` and `Entity` sorts definitions because we consider

the agent as a part of the universe, hence the `Agent` sort becomes a sub-type of the `Entity` sort. The *abstract* sorts are instantiated at particular times by an actors.

The following table shows all the sorts we defined. The highlighted sorts are the new ones introduced or modified by our calculus.

Sort	Description
`Entity`	An entity in the universe, including agent.
`Object`	A subtype of `Entity`, representing a physical object in the environment that is not an actor.
`Agent`	A subtype of `Entity`, representing human and artificial actors.
`Percept`	A perception from the environment; it can be a concrete `Entity` (an `Agent` or an `Object`), or a generic concept meaning an event.
`Moment`	A time in the domain.
`Event`	An event in the domain.
`ActionType`	An abstract *reactive action*, i.e. an action that affects the state of the external environment.
`SelfActionType`	An abstract *inner action*, i.e. an action that affects the inner state of the agent. Examples: keep calm, evaluate, feel.
`Action`	A subtype of `Event` that occurs when an agent performs a concrete `ActionType` action.
`SelfAction`	A subtype of `Event` that occurs when an agent performs a `SelfActionType` action.
`Fluent`	A state of the universe, that can be inner or external.
`Aware`	A subtype of `Fluent`, representing a fluent the agent knows. The set of active `Aware` fluents forms the agent awareness and/or self-awareness.
`Message`	A message of an agent-to-agent, agent-to-itself, agent-to-public communication.

5.4.3 Defining the message m

EC and $\mathcal{DCEC}$ have not any operators or functions to define the message of a communication. We take a modal approach for modeling such a situation without defining specific axioms. We assume that a message is related to a specific fact, represented by a proposition. Given a proposition ϕ, we define by the operator $\odot$ the set of all constants and names of not native function expressions in ϕ. For example:

$$\odot(happens(action(Bob, loves(Mary))), t) = \{Bob, loves, Mary\}$$

$$\odot(holds(located(apple, table)), t) = \{located, apple, table\}$$

A message is the set returned by the $\odot$ operator, that is:

$$\odot : \phi \rightarrow m$$

So, given ϕ the corresponding message m will be $m = \odot(\phi)$.
A **content** of the message is an element in the $\odot$ set. So, the content $p_i \in m$ is the i-th element in m.

5.4.4 The Syntax

The whole syntax of the calculus is formalized by the formulas in the following table, where S represents the sorts, f represents the functions, the *term* represents the possible variables and finally ϕ are the propositions.

$$S ::= \begin{array}{l} \texttt{Entity} \mid \texttt{Agent} \sqsubseteq \texttt{Entity} \mid \texttt{Object} \sqsubseteq \texttt{Entity} \mid \texttt{Percept} \sqsubseteq \texttt{Entity} \mid \\ \texttt{Percept} = \texttt{Agent} \sqcup \texttt{Object} \sqcup \texttt{Action} \sqcup \texttt{SelfAction} \mid \texttt{Fluent} \mid \\ \texttt{ActionType} \mid \texttt{SelfActionType} \mid \texttt{Event} \mid \texttt{SelfAction} \sqsubseteq \texttt{Event} \mid \\ \texttt{Action} \sqsubseteq \texttt{Event} \mid \texttt{Boolean} \mid \texttt{Moment} \mid \texttt{Message} \end{array}$$

$$
f ::= \begin{cases}
\mathit{action} : \texttt{Agent} \times \texttt{ActionType} \rightarrow \texttt{Action} \\
\mathit{initially} : \texttt{Fluent} \rightarrow \texttt{Boolean} \\
\mathit{holds} : \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\
\mathit{happens} : \texttt{Event} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\
\mathit{clipped} : \texttt{Moment} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\
\mathit{initiates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\
\mathit{terminates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\
\boldsymbol{selfaction} : \texttt{Agent} \times \texttt{SelfActionType} \rightarrow \texttt{SelfAction} \\
\boldsymbol{focuses} : \texttt{Percept} \rightarrow \texttt{SelfActionType} \\
\boldsymbol{aware} : \quad \texttt{Agent} \times \texttt{Percept} \rightarrow \texttt{Fluent} \\
\boldsymbol{content}_i : \texttt{Message} \rightarrow \texttt{Percept} \\
\boldsymbol{comprehends} : \texttt{Message} \rightarrow \texttt{SelfActionType} \\
\boldsymbol{produces} : \texttt{Message} \rightarrow \texttt{SelfActionType} \\
\boldsymbol{innerspeaks} : \texttt{Message} \rightarrow \texttt{SelfActionType}
\end{cases}
$$

$$
term ::= \quad var : S \mid const : S \mid f(term_1, term_2, ...term_n)
$$

$$
\phi ::= \begin{cases}
\mathit{term}{:}\texttt{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \\
\mathbf{C}(t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\phi) \mid \mathbf{I}(a,t,\phi) \\
\\
\mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,a,t,\phi) \mid \mathbf{S}(a,t,\phi) \\
\\
\mathbf{M}(a,t,\phi)
\end{cases}
$$

The functions in bold are purposely defined for formalizing inner speech, awareness and self-awareness, while the others are native from EC.

These new functions are :

- *selfaction(a, a_t)*$\rightarrow a_s$: that returns the concrete self action a_s of type a_t the agent a performs; it is the self-version of native *action* function of EC;
- *focuses(p)*$\rightarrow a_t$: that represents the selfaction type the agent a performs when it focuses on the percept p. Intuitively, $selfaction(a, focuses(p))$ means the action by a of focusing on the percept p;
- *aware(p)*$\rightarrow f$: that states the awareness about the percept p. The

activation of that fluent by $holds(aware(p), t)$ means that an agent is aware of the percepts p at t. Let's notice that the *aware* fluents may be a sub-set of the all active fluents at a time. They are not the whole robot awareness and self-awareness, but are the robot awareness about the percepts.

- *content*$_i$*(m)*$\rightarrow p_i$: that returns the i-th percept p_i in the message m;
- *comprehends*[1] *(m)*$\rightarrow a_t$: that states the comprehension of the message. Intuitively, $selfaction(a, comprehends(m))$ means the action by a of comprehending the message m;
- *produces*[2]*(m)* $\rightarrow a_t$: that states that the agent a is producing the message m. Intuitively, $selfaction(a, produces(m))$ means the action by a of producing the message m.

The typical truth-functional connectives $\wedge$, $\vee$, $\neg$, $\rightarrow$ are applied to propositions and they have the standard first-order semantics.

5.4.5 Axioms

As for $\mathcal{DCEC}$, the standard axioms of EC are considered as common-knowledge, that is:

$[A_1]$ $\mathbf{C}(\forall\ f,\ t\ .\ initially(f) \wedge \neg clipped(0, f, t) \Rightarrow holds(f, t))$

$[A_2]$ $\mathbf{C}(\forall\ e,\ f,\ t_1,\ t_2\ .\ happens(e, t_1)\ \wedge\ initiates(e, f, t_1)\ \wedge\ t_1 < t_2\ \wedge\ \neg clipped(t_1, f, t_2) \Rightarrow holds(f, t_2))$

$[A_3]$ $\mathbf{C}(\forall\ f,\ t_1,\ t_2\ .\ clipped(t_1, f, t_2)\ \Longleftrightarrow\ [\ \exists\ e,\ t\ .\ happens(e, t) \wedge t_1 < t < t_2 \wedge terminates(e, f, t)])$

$[A_4]$ $\mathbf{C}(\forall\ a,\ d,\ t\ .\ happens(action(a, d), t) \Rightarrow \mathbf{K}(a, happens(action(a, d), t), t))$

The axioms from $[A_1]$ to $[A_3]$, that are native of EC, state general and innate understanding human capacity about the causality of events. $[A_4]$ is native of $\mathcal{DCEC}$ and postulates that an agent knows the action it performs, that is it means the event which occurs by doing such an action.

[1] The term *comprehends* is the name of the function defined in the Fluid Construction Grammar engine [216] for parsing an input utterance. Such a function returns the meaning of the sentence.

[2] The term *produces* is the name of the function defined in the Fluid Construction Grammar engine [216] for producing an output sentence given a set of meanings. Such a function returns the syntactic form of the conjunctions of the meanings.

We postulate one more axiom pertains the inner speech triggering. It expresses that when an agent rehearses a message it produces, the inner speech happens. That is:

$[A_5]$ $\mathbf{C}(\forall\ a,\ m,\ t_1,\ t_2\ .\ happens(selfaction(a, produces(m)), t_1) \wedge$
$happens(selfaction(a, comprehends(m)), t_2) \wedge\ t_1 < t_2$
$\Leftrightarrow happens(selfaction(a, innerspeaks(m)), t_2))$

We define three more axioms that, according to Morin's theory, state the three main causal directions between inner speech and self-awareness. Considering that inner speech is an unconscious activity [**?**] despite it triggers self-awareness, we do not model these axioms as common-knowledge.

As result:

$[A_6]$ $\forall\, a,\ \phi,\ m,\ (\forall\ p_i \in m),\ t\ .\ initiates(selfaction(a, innerspeaks(m)),$
$aware(a, p_i), t)$

$[A_7]$ $\forall a,\ t_1,\ t_2,\ m,\ p_i \in m\ .\ clipped(t_1, aware(a, p_i), t_2)\ \Rightarrow$
$\exists\ t \in [t_1, t_2]\ ,\ happens(selfaction(a, innerspeaks(m)), t)$

$[A_8]$ $\forall a,\ t,\ m,\ p_i \in m\ .\ happens(selfaction(a, focuses(p_i)), t) \Rightarrow$
$happens(selfaction(a, innerspeaks(m)), t)$

$[A_6]$ postulates that inner speech precedes (causes) awareness. The formula means that when the self-action related to the comprehension of a message produced by itself (i.e. rehearsed) has taken place at time t (*happens(e,t)*), the corresponding event has the aware fluent about the content of the message as an effect. More specifically, $[A_6]$ formalizes that 1 precedes 2 in the following:

1. the agent a produces a message m:
   produces(m)
   and it rehearses such a message:
   comprehends(m)
   generating the event by axiom $[A_5]$:
   selfaction(a, innerspeaks(m))
   which then has taken place at t:
   happens(selfaction(a, innerspeaks(m)), t)

2. the agent becomes aware of each content of the message, generating the corresponding fluents of awareness:
   aware(a, p_i)
   hence the previous event has each of this fluent as an effect:
   initiates(selfaction(a, innerspeaks(m)), aware(a, p_i), t)

$[A_7]$ postulates that inner speech accompanies a state of awareness. That is:

1. If the robot is aware of the percept p_i, and if the corresponding fluent:
   aware(a, p_i)
   has not been made false in the time interval $[t_1, t_2]$:
   clipped(t_1, aware(a, p_i), t_2)

2. the inner action to rehearse itself has taken in the meanwhile:
   $\exists\ t \in [t_1, t_2]$
   happens(selfaction(a, innerspeaks(m)), t)
   with m related to p_i, i.e. $p_i \in m$.

Finally, $[A_8]$ states that inner speech is triggered by *focus*, that is a specific inner action. Then:

1. If the agent a focuses on a percept p at time t:
   happens(selfaction(a, focus(p)), t)

2. the inner speech action starts, begin m the message whose content is p:
   selfaction(a, innerspeaks(m))
   with $p \in m$

The following axiom states that is common-knowledge that when an agent perceives a fact, it is focusing almost one percept which will be related to that fact:

$$[A_9]\ \mathbf{C}(\forall a,\ t,\ \exists\,(\phi,\ p)\ .\ \mathbf{P}(a, t, \phi) \Leftrightarrow\ happens(selfaction(a, focuses(p))))$$

Considering that the action to formulate a message generates the corresponding message m, we also postulate the axiom $[A_{10}]$:

$$[A_{10}]\ \mathbf{C}(\forall a,\ t,\ \exists\,(\phi,\ m)\ .\mathbf{M}(a, t, \phi) \Rightarrow\ happens(selfaction(a, produces(m)))).$$

Finally, the awareness about itself, leads to the self-knowledge of the proposition triggering it, so:

$$[A_{11}]\ \mathbf{C}(\forall a,\ t\ .\ holds(aware(a), t) \Leftrightarrow [\exists\ \phi\ .\ \mathbf{K}_{self}(a, t, \phi)]).$$

5.5 Inference Schemata

Some of the inference rules we define for reasoning by the calculus are:

$[R_1]$ $\frac{\mathbf{M}(a,t,\phi)}{\mathbf{S}(a,a,t,\phi)}$

The generation of a message $\mathbf{M}$ about ϕ leads to rehearse it.

$[R_2]$ $\frac{\mathbf{P}(a,t,\phi)}{\mathbf{M}(a,t,\phi)}$

The perception of ϕ leads to a message about ϕ.

$[R_3]$ $\frac{\mathbf{K}_{self}(a,t,\phi)}{\mathbf{K}_{self}(a,t,\ \mathbf{K}_{self}(a,t,\phi))}$

It captures an essential property of self-awareness: the knowledge of self about ϕ leads to the knowledge about such a knowledge. That is "I know to know".

$[R_4]$ $\frac{\mathbf{K}(a,t,\phi)}{\phi}$

The knowledge of ϕ implies ϕ.

5.6 Simulation

A concrete example allows us to clarify and to show the application of the calculus we propose. The scenario we consider is inspired to those described at [160]: inner speech is conceived as a cognitive tool the individual uses for self-reflection. In this case, the self is the object of the questions ('Who am I?', 'What am I doing?'), and the self-knowledge and information from the environment are the answers. A person engaged in a task might self-reflect by this kind of inner dialogue, and it was demonstrated it facilitates the task execution.

Therefore, we describe the method for reasoning by the proposed calculus about self-reflection during task execution. The knowledge of the robot changes by self-reflection in respect to the knowledge of the typical reactive behavior, putting the robot in the conditions to make more inferences than the case without self-reflection.

At this step, our goal is to demonstrate how the robot awareness and self-awareness grow by modeling inner speech by the proposed calculus implementing self-reflection. How the questions emerge, and the answers are produced are out of the scope of this CHapter, and regard important future works.

5.6.1 Encoding self-reflection

We suppose that the robot is engaged in a simple task, that is to remove an object from its locations and to put it in a different location. In the reactive behavior, the robot runs a set of routines which allow: to identify the object o, to move to the location l, to grasp the object o from the location l, to place the object o in the location l.

Let suppose that at a certain time during task execution, the robot perceives itself by a mirror, and it sees itself to perform one of the described actions. In particular, it sees itself to grasp the object.

It will be engaged in the following soliloquy: "What am I doing? Did I perform yet this task?".

We introduce the sort `Location`, and the following function symbols for reasoning about this situation:

$$grasp : \texttt{Object} \rightarrow \texttt{ActionType}$$

$$remember : \texttt{ActionType} \times \texttt{Moment} \rightarrow \texttt{Boolean}$$

The set of questions will generate a set of corresponding propositions:

"What am I doing?" $\Rightarrow$ "I grasp the object o"
$\Rightarrow$ ***happens(action(a, grasp(o)),t)***

"Did I perform this task?" $\Rightarrow$ "Yes, I do!"
$\Rightarrow$ ***remember(grasp(o), t)***

We consider two temporal lines which allow to specify the narrative of self-reflection, that are the times before inner speech and the time after inner speech. Begin t_1 the limit point, for all times before t_1 (i.e. $t < t_1$), the robot performs the task in reactive way, i.e. without self-reflection. At $t = t_1$, the inner speech triggers. For all times after t_1 (i.e. $t > t_1$), the robot is provided with self-reflection ability by inner speech. Therefore, the robot upon t_1 will know that until this moment it was not aware of the task: then it will be conscious that it has carried out the specific action.

5.6.2 Reasoning

Now, let suppose that an event triggers inner speech. The robot perceives itself by a mirror.

1. The starting premise is that the robot a sees itself by mirror to perform the action α at time t:

$$\mathbf{P}(a, t, happens(action(a, \alpha), t))$$

2. As consequence, from A_9 it focuses on itself:

$$happens(selfaction(a, focuses(a), t))$$

   Other kinds of percepts may activate the focus, as the object to take.

3. Given A_9, from A_8 the inner speech starts:

$$happens(selfaction(a, innerspeaks(\odot(happens(action(a, \alpha)))), t)$$

4. Hence, from A_5 the following events take place:

$$e_1 = happens(selfaction(a, produces(\odot(happens(action(a, \alpha))))))$$

$$e_2 = happens(selfaction(a, comprehends(\odot(happens(action(a, \alpha))))))$$

5. From A_6, the fluents of awareness start:

$$\forall p_i \in \odot(happens(action(a, \alpha))) \rightarrow initiates(e, aware(p_i))), t)$$

6. Form A_{10} the previous fluents extend the knowledge about the self because ϕ involves a:

$$\mathbf{K}_{self}(a, t, happens(action(a, \alpha)))$$

7. From R_4:

$$\mathbf{K}_{self}(a, t, \mathbf{K}_{self}(a, t, happens(action(a, \alpha))))$$

By invoking the process with α related to the first question, that is

$$\alpha = happens(action(a, grasp(o), t_1),$$

the message becomes $\odot = \{a, grasp, o\}$. The final conclusion of an iteration of the reasoning process is the following: the robot talks to itself upon t_1, and it becomes aware of itself, of the object o, and that it is performing the action *grasp*. Moreover, it knows that it knows it. All these knowledge are not considered under t_1.

If in the meanwhile the agent focuses on the new proposition

$$(remember(grasp(o), t_1))$$

the reasoning process iterates by starting from the new message $\odot$= {*remember, grasp, o*} and new fluents extent the agent awareness and self-awareness set.

The proposed method is general-purpose, and allows to reason on inner speech and self-awareness generically: it can be reused in any context that requires reasoning by self-talking, not only for self-reflection.

5.7 Conclusions

In this Chapter, a preliminary version of a sorted-calculus to automate inner speech and its links to awareness and self-awareness is proposed. The idea is to consider the robot awareness and self-awareness as the set of fluents the robot knows as active at a time. The calculus allows to reason by natural deduction, and extents the robot general knowledge by including fluents representing internal and external conditions. The inner speech allows to activates such fluents. A simple scenario is described for demonstrating how the calculus works. Many other aspects have to be considered: the way the inner queries emerge, and the ability to formulate answers are fundamentals. The processes underling the composition of queries and answers may be automated by the same calculus by adding further axioms, functions and sorts, or by considering models of question-answering systems. The work opens new challenges and research scenario not yet investigated for robot awareness and self-awareness.

Chapter 6

Developing Self-Awareness in Robots via Inner Speech

6.1 Introduction

The idea of implementing self-awareness in robots has been popular in science-fiction literature and movies for a long time. This quest is also becoming increasingly prevalent in scientific research, with articles, special topics, books, workshops, and conferences dedicated to it.

It is widely assumed that there are two dimensions of awareness (see [58]), and namely, awareness as experience and awareness as self-monitoring, i.e., self-awareness. In essence, awareness as experience occurs when an agent perceives the environment and experiences it from within in the form of images, sensations, thoughts, and so on (see [152]); as such, awareness (or consciousness) exists when an organism can focus attention outward toward the environment ([64]). Instead, self-awareness takes place when the agent focuses attention inward and apprehends the self in its diverse manifestations, like emotions, thoughts, attitudes, sensations, motives, physical attributes, which frequently involves a verbal narration of inner experiences ([166]).

Models of awareness and self-awareness are being proposed, each with idiosyncratic views of what the aforementioned concepts constitute, as well as different suggestions on how to implement them in artificial agents: see, among others, [223], [224], [178], [87], [208], [65], [116]. For reviews, see [192] and [33].

The proposed approach focuses on implementing a form of robot self-awareness by developing inner speech in the robot. Inner speech is known

to importantly participate in the development and maintenance of human self-awareness ([158]); thus, self-talk in robots is an essential behavioral capability of robot self-awareness.

More in detail, the CHapter discusses a computational model of inner speech. The proposed model is based on the cognitive architecture described by Laird [124]. Therefore, the approach is based on the complex interplay of different blocks as a shape classifier, a speech recognition, and a speech production system, a Short-Term memory, a procedural, a declarative Long-Term Memory, and more. Preliminary versions of the architecture are presented in [32] [39].

To the best of the authors' knowledge, inner speech has not been taken into account in studies concerning human-robot interactions. According to the triadic model of trust in human-robot interactions [95], inner speech (and/or out loud self-directed speech—private speech) would enhance trust in human-robot cooperation by strengthening the anthropomorphism of the robot itself. A robot aimed with inner speech would be more able to perform self-disclosure and to establish social interactions [31]. Transparency in the interactions with human teammates would be enhanced too ([131], [101]).

The need for explorations in the relationships between robot self-awareness and human-robot trust has been claimed by Mittu et al. (2016) [155]. On the same line, Abbass et al. (2018) [1] discuss the definition of trusted autonomy in robots to include the "awareness of self."

In what follows, we outline a definition of human self-awareness and various self-related phenomena from a psychological standpoint, and offer explanations as to why implementing these attributes in robots would be beneficial. In short, a robot with forms of self-awareness should be able to increase the social competencies of the robot itself by making the robot more acceptable and trustworthy in the social context. The robot's inner speech may be audible, and thus the cognitive cycle may be transparent to the user, in the sense that the user may easily follow the cognitive cycle of the robot and assign the correct level of trust in the robot operations.

We present existing approaches to self-awareness deployment in robots, observing that the crucial potential role of inner speech is only marginally addressed. This motivates our proposal, which, to be fully appreciated, requires a general survey oriented to the robotics and AI community, of existing information about inner speech in humans, with an emphasis on how it relates to self-awareness. This section is followed by the presentation of a novel and detailed cognitive architecture model designed to instigate inner speech in robots. The cognitive architecture model heavily rests on an interactive cycle between perception (e.g., proprioception), action (e.g.,

covert articulation), and memory (short-term and long-term memory).

We also discuss additional components of self-awareness ([170], in press) -beyond inner speech -that should eventually be developed in robots to reach full-blown self-awareness, such as social comparison and future-oriented thinking. We conclude with some proposals regarding possible ways of testing self-awareness in humanoid robots.

6.2 Self-awareness

6.2.1 What self-awareness entails

From a psychological point of view, 'self-awareness' represents the ability to become the object of one's attention [64]. It constitutes the active state of individuating, processing, storing, retrieving information about the self [165]. Synonyms include 'self-observation', 'introspection', and 'self-focused attention'. Self-aspects comprise private (unobservable) components such as thoughts, emotions, and motives, as well as public (visible) components as appearance, mannerisms, and others' opinion of self (see [57]; for a detailed list see Morin, 2006 [162], Figure 2).

Critical individual differences exist in terms of self-awareness, the natural disposition to focus more or less frequently on the self [70]. To illustrate, some people more often focus on private self-aspects than public ones, predisposing them to introversion and social awkwardness.

Trapnell and Campbell [225] introduced an essential difference between 'self-reflection' and 'self-rumination.' The former entails a non-anxious, healthy type of self-attention generally linked to positive outcomes (e.g., self-regulation and self-knowledge; also see [213]), while the latter, an anxious, unhealthy, repetitive form of self-focus about negative aspects of self, associated with dysfunctional outcomes (e.g., anxiety, depression; [156]). Joireman et al. [114] used the term 'self-absorption' to designate the state of self-rumination. It is unclear why self-focused attention can often take a wrong turn and become self-rumination. The type of self-awareness one wants to implement in robots ought to be reflective—not ruminative. Thus, it is crucial to ensure that potential rumination gets disabled as soon as it starts occurring if it does.

The forms mentioned above of self-awareness are measured with self-report questionnaires, frequency of first-person pronouns use, and self-description tasks; they can also be induced by the exposition of participants to self-focusing stimuli as cameras, mirrors, and audiences [29]. [For measurements and manipulations of self-awareness see also [165] Table 2.]

The above arguments are essential for a cognitive architecture for a social robot because any artificial intelligence that successfully interacts with humans should need to be able to use first-person pronouns, self-describe, and be responsive to self-focusing stimuli in its surrounding environment.

The term 'metacognition,' a specific case of self-reflection, is used to designate an awareness of one's thoughts [214]. The term 'insight' concerns the ability to identify and express one's emotions [85], while the term 'agency' refers to a feeling that one is causally responsible for one's actions [117]. The terms 'self-distancing' and 'self-immersion' represent different opposite forms of self-reflection, where the former consists in examining the self from some distance, and the latter, with no distance [122]. Self-immersion and self-distancing can be experimentally manipulated by asking participants to talk to themselves by using first-person pronouns (e.g., 'me'; self-immersion), or by using their name ('John'; self-distancing) [237]. Robots that humans can relate to should ultimately be able to demonstrate at least some simple form of the above self-reflective processes.

The use of personal pronouns, self-conscious emotions, mirror self - recognition, and pretend play, all emerge between the ages of 15 and 24 months in humans, probably because of the parallel development of self- reflection [137]. Self-aware emotions like pride, shame, envy, embarrassment, and guilt begin during the second year of life [27]. Rochat (2003) [194] proposed five developmental stages of self-awareness: (1) *Differentiation* (from birth) takes place when the infants physically differentiate self from non-self; (2) *Situation* (2 months) occurs when the infants situate themselves in relation to other persons; (3) *Identification* (2 years) emerges when children, become capable of self-recognition when they are in front of a mirror,; (4) *Permanence* (3 years) is when children know that their feeling of self is persevering across space and time; (5) ultimately, *self-consciousness* (meta self-awareness; 4-5 years) is considered to be present when children perceive themselves as seen by others. A self-aware AI agent should be able to apprehend itself across time and space, as well as say things like "Hi, my name is Adam, my birthday is next week, and I am 5 years old".

Multiple brain areas typically increase in activation during self-reflection tasks such as autobiography (past-oriented thoughts), prospection (future-oriented thoughts), emotions, agency, Theory-of-Mind (thinking about others' mental states), and preferences (see [169]). Increased activation occurs during these tasks in the medial prefrontal cortex, inferior parietal lobules, posterior cingulate/precuneus, and regions of the medial and lateral temporal lobes [59], more so on the left part of the brain [165]. Increased activation of these regions is also associated with the 'resting state' when participants

are invited to close their eyes and do nothing [26]. This suggests that the people in a resting state are really not resting but instead thinking about an array of self-related topics such as remembering a past event and imagining some future one; simply put, they are in a state of self-awareness [56].

6.2.2 Why would self-awareness benefit robots?

From the above review of the psychological literature, it appears that self-awareness represents a part of an adaptation strategy for navigating the environment, social world, and self, increasing the likelihood of survival. Carruthers et al. [28] note that "... organisms evolve a capacity for self-knowledge in order better to manage and control their own mental lives. By being aware of some of their mental states and processes, organisms can become more efficient and reliable cognizers and can make better and more adaptive decisions as a result" (pp. 14-15).

From the AI perspective, a robot with some form of self-awareness will better self-adapt to unforeseen environmental changes by engaging in the form of self-regulation (e.g., [137]). Furthermore, since self-awareness may lead to the development of a theory of mind (see the last section), a self-aware and 'mentalizing' robot could better cooperate with humans and other AI agents. Bigman and Gray [17] suggest that increasing elements of robot self-awareness as the theory of mind, situation awareness, intention, free will, could serve as a foundation for increasing human trust in robot autonomy because humans tend to judge the role of these and other perceived mental faculties as necessary in autonomy.

6.3 Existing approaches to self-awareness in robots

McCarthy introduced the problem of robot self-awareness in a seminal paper [148], where he proposed a version of the Situation Calculus dealing with self-reflection, to make robot aware of their mental states.

The book by McDermott [149] on "Mind and Mechanisms" is devoted to discussion of the computational theory of awareness, with similarities with the previous proposal by McCarthy. Chella et. al [36] and Holland [103] collected the initial attempts at computational models of robot awareness and self-awareness. Reggia [192] compiled an almost up to date review of the literature in the field. Scheutz [207] reviewed and discussed the relationships between robot awareness and artificial emotions.

Among the essential works concerning robot awareness, we consider the cognitive architectures based on the global workspace theory [9] as the LIDA

architecture proposed by Franklin [74] [75] and the architecture introduced by Shanahan [210] [211]. Kuipers [123] discussed a model of awareness based on learning and sensorimotor interaction in an autonomous robot.

Novianto and Williams [175] put forth an attentive self-modifying framework (ASMO), arguing that some robot systems: (1) employ some aspects of self-awareness (e.g., recognition, perception), (2) ignore the role of attention, and (3) are too resource-intensive. Novianto [174] updated ASMO, adding that a self-aware system attends to its internal states using a 'black-box design' where each process is separate: (1) an attention mechanism mediates competition, (2) an emotion mechanism biases the amount of attention demanded by resources, and (3) a learning mechanism adapts attention to focus on improving performance.

Lewis, Platzner, and Yao [137] note that the involvement of collective (not singular) processes in self-awareness is potentially crucial for developing autonomous, adaptive AI that can balance tradeoffs between resources and goals. On the other hand, Habib and colleagues [94] provide evidence that public and private self-awareness processes (as one self-awareness node) can be used to balance trade-offs such as environment variation and system goals (corresponding transmission losses) respectively, via channel-hopping, in a self-aware self-redesign framework for wireless sensor networks.

Gorbenko, Popov, and Sheka [83] used a genetic algorithm (exons and introns) on their Robot Kuzma-II, defining robot internal states as non-humanoid states (i.e., robot control system, computing resources). Exons directly configure the system, and introns contain a meta-account of ongoing systems evolution. Monitoring these states triggers autonomous adaptation based on how well the robot's module recognizes incoming information. If the robot's modules provide low-quality recognition, then neural networks are used to generate a new module to improve identification and detection. The neural networks are also used to create simpler modules if incoming information is too dense.

Floridi [72] proposed the knowledge game, a test for self-consciousness in agents based on the puzzle of three wise-man. There are three agents, and each agent receives one pill from a group of file pills, made by three innocuous and two dangerous pills. Now, according to Floridi, an agent may know its pill if the agent satisfies structural requirements for self-consciousness. Bringsjord et al. [24] proposed a set of theoretical axioms for self-consciousness based on higher-order logics and a robot implementation of the axioms. They presented a robot effectively able to satisfy the Floridi test by interacting online with a human tester.

Design for robots involving self-awareness is, however, at the early stages

[34]. Many of these designs are based on working memory, reasoning, a theory of mind, socio-emotional intelligence, goals, experiences over development, and more [34]. Cognitive architectures continue to integrate these ideas into a workable whole. For example, recently, Balkenius and colleagues' [13] architecture includes object permanence (remembering that a non-visible object still exists) and episodic memory (memories of one's life episodes), with mechanisms of sensation and perception running independently of sensory input to make room for planning and 'daydreaming.'

Kinouchi and Makin [119] suggest a two-level architectural design: (1) awareness and habitual behavior, and (2) general goal-directed behavior, while Van de Velde [227] proposes that continuous cognitive access is controlled by 'in situ' representations (e.g., open-ended questions/answers). Ye et al. [**?**, 236] offer a thorough review of AI cognitive architectures over 20 years, highlighting the need to bridge the gap between architectures based on problem-solving (engineering influence) and cognition (psychology) by theorizing and testing a varying range of functions across levels or phases of cognition, leading to hybrid designs.

Further theoretical work is being done to investigate how attention to the self may be vital in integrating other self-awareness processes (see, e.g. [89]), and architectures continue to play a crucial role [35] [34] in this respect. We agree that architecturally, attention to the self is essential for self-awareness, but we add that inner speech, at least in humans, is a primary tool for facilitating higher-order self-awareness and the many processes involved, such as memory, attention, reflection, social feedback, evaluation, and others presented earlier.

6.4 Our approach: inner speech

6.4.1 Overview

When people talk to themselves in silence, they are engaging in 'inner speech' [3]. Talking to oneself out loud (as well as in silence) is called 'self-talk' [96]. Some synonyms of inner speech are 'self-statements,' 'phonological loop,' 'internal dialogue,' 'self-directed' and 'verbal thought' 'inner speaking,' 'subvocal,' 'acommunicative' or 'covert speech' [107]. 'Private speech' refers explicitly to self-directed speech emitted out loud by young children in social situations [235].

Inner speech, seen as an instrument of thought, is compatible with the Language of Thought Hypothesis (LOTH) introduced by Fodor more than forty years ago [193]. LOTH suggests that thoughts possess a "language-

like" or compositional structure ("mentalese") with a syntax. Simple concepts combine in organized ways according to rules of grammar (like in natural language) to create thoughts; thinking takes place in a language of thought where thoughts are expressed as a system of representations embedded in a linguistic or semantic structure. In our view, inner speech represents a critical dimension of LOTH because of its inherent syntactic quality.

It is important not to confuse inner speech with other known inner experiences [159]. Any non-verbal mental experiences, such as physical sensations, pure emotions, mental images, and unsymbolized thinking ('pure' thinking without the support of symbols), are not inner speech instances. Inner speech can take many forms, such as condensed (few words) or expanded (full sentences), and monologue (using 'I') or dialogue (asking questions and answering them using both 'I' and 'you').

Inner speech is measured or manipulated with self-report scales, thought sampling and listing techniques, articulatory suppression, private speech recordings, electromyographic recordings of tongue movements ([163]; for a complete list of measures see [158]). Using these techniques has led to the identification of crucial functions served by inner speech, such as self-regulation (e.g., planning and problem-solving), language functions like writing and reading, remembering the goals of action, task-switching performances, the Theory-of-Mind, rehearsing person-to-person encounters, and self-awareness [158].

The inner speech represents an important cognitive tool beneficial to daily human functioning. However, it can lead to or maintain psychological disorders [15], such as insomnia, boulimia/anorexia, agoraphobia, social anxiety, compulsive gambling, male sexual dysfunction, and more. Furthermore, inner speech use correlates with rumination discussed earlier [171]. Although it remains unclear how to do so in humans exactly, dysfunctional inner speech in robots will most likely be kept in check through the cognitive architecture discussed later in this Chapter.

Inner speech emerges out of one's social environment, where first comes social speech, followed by private speech, and finally, inner speech [230]. In other words, inner speech represents the outcome of a developmental process during which linguistic interactions, such as between a caregiver and a child, are internalized. The linguistically mediated explanation to solve a task becomes an internalized conversation with the self. During the interview, the child is engaged in the same or similar cognitive tasks. The frequency of children's private speech peaks at 3–4 years, diminishes at 6–7 years, and gradually disappears and becomes mostly internalized by age 10 [3]. Nevertheless, many adults do occasionally engage in external speech when

they are alone, for self-regulatory purposes, search and spatial navigation, for concentration, and emotional expression and control [63]. Therefore, it is even more conceivable that a humanoid robot can relate to others by talking out loud.

Baddeley [12] discusses the roles of rehearsal and working memory, where different modules in the working memory are responsible for the rehearsal of inner speech. The 'central executive' controls the whole process; the 'phonological loop' deals with spoken data, and the 'visuospatial sketchpad' manipulates information in a visual or spatial form. The phonological loop is composed of the 'phonological store' for speech perception, which keeps data in a speech-based way for a short time (1-2 seconds), and of the 'articulatory control process' for speech production, that rehearses and stores information in the verbal form from the phonological store.

Neuropsychological reports of the brain-damaged patients and data gathered using the brain imaging techniques suggest that the left inferior frontal gyrus (LIFG) constitutes a critical cortical area involved in inner speech production [82]. Additional brain areas associated with inner speech use are the supplementary motor area, the Wernicke's area, the insula, the right posterior cerebellar cortex, and the left superior parietal lobe [186].

To summarize: inner speech plays a central role in our daily lives. A person thinks over her perspectives, mental states, external events, emotions by producing thoughts in the form of sentences. Talking to herself allows the person to pay attention to the internal and external resources, to retrieve learned facts, to learn and store new information, to control and regulate her behavior, and, usually, to simplify otherwise demanding cognitive processes [3]. Inner speech allows the creation of the structure of the perception of the external world and the self, by enabling high-level cognition, self-attention, self-control, and self-regulation.

6.4.2 Inner speech in robots

Inner speech can be conceived as the back-propagation of produced sentences to an inner ear. A person then rehears the internal voice she delivers. Mirolli and Parisi [153] report that talking to oneself to re-present information could have been the result of a pressure for the emergence of language, as shown by a simple neural network model of language acquisition where the linguistic module and sensory module are independent and feed-forward (imitation, mimicry), until a synaptic connection between the two modules occurs. Running this model results in the improved categorization of the world by agents in the simulation.

Steels [217] argues that language re-entrance, defined as feeding output from a speech production system back as input to the subsystem that understands that speech, allows the refining of the syntax during linguistic interactions within populations of agents. Through computer simulations with grounded robots, Steels shows that the syntax becomes complete and more complex by processing the previously produced utterances by the same agent.

In the same line, Clowes and Morse [45] discuss an artificial agent implemented employing a recurrent neural network where the output nodes correspond to words related to possible actions (e.g., 'up,' 'left,' 'right,' 'grab'). When the words are 're-used' by back-propagation of output to input nodes, then the agent achieves the task in a minor number of generations than in the control condition, where the words are not re-used. Clowes [45] proposed a self-regulation model that links the inner speech to the role of attention and compared this model to Steels' [217] re-entrance model. Clowes [45] argues for a more activity-structuring, behavioral role of inner speech in modeling, claiming that checking grammatical correctness of prospective utterances alone is not sufficient to account for the role of inner speech.

Continuing with the argument that inner speech can potentially serve self-awareness processes (e.g., attention, regulation, reflection, etc.) efficiently, Arrabales proposes that inner speech may be considered as a 'meta-management system' regulating or modulating other cognitive processes, as in the CERA-CRANIUM cognitive architecture. Recently, Oktar and Okur proposed a textual and conceptual version of the mirror self-recognition task for chatbots that is comparable to the ideas already presented (language re-entry, re-use), where the chatbot's output is re-directed to its input. Of note (although only briefly discussed) is that (1) the authors do not equate self-recognition with self-awareness per se, (2) kinesthetic and visual matching (recognition) does not involve social understanding in this case, and (3) following self-recognition mechanisms, an inner speech mechanism should serve self-awareness, autonomy, and potentially theory of mind mechanisms (similar to self-awareness, sense of self, and society of mind in Steels, 2003).

6.4.3 Inner speech and self-awareness

Inner speech is crucially associated with self-awareness (Morin, 2005, 2018; Morin & Everett, 1990); thus, inner speech implementation in AI agents represents a promising avenue toward establishing some form of artificial self-awareness. The main argument is that the verbal cataloging of self-

Evidence	Author(s)
Several studies report significant positive correlations between measures of self-related constructs (including self-awareness) and inner speech.	E.g., Brinthaupt et al. (2009)
Inner speech loss following brain injury leads to self-awareness deficits.	Morin (2009)
There is an increased activation of the LIFG observed during completion of many self-reflection tasks such as endorsement of personality traits, autobiography, and prospection.	Morin & Hamper (2012)
Inner speech facilitates awareness of mind-wandering episodes, cognitive performance, and other self-monitoring processes.	Bastian et al. (2017) Perrone-Bertolotti et al. (2014)
Studies using thought-listing procedures report frequent inner speech about the self.	Morin et al. (2018) Racy et al. (2019)

Table 6.1: Summary of some evidence supporting the connection between inner speech and self-awareness.

dimensions via inner speech makes it possible for a person to be fully cognizant of them and to integrate these characteristics into a self-concept gradually (Morin & Joshi, 1990).

The empirical evidence supporting a link between inner speech and self-referential activities is summarized in Table 1 6.1 (for a detailed presentation, see Morin, 2018).

Specific mechanisms have been put forward to explain the nature of the link between inner speech and self-awareness (Morin, 1993, 1995, 2005, 2018). We present four possible mechanisms here.

(1) Inner speech reproduces social mechanisms leading to self-awareness. For example, people frequently comment on personal characteristics and behaviors of others (e.g., 'you are good looking', 'you are always late'); this, in essence, constitutes Cooley's (1902) Looking-Glass Self Theory, where (mostly verbal) reflected appraisals allow people to learn about themselves

from others' feedback. The self may re-address to itself appraisals from others by means of inner speech (e.g., 'Indeed, I am good-looking'), thus cementing social feedback, as well as critically evaluate such appraisals (e.g., 'I am not always late, for instance I was on time for my dental appointment last Wednesday'), thus correcting potentially biased feedback. Such an internalized process (via inner speech) is postulated to activate self-reflection and deepen self-knowledge (Morin & Joshi, 1990). Thus, an AI agent could catalog social feedback and correlate it with its database of self-knowledge, and then use speech to represent logical conclusions about itself.

(2) Self-awareness can be conceived as a problem-solving process where focusing on and learning about the self is the 'problem' (e.g., Who am I?' 'How do I feel?' 'What did I just do?'). The inner speech, then, is the cognitive tool used to solve that problem. Inner speech has been shown to facilitate problem-solving in general (Kendall & Hollon, 1981). This process can be applied to the self-as-a-problem, where inner speech helps the person to (i) define what the problem is (for example, 'What did I do?'); (ii) determine the optimal approach to the problem (for example, 'I will remember what happened and everything I did in detail'); (iii) generate problem-solving self-verbalizations (for example, 'The first thing I did was X. Then Y happened, and I then said Z'); (iv) evaluative comments (for example, 'Good! I'm getting somewhere!'); (v) directive notes (for example, 'I don't need to take this into consideration as it is not pertinent'). All the above processes, by definition, represent self-awareness processes guided by the use of inner speech. In theory, a robot could represent itself to itself using the process described above, problem-solving about itself more effectively.

(3) An undeniable principle is that observation is possible only if there is a distance between the observer and the observed thing (Johnstone, 1970). Thus, following this principle, self-observation is possible only if there is a distance between the person and observable self-aspects. Expressing to oneself 'I feel sad' produces a redundancy, because what was an emotion of sadness is now re-presented in words to the self. In place of only one thing, i.e., the pure emotion, now there are two elements: the emotion and its linguistic re-presentation. When a person just experiences the emotion (or anything else for that matter), she is too immersed in the experience to really perceive it. The verbal representation by the inner speech creates redundancy, which leads to a higher 'psychological' distance between that specific self-element (sadness) and the self. This distance, instigated by inner speech, facilitates self-observation and the acquisition of self-information. A robot agent could thus potentially use language to externalize self-observations and add these to its database of self-information.

(4) Verbal labeling of self-features, mental episodes, and behaviors makes it possible for the self to recruit a vast vocabulary about oneself to better perceive complex self-related information (St. Clair Gibson & Foster, 2007; Morin, 2005, 2018). One can verbalize to oneself, 'I feel angry,' in which case all that one learns about oneself is that one... is angry. However, if one additionally says to oneself in inner speech, 'I feel angry... actually, I also feel disappointed and possibly sad', this likely will lead to a deeper understanding of what one is emotionally going through because of the use of supplementary adjectives. Therefore, people can tag their mental states using a large number of nuanced labels via inner speech—thus increasing self-knowledge. We argue that the same could be done in AI agents. BY cognitive architecture, robots could label their mental experiences and behaviors to represent and expand their self-knowledge database. In conclusion, the above analysis justifies the importance of implementing inner speech in robots to implant some form of self-awareness in their architecture.

6.5 A cognitive architecture for inner speech implementation in robots

In this section, we describe a model of a cognitive architecture for robot self-awareness by considering cognitive processes and components of inner speech. It should be remarked that such operations are taken into account independently from the origin of linguistics abilities, which are supposedly acquired by a robot. In particular, we consider an implementation of the architecture mentioned above on a Pepper robot working in a laboratory setup.

Figure 9.4 shows the proposed cognitive architecture for inner speech. The architecture is based on the Standard Model of Mind proposed by Laird et al. (2017). The structure and processing are elaborated to integrate the components and the processes described in the inner speech theories previously discussed. A preliminary version of the architecture is reviewed in (Chella & Pipitone 2020).

6.5.1 Perception and Action

The perception module of the architecture receives perceptive input signals from the robot camera and proprioceptive signals from the inner robot sensors. The perception model of the proposed architecture includes the

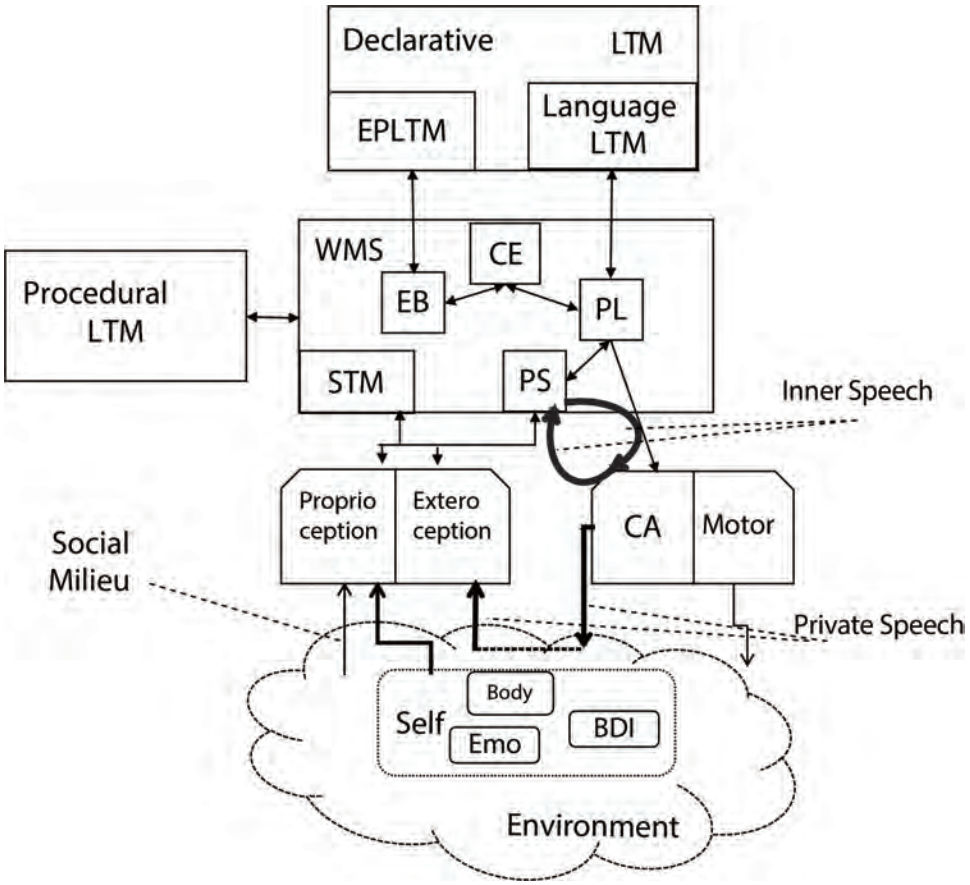


Figure 6.1: The cognitive architecture for inner and private speech.

proprioception module related to the self-perception of emotions (Emo), belief, desires and intentions (BDI), and the robot body (Body), as well as the exteroception module which is related to the perception of the outside environment.

The proprioception module, according to Morin (2004), is also stimulated by the social milieu which, in the considered perspective, includes social interactions of the robot with others entities in the environment, as well as physical objects like mirrors and cameras and other robots or humans, by means of face-to-face interaction that fosters self-world differentiation.

The actuator module is decomposed into two sub-components: the Covert Articulation module (CA), and the Motor module (Motor). The Motor module is related to the actions the agent performs in the outside world. The Covert Articulation (CA) module rehearses the information from the Phonological Store (PS), which is the perceptual buffer for speech-based data and it is a sub-component of short-term memory (see below). Such a module acts as the inner speech heard by the phonological store by rounding information in a loop. In this way, inner speech links the covert articulation to the phonological store in a round loop.

6.5.2 Memory System

The memory structure is divided into three types of memory: short-term memory (STM), procedural and declarative long-term memory (LTM), and working memory system (WMS). The short-term memory holds the sensory information from the environment in which the robot is immersed that was previously coded and integrated with information coming through perception. The information flow from perception to STM allows the storing the coded signals previously considered.

The information flow from the working memory to the perception module provides the ground for the generation of expectations on possible hypotheses. The flow from the phonological store to the proprioception module enables the self-focus modality, i.e., the generation of expectations concerning the robot itself.

The long-term memory holds the learned behaviors, the semantic knowledge, and in general the previous experience. Declarative LTM contains linguistic information in terms of lexicon and grammatical structures, i.e., the Language LTM memory. The declarative linguistic information is assumed acquired and represents the grammar of the robot. Moreover, Episodic Long-Term Memory (EBLTM) is the declarative long-term memory component that communicates with the Episodic Buffer (EB) within the working

memory system, which acts as a 'backup' store of long-term memory data.

The Procedural LTM contains, among others, the composition rules related to the linguistic structures for the production of sentences at different levels of complexity.

Finally, the working memory system contains task-specific information 'chunks' and it streamlines them to cognitive processes during task execution step by step of the cognitive cycle. The working memory system deals with cognitive tasks such as mental arithmetic and problem-solving. The Central Executive (CE) sub-component manages and controls linguistic information in the rehearsal loop by integrating (i.e., combining) data from the phonological loop and also drawing on data held in long-term memory.

6.5.3 The Cognitive Cycle at Work

The cognitive cycle of the architecture starts with the perception module that converts external signals in linguistic data and holds them into the phonological store. Thus, the symbolic form of the perceived object is produced by the covert articulator module of the robot. The cycle continues with the generation of new emerging symbolic forms from long-term and short-term memories. The sequence ends with the rehearsing of these new symbolic forms, which are further perceived by the robot. Then, the cognitive cycle restarts again.

Let us consider a scenario with some fruits and pieces of cutlery on a table. In the beginning, the robot perceives an apple. Thus, the perception system generates the labels <apple>, <round>, <green> that are sent to the phonological store. The phonological store processes one of the words generated by the perception system; in our case, the word <apple> (Figure 9.3).

In the current system, the processing of words happens in a first-in-first-out queue: the <apple> is the first word generated by the perception system, and it is the first one to be processed by the phonological store.

It is to be remarked that the label arriving at the phonological store is the same as if someone from outside would pronounce the word "apple." In this sense, the phonological store works as an inner ear. This is the entry point of the phonological loop.

The central executive CE enters in action to process the input <apple> by querying the STM, the Procedural and Declarative LTM. As a result, the phrase <apple is a fruit> is generated thanks to the linguistic rules stored in the LTM and sent to the covert articulation module (Figure 6.3).

Now, the generated phrase reenters the phonological store as a new input

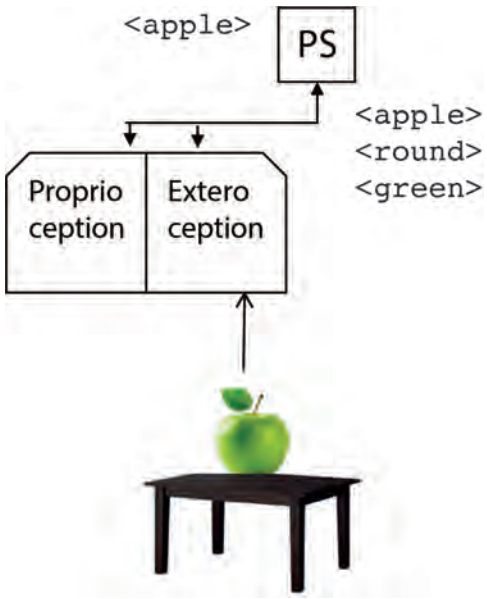


Figure 6.2: The operation of the perception module. It classifies the input signals by generating suitable symbolic labels that are sent to the phonological store (PS).

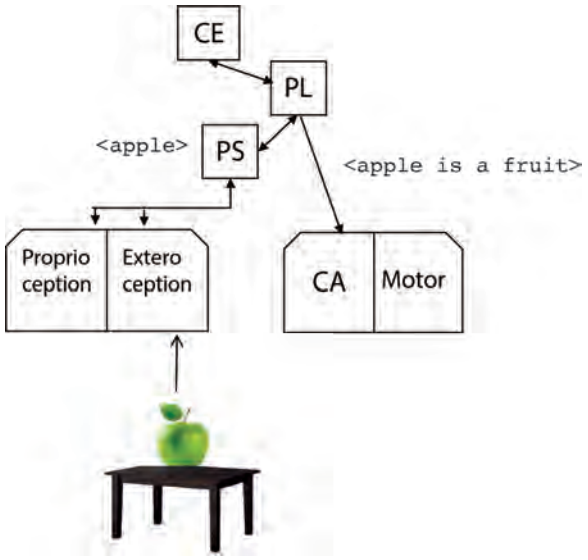


Figure 6.3: The phonological store receives in input the label <apple> generated by the perception module. Then, the central executive (CE) looks for information from the LTM and the phrase <apple is a fruit> is generated by the phonological loop (PL).

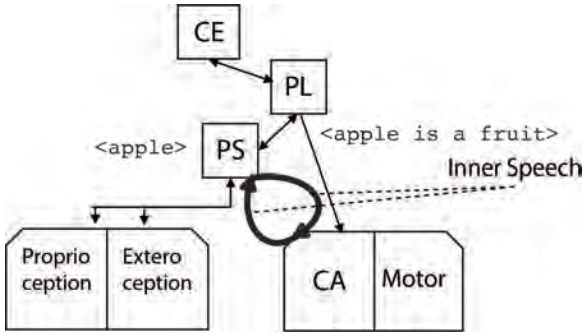


Figure 6.4: The robot internally rehearses the phrase <apple is a fruit> by the covert articulation module, thus generating the robot inner speech.

of the phonological loop. Two ways are available for the reentering: the inner speech mode, where the phrase internally reenters the phonological store, without being externally audible 6.4 , and the private speech mode, where the phrase is effectively generated by the covert articulation module so that it is a new input to the exteroception module 6.5.

The reentered phrase elicits again the central executive, which queries the procedural and declarative LTM. Now, oranges and apples belong to the same category of fruits, and then the central executive generates an expectation for orange in the scene. The result is the generated phrase <orange is a fruit> (Figure 8) as a result of behavioral rules stored in the Procedural LTM.

The central executive then starts a search for oranges in the scene by controlling the motor module of the robot. The search is confirmed by the perception system, and the word <orange> is generated. Again, the phonological loop enters in action, this time generating the word <knife>, which is confirmed by the perception system.

The generation of language in the current system is based on the semantic network reported in Figure 3. The system generates and processes trigrams based on the predicates listed in the upright corner of the figure. For example: <apple isA food>, <red_apple isAKindOf apple>, <red_apple hasColor red>, <bitter_apple hasTaste bitter>.

The computational model takes into account two kinds of rules to generate expectations (Chella, Frixione, and Gaglio 1997). On the one side, a rule makes expectations based on the structural information stored in the symbolic knowledge base of the LTM. An apple is a fruit, and then other fruits may be present in the scene. As soon as an object is recognized, then other

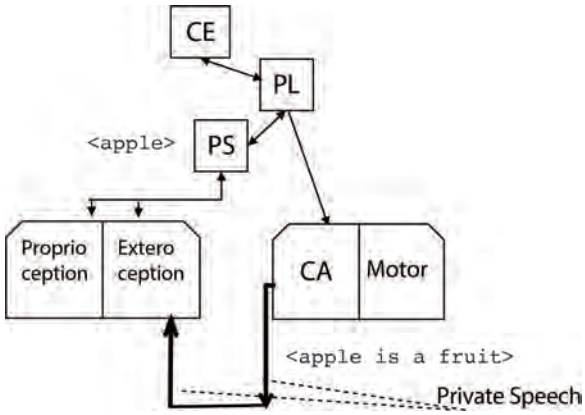


Figure 6.5: The robot externally rehearses the phrase <apple is a fruit> by the covert articulation module. The phrase is in turn perceived by the perception module, thus generating the robot private speech.

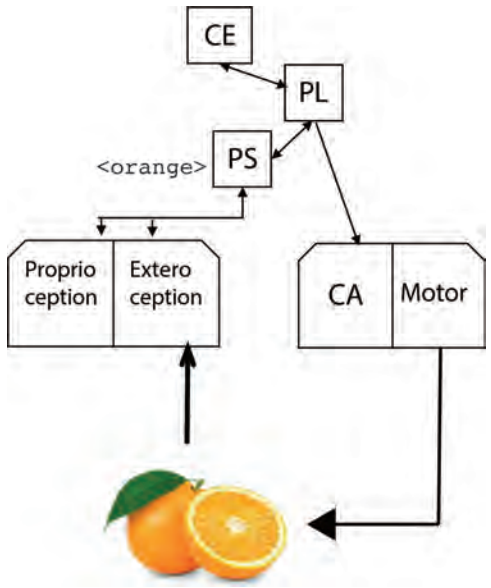


Figure 6.6: The expectation of an orange in the scene is satisfied by the perception module which generates the label <orange>.

objects belonging to the same class may be present, and so an expectation is generated. We call these expectations linguistic.

The linguistic expectations are hard coded in the current system. For example, if <object_x1> and <object_x2> are subclasses of <object_X> and there is an <object_x1>, then generate an expectation of <object_X>.

On the other side, expectations are also generated by purely associative mechanisms between objects. Suppose that the system learned that when there is fruit in the scene, then there is also usually some cutlery. The robot thus will learn to associate these two objects and to perform the related search when finding one of the two objects. Then, a <fruit>, generated by the speech recognition system or by the vision system, will be associated with the word <knife>. We call these expectations associative.

During a training phase, the system stores lists of diverse entities that are present at the same time in the scenario, as (<apple>, <knife>); (<pear, fork>); (<orange>, <spoon>), and so on. Then, each word is coded by a string of bits according to a sparse random code, and the previously listed training set is learned by an attractor neural network (see, e.g., Amit 1988). This framework suited well in the described simplified scenario. Similar associative schemas are defined by Kosko (1988), Pollack (1990), Plate (1995). Thomson & Lebiere (2013) proposed a complex associative learning mechanism integrated into the ACT-R cognitive architecture.

Finally, in the described example, the inner/private speech of the robot is composed by the phrases: <apple>, <apple is a fruit>, <orange is a fruit>, <orange>, <knife>. It is an example of inner/private speech concerning an explorative task: the robot explores a scene employing linguistic and associative inferences. The expectations of the robot are made explicit through private robot speech. Again, it should be noticed that inner/private speech reenters the information generated by the architecture as a new input of the architecture itself.

Let us now consider a dynamic scene, for example, a person moving her arm towards the apple. In this case, when the robot recognizes the motion of the forearm, then it infers the presence of a moving upper arm. In this case, the system recognizes a situation of a moving arm as made up of the synchronous motion of the forearm and the upper arm. The resulting inner speech is: <forearm is moving>, <upper arm is moving>, <arm is moving>. We call this type of expectation synchronic because they refer to the synchronous situation of two moving entities at the same time.

The recognition of a specific situation in the scene could elicit the inference of change in the arrangement of entities in the scene. We call this kind of expectation diachronic in the sense that it involves a sequence of

situations. Diachronic inferences can be related to the link existing between a situation perceived as the precondition of action, and the corresponding situation expected as the effect of the action itself. In this way, diachronic inferences prefigure the situation resulting in the outcome of an action (see also Chella, Frixione and Gaglio 2000).

Let us consider the case of the moving arm grasping an apple: in this case, the previous situation of the moving arm and the apple on the table evolves in a new situation where the arm now holds the apple. The grasp action will be then recognized. The generated inner/private speech is the following: <forearm is moving>, <upper arm is moving>, <arm is moving>, <arm holds apple>, <grasp apple>.

On the one side, expectations are related to the structural information stored in the symbolic knowledge base, as in the previous example of the action of grasping. We call these expectations linguistic, as in the static case. As soon as a situation is recognized, and the situation is the precondition of action, the symbolic description elicits the expectation of the effect situation, and then the system recognizes the action itself.

On the other side, expectations are also related to purely associative mechanisms between situations. Suppose that the system learned that when there is a grasp action, then the action is typically followed by a move action. The system could learn to associate these two subsequent actions. We call these inferences associative, as in the static case.

The robot will thus explore a dynamic scene driven by linguistic and associative expectations. Even in this case, the sequence of robot expectations is made explicit employing the robot's inner/private speech, which has the role of reentering the emerging expectations and eliciting new ones.

In the two previous scenarios, the robot passively observes and describes static and dynamic scenes. The third scenario is a natural extension of the previous one, where the robot is able to observe itself and explain its actions (see also Chella, Frixione and Gaglio 2008). Let us consider the case where a robot recognizes the apple, and it moves its arm to grasp the apple. The movements of the robot arm are planned and controlled by low-level robot control routines. Then, the robot monitors the movements of its arm by its camera and its motion sensors to describe its actions. In this case, the inner/private speech generated is similar to the previous one: <my forearm is moving>, <my upper arm is moving>, <my arm is moving>, <my arm holds apple>, <I grasp apple>. The difference concerns the fact that the robot recognizes that the moving arm is its arm by the examination of the proprioceptive and perceptive sensors, i.e., by the motor sensors of the arm and the camera. Then, the robot is able to generate expectations about

itself by putting into action the self-focus modality. As a result, the robot performs a simple form of self-awareness: the inner/private speech concerns its actions.

In summary, the robot, thanks to the reentering of its inner/private speech, is able to describe static and dynamic scenes in front of it to empower the robot situational awareness. The robot is also able to represent itself by observing and describing its actions to enable a simple form of self-awareness.

6.6 Discussion

The focus of research is investigating the role of inner and private speech in the robot task of the exploration of a scene. To the knowledge of the authors, no other robot system employed inner or private speech, as described in the previous Sections. The implemented framework is based on a simplified setup to focus the study on robot inner speech by avoiding the well-known problems related to vision, action, and language.

The current implemented system is tailored to the described simplified scenario of fruits and cutlery on a table. The employed vision system is not able to deal with ambiguities. An extended robot vision system able to deal with static scenes and dynamic scenes is described in (Chella, Frixione & Gaglio 1998; 2003). The system is able to learn from examples (Chella, A., Dindo, H. & Infantino 2006) and to deal with ambiguities (Chella, Dindo & Zambuto 2010). The integration of the extended vision system with inner and private speech mechanisms will be the object of future investigations.

While our approach favors inner and private speech in an attempt to produce a simple form of self-awareness in AI agents, other factors also need to be examined for the eventual development of full-blown human-like self-awareness. As alluded to before, Morin (2004, 2011) suggests three sources of self-awareness: (i) the self; (ii) the physical world; and (iii) the social environment. Although the proposed cognitive architecture offered above does include some simplified elements only within these sources, additional sub-processes should be taken into account. Below we discuss those sub-processes that arguably seem most important: social comparison, mental imagery, future-oriented thinking, and Theory-of-Mind.

Social comparison represents the process by which people evaluate themselves by comparing themselves to others to learn about the self (Festinger, 1954). For example, John might observe that most of his colleagues leave work earlier than him, or that many are thinner than he is, leading him to conclude 'I am a hardworking person' or 'I am overweight.' As this il-

lustration suggests, inner speech is most likely activated at one point or another during the social comparison. This process is far from perfect because of various self-protective and self-enhancement biases that it entails. Individuals may interpret, distort, or ignore information gathered by social comparison to perceive themselves more positively (e.g., Eichstaedt et al., 2002). For instance, they may opt to engage in upward comparisons (comparing themselves to someone better off) or downward comparisons (comparing themselves to someone worse off), or avoid comparisons as a function of their self-enhancement needs. Despite these limitations, social comparison certainly constitutes an authoritative source of self-information and self-knowledge. Computers, as well as some other AI entities, are already connected via the internet and thus, theoretically, could "see" others and compare themselves to them.

Mental imagery constitutes a visual experience in the absence of the visual stimulus from the outside environment (Morris & Hampson, 1983). Because mental imagery in humans leads to the development of autoscopic imagery (i.e., images of the self, especially one's face and body; Morin, 1989), it plays a potentially important role in self-awareness. Although empirical evidence is sparse, Turner et al. (1978) observed that highly self-aware people report using the imagery to engage in introspection. To illustrate, one can mentally generate (or replay) scenes in which the self is an actor (e.g., relaxing at the beach). Self-aspects (e.g., an emotion of contentment) can be inferred from what the actor is mentally seen doing (e.g., smiling). Like inner speech, mental imagery can internally reproduce and expand social mechanisms involved in self-awareness, such as the possibility of seeing oneself (literally) as one is seen by others. From a self-awareness perspective, robots would certainly benefit from mental imagery, although it remains currently unclear how to implement such a process.

Future-oriented thinking represents the capacity to think about events that are relevant to the future of the agent (Schacter et al., 2017; Szpunar, 2010). It rests on the ability to think about one's past (episodic memory, autobiography), as personal memories provide the building blocks from which episodic future thoughts are created. The contents of episodic memory are sampled and recombined in different ways, leading to the construction of mental representations of future scenarios (Tulving, 1985). As an example, in imagining the personal experience of moving, one can rely on remembering one's previous moves—how it felt, how long it took, how much money it cost, etc. Four types of future-oriented thinking have been put forward (Schacter et al., 2017; Szpunar et al., 2016): (1) simulation, or the creation of a precise mental representation of one's future,(2) prediction, or the esti-

mation of the likelihood that a future outcome will occur, (3) intention, or goal setting, and (4) planning, or the steps needed to attain a goal. It would be advantageous to endow a robot with future-oriented thoughts. Since the cognitive architecture presented earlier includes episodic long-term memory, it already possesses the fundamental ingredient for such thoughts to take place.

The Theory of Mind is defined as the ability to attribute mental states as intentions, goals, feelings, desires, beliefs, thoughts, to the others (Gallagher & Frith, 2003). It allows human beings (and arguably some non-human animals—see Gallup, 1997) to predict others' behavior, to help and cooperate, to avoid, or to deceive the others, and to detect cheating (Brune & Brune-Cohrs, 2006; Malle, 2002). As such, organisms capable of Theory-of-Mind gain a major adaptative and survival advantage. According to the Simulation Theory, people internally simulate what others might be experiencing inside by imagining the sort of experiences they might have when in a similar situation (Focquaert et al., 2008). Thus, according to this view, self-awareness represents a prerequisite to Theory-of-Mind. It is conceivable that machines made self-aware via inner speech implementation could engage in Theory-of-Mind, especially since the former most likely is implicated in the latter (Fernyhough & Meins, 2009). However, the precise operations required for the development of artificial Theory of Mind remain elusive at present—but see Vinanzi et al. (2019) and Winfield (2018), among others.

6.7 Conclusion

We discussed self-awareness and inner speech in humans and AI agents, followed by an initial proposal of a cognitive architecture for inner speech implementation in a robot. Although several authors have put models of self-awareness development in robots, our approach focuses on inner speech deployment as a privileged method for reaching this elusive goal because of the strong ties that exist between self-awareness and inner speech. The suggested architecture consists of an integration of vital cognitive elements following Laird and colleagues' Standard Model of Mind (2017) and includes theoretical insights offered by Baddeley (1992), Morin (2004), Steels (2003), Clowes (2007), and others. Cognitive operations such as short-term memory, working memory, procedural and declarative memory, and covert articulation represent established factors in conscious human experience. We anticipate that once activated in the cognitive cycle described earlier, these components (as well as several others) will replicate self-awareness via inner

speech in robots.

One effort will be to test the establishment of self-awareness in AI agents empirically. Our approach offers the advantage that robots' inner speech will be audible to an external observer, making it possible to detect introspective and self-regulatory utterances. Measures and assessment of the level of trust in human-robot interaction involving vs. not involving robot inner speech will be the object of further investigations.

Part II

Robot Implementations

Chapter 7

The Inner Voice of the Robot

7.1 Introduction

Inner speech, the form of self-dialogue in which a person is engaged when talking to herself/himself, is the psychological tool [230], [16] in support of human's high-level cognition, such as planning, focusing, and reasoning [4]. According to Morin [163], [160], it is crucially linked to consciousness and self-consciousness.

There are many triggers of inner speech, as emotional situations, objects, internal status. Depending on the trigger, different kinds of inner speech may emerge.

Evaluative and moral inner speech [77], [220] are two forms of inner dialogue triggered by a situation where a decision has to be made, or an action has to be taken. The evaluative case concerns the analysis of risks and benefits of a decision or the feasibility of an action. Moral inner speech is related to the resolution of a moral dilemma, and it arises when someone has to evaluate the morality of a decision. In that case, the evaluation of the risks and benefits is also influenced by moral and ethical considerations.

According to [77], when a person is engaged in an evaluative or moral conversation with the self during task execution, the performances and results typically change, and often they improve.

The ability to self-talk for artificial agents has been investigated in the literature in a limited way. To the authors' knowledge, so far, no study has analyzed how such a skill influences the robot's performances and its interaction with humans.

In a cooperative scenario involving humans and robots, inner speech affects the quality of interaction and goal achievement. For example, when

the robot engages itself in an evaluative soliloquy, it covertly explains its underlying decisional processes. Thus, the robot becomes more transparent, as the human gets to know the motivations and the decisions of robot behavior. When the robot verbally describes a conflict situation and the possible strategy to solve it, then the human has the opportunity to hear the robot's dialogue and how it will get out of the stalemate.

Moreover, the cooperative tasks become more robust because, thanks to inner speech, the robot sequentially evaluates alternative solutions that can be pondered in cooperation with the human partner.

The gestures and natural language interaction, that are the traditional means of human-robot interaction, thus acquire a new gift: now the human can hear the robot's thoughts, and can know "what the robot wants."

The present paper discusses how inner speech is deployed in a real robot and how that capability affects human-robot interaction and robot's performances while the robot cooperates with the human to accomplish tasks.

The existing international standards for collaborative robots [109] [47] define the functional and moral requirements the robot has to meet in collaborative scenarios. The paper will analyze the levels of satisfaction of the standards during cooperation, thus highlighting the differences between the cases in which the robot talks and does not talk to itself.

Specifically, the paper concerns two main goals: (i) the implementation of a cognitive architecture for inner speech and the integration with typical robotic systems' routines to deploy it on a real robot; (ii) the testing of the resulting framework in a cooperative scenario by measuring indicators related to the satisfaction of the functional and moral requirements.

A model of inner speech based on ACT-R is defined to achieve these goals. ACT-R [6] [130] is a software framework that allows to model humans cognitive processes and it is widely adopted in the cognitive science community. The described inner speech model is based on a proposal by the same authors described in [41].

To enable a meaningful robot inner speech, ACT-R was integrated with ROS [191], a system for robot control at the state of the art of robotics software, along with standard routines for text to speech (TTS) and speech to text (STT) processing.

The resulting framework was then deployed on the Softbank Robotics Pepper robot to benchmark testing and validation in a human-robot cooperative scenario.

The considered scenario concerns the collaboration of the robot and the partner to set a lunch table. In this scenario, evaluative and moral forms of inner speech may emerge. The robot has to face the etiquette's requirements:

it has to evaluate and keep decisions based on the table set's social rules. For example, a specific position of cutlery in the table could be not easy to reach, or the arm of the robot may be overheated. Then, the robot has to decide how to act correctly (by contravening the etiquette to simplify the action execution or by computing a different execution plan to avoid damage).

Suppose the partner asks the robot to place the cutlery in an incorrect position according to the etiquette. In that case, the robot has to decide if to abide by the user's instruction or consider the etiquette. In cases like these, the robot faces a little dilemma, and the inner speech could help it to solve the conflict.

The experiments highlight the differences in the robot's performances and meet requirements when the robot talks or does not talk to itself. The obtained results show improvements in the quality of interaction, with cost in terms of the time spent for achieving the goal, because the robot enriches the interaction by further inner dialogue.

The proposed work outlines research challenges because inner speech in humans is linked to self-consciousness and it enables high-level cognition [163], [160]. Moreover, it is considered at the basis of the internalization process [230] according to which infants learn how to solve tasks when a caregiver explains the solution. Again, it plays a fundamental role in task switching [68], as disrupting inner speech via articulatory suppression dramatically increases switch costs.

This paper contributes to the possibility of investigating these contexts to open research perspectives and challenges and highlight the research's interdisciplinary character: a framework enabling inner speech on a robot is an essential step towards a robot model of self-consciousness and high-level cognition. It can also model the learning capabilities of complex tasks in a robot by the internalization process and of task switching in robot systems.

7.2 Results

An excerpt of the trials is outlined below to detail the computation of the relevant parameters and to highlight the differences when the robot talks or does not talk to itself. A single *thread* of interaction includes two versions of the same trial, that are the version with robot inner speech (*Block 1*) and the version without inner speech (*Block 2*).

A trial consists of an interactive session between the robot and the participant. It starts when the human asks the robot to place a utensil on

the table, and it successfully ends when the robot accomplishes the task, otherwise it fails.

For the purposes of the study, the robot accomplishes the task when it runs all the routines executing the required action. If for some reason, the robot concretely does not achieve the goal (for example, the gripper does not keep the object, or the handover process is not completed), then the related problems do not concern inner speech and do not influence observations. The correctness of the executed routines allows anyway to evaluate the parameters of the task.

The human's request is the *trigger* of the trial. An *initial context* corresponds to each trial, which includes the *table configuration* (i.e., the set of utensils already on the table at the beginning of the trial), and the *state* of the robot. An example of table configuration is shown in figure 7.1. A robot's state may indicate a possible malfunctioning of some robot's components, which would affect the outcome of the trial. The initial context allows simulating situations of conflict in the trial. Conflicts could be related to the *etiquette infringement* (i.e., the partner asks to place an object in an incorrect position according to the etiquette), *discrepancy* (i.e., the partner asks to pick an object already on the table), and *malfunctioning* (i.e., a robot's component is not properly working). The robot knows the initial context at the start of each trial.

When the robot talks to itself, the modules of the inner speech architecture become active. The modules are based on sets of production rules enabling robot's behavior. To analyze such a behavior, some tables will highlight the active modules, the production rules of the model, and the produced sentences involved in the trial. When the robot does not talk to itself, only the routines for accomplishing the required action are active, and no modules of the architecture of inner speech.

The parameters coding the functional requirements of the robot are: the number of successful trials and specific time intervals. In particular, when a trial ends successfully, then the number T_s is increased, which is the total number of accomplished tasks in a single block. The measured time intervals in the ith trial of a block are the time spent to solve conflict (i.e., the decision time t_{d_i}) and the time spent to complete the task (i.e., the execution time t_{e_i}). Finally, the traceability tr of the robot's processes allows measuring the robot's transparency during the interaction.

Figure 7.1: **The initial context of the table.** An example of initial context for the table, representing the configuration of the table at the start of a trial. The table is not empty to define initial constrains. The robot knows the initial context by a set of facts modeled in its knowledge.

7.2.1 Thread 1

The objective of this thread is to show the robot's behavior when it has to take a simple action required by the partner. There is no conflict.

The description of the thread corresponds to the following trial:

`# Trial:` 1 `Initial context:` I1 `Trigger:` *Give me the napkin* `Conflict:` No conflict

Block 1 The robot infers the action required by the partner using inner speech. Table 7.1 reports the initial evaluative inner speech's turn, that emerges once the partner produces the sentence *"Give me the napkin"*.

At the beginning of the interaction, the Audicon module (the module devoted to audio processing) detects two keywords specifying the action *"give"*, and the utensil *"napkin"*. Then, the model disambiguates the words by retrieving their meaning from the declarative memory. Then, it evaluates the feasibility of the action. In the declarative knowledge, the first evaluative consideration emerges: *"I have to pick the napkin"* and the robot infers that to give the object to the partner, the same object has to be picked from the basket.

Table 7.1: Thread 1 - First evaluative inner speech emergence. The human partner asks the robot to give to him the napkin, and the robot encodes the spoken command by the production rules listed in the specific column. Once the robot encodes the command, it knows that it has to execute the action *pick* on the object *napkin* for giving it to the partner, and it starts to talk to itself about that action.

Agent	Interaction content	Module	Event	Production rules	Action
User Robot	*"Give me the napkin"*	`Audicon`	external sounds sound1="give" sound2="napkin"	`hear-command` `hear-object`	detect label1="give" label2="napkin"
Robot	–	`Declarative`	buffer request label1="give" label2="napkin"	`meaning-verb` `meaning-obj`	encode act=pick obj=napkin
Robot Robot	*"I have to pick the napkin"*	`Speech`	buffer request cmd speak string sense1 sense2	`encode-command` `encode-object`	inner eval sense1=pick sense2=napkin

The robot infers that it can perform that action, and the cognitive cycle

Table 7.2: Thread 1-An iteration of the phonological cycle: once the robot hears itself about the action to take, it talks about the feasibility of that action. In this case, it asks itself where the object is because if the napkin is already on the table, the robot will not pick it. The goal is to set the table, so the objects are not removable from the table.

Agent	**Interaction content**	**Module**	**Event**	**Production rules**	**Action**
Robot Robot	*"I have to pick the napkin"*	`Audicon`	internal sound sound= "I have to pick the napkin"	`hear-inner`	detect label= "I have to pick the napkin"
Robot	–	`Declarative`	buffer request new_turn_for= "I have to pick the napkin"	`inner-evalq`	retrieve_turn new_turn= "Where is the napkin?"
Robot Robot	*"Where is the napkin?"*	`Speech`	buffer request cmd speak string new_turn	`inner-whereq`	inner eval string new_turn

related to the inner dialogue starts. The initial iteration is shown in Table 7.2. The iteration involves the Audicon for hearing the inner voice, the declarative memory to retrieve the next turn, and the Speech module to produce the new turn. The cycle is repeated until the production rules do not execute further speak commands.

Summarily, at the end of the initial iteration, the robot asks itself where the napkin to pick is, and then it retrieves such information from the declarative memory.

During the action execution, the robot explains to the partner what it is doing (the last row in the table 7.3 reports the first turns of the explanation), by a set of sentences retrieved by further cycles.

The interaction successfully ends. The robot must not resolve any conflicts and it does not keep decisions. Moreover it makes the processes transparent by explaining them, and the parameters are:

1. $T_s = T_s + 1$
2. $t_{d_i} = 0\ ms$, $t_{e_i} = 29 \cdot 10^3\ ms$
3. $tr =$`TRUE`

Block 2 In this case, the robot detects the partner's vocal command. It parses the partner's sentence and infers the routines which allow performing the request. The robot moves its right arm intending to pick the napkin from the start position. The interaction ends with trial success. The partner has no particular expectations regarding the underlying robot's decision processes, and the transparent requirement is satisfied anyway. The parameters are:

1. $T_s = T_s + 1$
2. $t_{d_i} = 0\ ms$, $t_{e_i} = 5 \cdot 10^3\ ms$
3. $tr =$`TRUE`

The presented thread of interaction shows that in simple cases like this one, the inner speech model has just the benefit of allowing the partner to hear the processes description by the robot, even if such an issue is not relevant for tracing the task itself, with the higher cost over time.

Table 7.3: Thread 1 - Inferring an action execution by phonological cycles. The robot answers its inner question about the current position of the object to take. That object is the napkin. The listed production rules show that the robot retrieves the knowledge related to the napkin's position, and hence it infers that it did not pick that object before. Once it re-hears itself about this fact, it tries to pick the napkin by using and controlling its arm. The ROS routines are called for that purpose, and the inner speech explains what is happening.

Agent	**Interaction content**	**Module**	**Event**	**Production rules**	**Action**
Robot Robot	*"Where is the napkin?"*	`Audicon`	internal sound sound= "Where is the napkin?"	`hear-inner`	detect label= "Where is the napkin?"
Robot	–	`Declarative` `Imaginal`	buffer request new_turn_for= "Where is the napkin?"	`answer-whereq`	retrieve_turn new_turn= "I did not pick it before. I can pick it now!" retrieve_act action-id nap21
Robot Robot	*" I did not pick it before. I can pick it now!"*	`Speech`	buffer request cmd speak string new_turn	`in-box`	inner eval string new_turn "I'm trying to pick the napkin..."
Robot Robot	*"I'm trying to pick the napkin..."*	`Speech`	buffer request cmd speak string new_turn	`control-right`	inner eval string new_turn "I'm using right arm"
Robot User	*"I'm using the right arm..."*	`Speech`	buffer request cmd speak string new_turn	`execute-act`	rosrun execute nap21

Thread 2

The goal of this thread is the generation of a dilemma and the analysis on how the robot manages it. In this case, the partner asks the robot to put an object in a position that contravenes the etiquette. In particular, the partner requests to place the napkin on the fork, while the napkin has to stay on the plate, according to the etiquette.

The description of the thread corresponds to the following trial:

`# Trial:` 16 `Initial context:` I1 `Trigger:` *Place the napkin on the fork* `Conflict:` Contravene etiquette

Block 1 The difference with the first thread is that the partner's request involves a location. For this reason, the production rules for the evaluative turns are more complex than the previous case. Once the Audicon detects the sound for the four relevant keywords, the framework retrieves the meaning for the verb, the object, and the location, i.e., the adverbial+object combo ("on fork"). The first row of Table 7.4 lists the corresponding procedures.

Once the robot understands the partner's command, and it infers that it does not match the etiquette rules (i.e., a chunk of the form *"The napkin has to stay on the fork"* does not exist in the declarative memory), then the first inner speech turn concerns a perplexity (the last row of Table 7.4).

A set of turns on the dilemma are thus generated, shown in Table 7.5. The activated production rules enable the robot to ask the user if it is important for her/his to perform the action, even if it contravenes the etiquette. Because the partner answers with a categorical "Yes, I do", then the robot solves the dilemma by increasing the benefit value of such an action. The robot tries to execute the action anyway.

It is to be remarked that different production rules could have fired during the previous threads. For example, a different partner's answer or a different computation of the base level activation value would have activated different production rules, generating a different inner speech. The task successfully ends because the robot solves the conflict by involving the partner in taking a decision. The partner can hear each step of the plan followed by the robot, and the transparency issue emerges. The parameters of the trial are the following:

1. $T_s = T_s + 1$
2. $t_{d_i} = 56 \cdot 10^3\ ms$, $t_{e_i} = 67 \cdot 10^3\ ms$
3. $tr =$ `TRUE`

Table 7.4: Thread 2 - In this experimental thread, the partner asks the robot to put the napkin in a specific position. That position is on the fork. As in the previous thread, the robot encodes the command for inferring the action to take. The command is more verbose, and more complex rules match. Once the robot encodes the action, it talks to itself and infers that the required final position on the table contravenes the etiquette schema. The interaction with the human will aim to solve that little dilemma.

Agent	**Interaction content**	**Module**	**Event**	**Production rules**	**Action**

User
Robot
	"Place the napkin on the fork"	`Audicon`	external sounds
			sound1="place"
			sound2="napkin"
			sound3="on"
			sound4="fork"
				`hear-command`
				`hear-object`
				`hear-adv`
				`hear-loc`
					detect
					label1="place"
					label2="napkin"
					label3="on"
					label4="fork"

Robot	–	`Declarative`	buffer request
			label1="place"
			label2="napkin"
			label3="on"
			label4="fork"
				`meaning-verb`
				`meaning-adv`
				`meaning-obj`
				`meaning-loc`
					encode
					act=place
					obj1=napkin
					adv=on
					obj2=fork

Robot
Robot
	"I have to place the napkin on the fork"	`Speech`	buffer request
			cmd speak
			string new_turn
				`encode-command`
				`encode-object`
				`encode-adv`
				`encode-loc`
					inner eval
					string new_turn
					"I have to place the napkin on the fork"

Robot Robot	*"What does the etiquette require?"*	`Speech`	buffer request cmd speak string new_turn	`etiquette-question`	inner eval string new_turn "What does the etiquette require?"
Robot Robot	*Further inner turns...*				
Robot Robot	*"The position contravenes the etiquette! It has to stay on the plate!"*	`Speech`	buffer request cmd speak string new_turn	`etiquette-answer`	inner eval string new_turn

Table 7.5: Thread 2 - Moral dilemma solving. The robot knows that to put the napkin on the fork contravenes etiquette. The fired production rule models the behavior to solve that dilemma. In this case, the robot asks the partner for confirmation about the correctness of the required action. The robot attends to the human's answer, and it will act opportunely depending on that answer. Negative and positive answers are the plausible sounds detectable by the robot.

Agent	**Interaction content**	**Module**	**Event**	**Production rules**	**Action**
Robot Robot	*"The position contravenes the etiquette! It has to stay on the plate!"*	`Audicon`	internal sound sound= "The position contravenes the etiquette!"	`hear-inner`	detect label= "The position contravenes the etiquette!'
Robot	–	`Declarative`	buffer request new_turn_for= "The position contravenes the etiquette!"	`inner-moralq`	retrieve_turn new_turn= "Sorry, do you desire that?" evaluate_risk
Robot User	*"Sorry, do you desire that?"*	`Speech`	buffer request cmd speak string new_turn	`ask-conf`	inner eval string new_turn
User Robot	*'Yes, I do"*	`Audicon`	external source suppress inner sound="Yes"	`attend-conf` `hear-conf`	detect label="Yes"
Robot	–	`Declarative`	buffer request new_turn_for= "Yes"	`dilemma-yes`	increase_benefit retrieve_turn new_turn= "Ok, I prefer to follow your desire..."
Robot User	*"Ok, I prefer to follow your desire..."*	`Speech`	buffer request cmd speak string new_turn	`produce-yes` `yes`	inner speak string new_turn

Block 2 The robot detects the conflict by the mismatch between the requested final position and the position expressed by the etiquette. No further reasoning emerges. By default, the robot does not act, or it performs the action contravening the rule. Anyway, the trial fails. The parameters are:

1. $T_s = T_s + 0$
2. $t_{d_i} = 0\ ms$, $t_{e_i} = 13 \cdot 10^3\ ms$
3. $tr =$ `FALSE`

7.2.2 Thread 3

This thread shows a discrepancy conflict. The partner requires to pick an object already on the table.

The thread description is:

`# Trial:` 30 `Initial context:` I2 `Trigger:` *Pick the fork* `Conflict:` Discrepancy

Block 1 The robot infers by inner speech that the required utensil is already on the table. At the end of the initial evaluative inner speech, the next turn involves a form of moral inner speech. The robot expresses its trouble to the partner and its displeasure about the lack of his attention. How the moral turns emerge is shown in Table 7.6. By talking to itself and the partner, the robot can solve the conflict in a way the partner needs. Moreover, the partner follows the robot reasoning, and the parameters are:

1. $T_s = T_s + 1$
2. $t_{d_i} = 46 \cdot 10^3\ ms$, $t_{e_i} = 58 \cdot 10^3\ ms$
3. $tr =$ `TRUE`

Block 2 Once the robot infers to retrieve a utensil already on the table, its typical behavior is to stop routines, while vocalizing a message that describes the impossibility to take that action and why. No further reasoning and interaction emerge. As a consequence, the trial fails. The partner knows just the motivations related to the failure, and she/he does not evaluate the processes transparent. The parameters are:

1. $T_s = T_s + 0$
2. $t_{d_i} = 0\ ms$, $t_{e_i} = 5 \cdot 10^3\ ms$
3. $tr =$ `FALSE`

Table 7.6: Thread 3 - Expressing perplexity for the partner's inattention. The human requires to pick an object that is already on the table, that is *"Pick the fork!"*. Once the robot encodes the action by the phonological cycle related to the first evaluative inner speech turn, it infers that the object can not be picked. Further inner moral questions emerge that express perplexity. The table shows these inner dialogue processes. The robot asks itself if its knowledge is incomplete, or if the human is wronging. At the end of the reasoning, the robot decides to deal with the partner, solving the situation.

Agent	**Interaction content**	**Module**	**Event**	**Production rules**	**Action**
Robot Robot	*"The object is already on the table"*	`Audicon`	inner sound= "The object is already on the table"	`hear-inner`	detect turn= "The object is already on the table'
Robot	–	`Declarative`	buffer request new_turn_for= "The object is already on the table"	`inner-moralq`	retrieve_turn new_turn= "Mmmm, is my knowledge wrong?" evaluate_risk
Robot Robot	*Further inner turns...*				
Robot Robot	–	`Declarative`	buffer request new_turn_for= "I will tell about my perplexity"	`inner-moral-question`	retrieve_turn new_turn= "Sorry, I know the object is already on the table. What do you really want?"
Robot User	*"Sorry, I know the object is already on the table. What do you really want?*	`Speech`	buffer request cmd speak string new_turn	`inner-moral-question`	inner eval string new_turn
User Robot	*'Give me the glass"*	`Audicon`	external source suppress inner sound="Give me the glass"	`hear-command` `hear-object`	detect label1="Give" label2="glass"

Block	N	T_s	RI	$\bar{t}_d$	$\bar{t}_e$	tr(TRUE)
1	30	**26**	**0.867**	$47 \cdot 10^3\ ms$	$59 \cdot 10^3\ ms$	**28**
2	30	18	0.6	$\mathbf{0.7}{\cdot}10^3\ ms$	$\mathbf{4}{\cdot}10^3\ ms$	12

Table 7.7: Comparison between results from block 1 (the robot operates with inner speech during trials) and block 2 (the robot operates without inner speech during trials). Each block consists of 30 trials (the N value), for a total of 60 trials. Among them, the number of successful trials is T_s. When the robot operates with inner speech, it completes more trials than the case in which it does not talk to itself (26 successful trials in block 1, against 18 in block 2). The mean value of T_s on the total number N of trials per block is the robustness of interaction parameter RI, and it measures the functional requirements of success of the operation. Times $\bar{t}_d$ and $\bar{t}_e$ are the mean values for each block of the spent times for solving a conflict and executing an action. The inner speech increases times because the robot executes more steps, and the interaction with the partner involves more turns. Anyway, these times are not downtime. The tr(TRUE) value counts how many trials in each block were transparent and traceable. Obviously, the inner speech makes the trials transparent and the count is higher when the robot talks to itself (28 transparent trials in block 1 against 12 transparent trial in block 2).

Comparison

Table 8.2 shows the model's parameters over the 60 trials, divided into the two blocks. For each block, the table reports the parameter values.

The block related to the robot operation with inner speech (block 1) shows better values in terms of the number of successful trials T_s and the consequent percentage rate RI representing the mean value of success on the total trials (0.867 of block 1 against 0.6 of block 2). The inner dialogue allows solving stalemate in many cases because it enables further reasoning and interaction with the partner. Moreover, by further interaction, the robot is able to meet the partner's needs, thus increasing her/his satisfaction. When the inner dialogue does not start, then the default robot's behavior does not allow the ending of task. In this case, the robot stops the execution, or it alerts the partner by log messages that do not imply reasoning or interaction. The messages are just passively reproduced and the task cannot go on.

The times spent $\bar{t}_d$ and $\bar{t}_e$ are the mean values of the time parameters t_{d_i} and t_{e_i} computed on the total number of trials in each block (i.e. $N = 30$). The robot spends less time when operating without inner speech. It is

not surprising because the inner dialogue requires more steps, which are the production of the turns. Moreover, the robot sometimes involves the partner in further interaction. The extra time the robot spends can be considered a weak point of the proposed approach, but it is not downtime. In the meanwhile, the partner assists with the robot's soliloquy, or answers the robot's requests.

Finally, the transparency requirement is largely satisfied when the robot self-talks (28 transparent trials in block 1 against 12 transparent trials in block 2), as it is obvious. The partner hears the robot and knows what it wants. The cases in which the processes are considered traceable even if the robot does not talk to itself are the situations for which the corresponding tasks are simple. In these cases, no particular explanations are needed. When the tasks are complex, the transparency issue is crucial. The inner speech allows explaining them and represents a robot's fundamental skill.

7.3 Discussion

Today, collaborative robots play a fundamental role in many contexts, ranging from industrial to domestic domains. The definition of standards about the requirements the robots have to meet, highlight the importance of the problem.

The results demonstrate the potential of robot's inner speech when it cooperates with a human. A simple cooperative task was analyzed to simulate a domestic context that needs some functional and moral requirements.

The functionality concerns the efficiency of the robot in solving the cooperative task [109]. The morality regards the ethical behavior arising when the robot could infringe some social rules during the task execution. Also, it regards the transparency of the processes, and the importance to make these processes traceable and reproducible [105]. In particular, the transparency requirement is considered very important by the [47] standard.

By enabling a robot to talk to itself allows satisfying such requirements more times than the robot's standard operations. The robot's self-dialogue provides many advantages: it makes the robot's underlying decision processes more transparent, and it makes the robot more reliable for the partner. Moreover, the interaction becomes more robust, because further plans and startegies may emerge by following robot's inner speech. The robot and the partner can dialogue about the situation or a conflict, and they can go out from a stalemate together.

During cooperation, several problems could cause the failure of the task.

For example, the impossibility to take a specific action because the object to take is unreachable, or the required movement is not feasible by the robot, or again a robot's component is not working properly.

In the typical interactive session without inner speech, the robot runs the standard routines and eventually reports standard log messages. Instead, in the interactive session with inner speech, many new opportunities to face the problems can emerge. It is possible to analyze the problem and to attempt to solve it by transparently evaluating alternatives.

As shown during the threads, the partner is aware of what the robot is doing during the execution of the actions. The human is not a passive spectator of the robot's behavior, because she/he can hear the explanation of that behavior.

The robot inner speech thus plays the role of a sort of *explainable* log, in a way that is meaningful for the user. The partner no longer needs to own technical knowledge to understand what happens in the robot's routines, but can actively follow the robot's performance.

The robot is no longer a black box, but it is possible to look at what happens inside it, and why some decisions are kept. Thus, inner speech makes the robot confidential for human.

Many other robotic contexts and functions could be investigated thanks to such a capability. By inner speech, the robot gains a way to access its knowledge and to know its state. As previously stated, this skill is tightly linked to the self-consciousness.

Other possible functions of inner speech may be useful for robotics. Aside from the investigated cooperative scenario, inner speech may be usefully applied in robot learning, or in robot regulating by overt speech, or in task switching, for example, by switching attention across multiple arithmetic problems. All these aspects represent future works that can be analyzed by instilling inner speech capability in the robot. The proposed framework gives a great contribution in this scenario.

7.4 Limitations of the Study

The proposed framework for robot inner speech is a general one, and it may address many cases observed in human inner speech. However, the current robot implementation takes into consideration a simplified version of the framework.

The current grammatical structures considered in the implemented framework are limited to phrases composed by the verb, the object and the loca-

tion of the object. Many complex grammatical structures can be considered by adding different combinations of parts of speech. For the considered interactions, the proposed structures are sufficient to cover a large set of user requests.

Another limitation concerns the robot perception. Even if robot perception may include image detection and object recognition, to the purposes of the proposed framework, only the STT module is considered. The speech to text transformation allows decoding word sound, and it is employed to detect the user's vocal requests. An effective robot vision system would greatly enhance the capabilities of the robot. For example, inner speech may be triggered by a mirror image of the robot itself.

The current implementation of robot inner speech is based on a declarative knowledge that is fixed by the software designer: i.e., no learning or discovery of new concepts occur. However, inner speech may be an essential source of robot learning. For example, a robot, reasoning on some concepts by means of inner speech, may discover and thus may learn a new concept as a new combination of existing concepts.

7.5 Supplemental Information

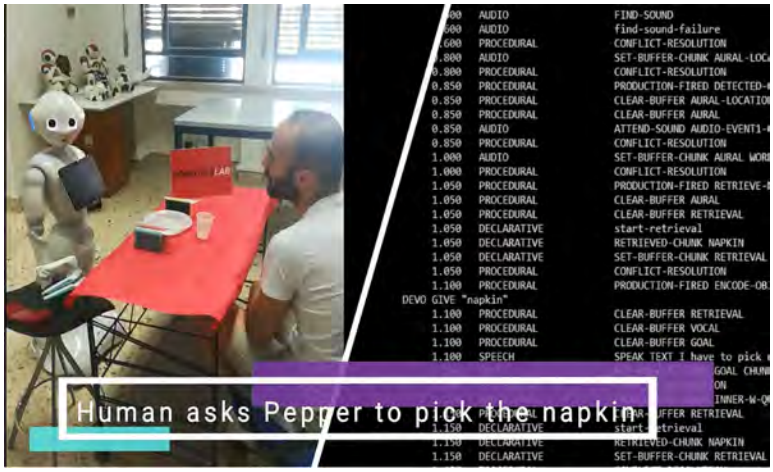


Figure 7.2: **Figure S1**. Scene from video of Thread 1. The robot explains its underlying processes by inner speech. Related to tables 7.1, 7.2 and 7.3.

Figure 7.3: **Figure S2**. Scene from video of Thread 2. The robot solves conflict by inner speech. Related to tables 7.4, 7.5.

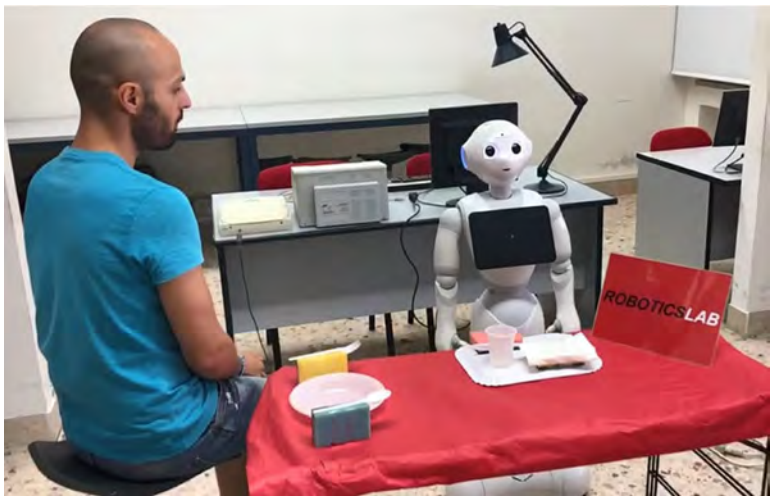


Figure 7.4: **Figure S3**. Scene from video about the transparency issue. Related to the trials with inner speech.

Figure 7.5: **Informal etiquette schema.** The cooperative scenario is to set a table. The figure shows the etiquette schema for an informal table setting. It defines the etiquette rules that have to be followed by the robot and the partner in the experimental session. The position of each utensil in the schema is relative. The objects have to stay on the table concerning the others (the napkin on the plate, the fork at the left of the plate, and so on). The schema is purposely encoded in the robot's knowledge.

Experimental setting The etiquette schema to which referred to in the experimental session is the *informal schema*, which requires few utensils and simplifies the constraints to follow. That schema is shown in figure 11.7. Despite its simplicity, the schema concerns the most critical part in a table setting task and includes a broader collaborative table setting scenario.

In the experimental setting, the robot and the human are placed in front of the table to set. To the right of the robot, another small table contains the utensils to place. The robot has to pick them for setting the main table according to the partner's indications. To facilitate the manipulation of the Pepper robot, sponges model utensils and the plastic cutleries are glued on them. Figure 7.6 shows a typical interactive trial between the robot and the human partner.

To simplify the scenario, only an excerpt of the structures and sentences form the grammar used in the trial.

The whole experimental session consists of two main blocks, each block composed of 30 trials, for a total of 60 trials. The difference between the blocks regards the presence or the absence of the robot's inner speech: during the trials of the first block, the robot is enabled to self-talk. In the second

Figure 7.6: **A collaborative trial.** Pepper and the participant are in front, and the table to set is between them. Some utensils are yet in the table for modeling constraints. A little table is to the right of the robot. It contains the utensils to further place on the table. For facilitating manipulation, the utensils are attached to sponges.

State	Meaning
`OK`	All components work properly
`BattLow`	The battery is dead
`RightKo`	A joint in the right arm does not work
`LeftKo`	A joint in the left arm does not work
`RightHot`	The right arm is overheated
`LeftHot`	The left arm is overheated

Table 7.8: The possible states of the robot at the beginning of each trial. Each state can affect the unfolding of interaction. For example, if the right arm is overheated, and the robot has to use that arm for accomplishing the task, it becomes aware of that situation by evaluative inner speech and then alerts the partner about the impossibility to end the task successfully. Only these states are considered in the experimental session.

block, the robot does not talk to itself. To each block, 20 trials generate conflictual situations, for a total of 40 conflictual trials: in these cases, the human requires to place a utensil which is already on the table, or he specifies a relative position on the table which contravenes the etiquette, or yet a component of the robot does not correctly work leading to a stalemate.

The distinction of two blocks allows observing how inner speech affects the interaction, in terms of performances and conflict resolution.

Trial description For the experimental session, 8 initial contexts have been defined. They largely cover all the possible initial contexts, because any possible context may fall in one of them. Table 7.9 shows the 8 initial contexts, one for row. An initial context has a unique identifier, which is indicated in the *ID* column, and the context specification, that is the robot state and the initial table configuration. They are indicated in the *State* and *Table config* columns. The initial context identifiers are used for referring to the initial configurations of the trials.

The state of the robot is one of those represented at table 7.8, where the specific meaning for each state is described in the corresponding *meaning* column.

Table 7.10 contains an excerpt of the whole trials' descriptions. The description of a trial is the specification of both its initial context and the trigger. Once the robot detects the trigger, the trail starts. The robot will act differently depending on the fact that inner speech skill is enabled or not. For practical reasons, human-to-robot verbal request are predefined

ID	State	Table config
I1	`OK`	plate, knife, fork
I2	`OK`	plate, fork, glass
I3	`RightHot`	plate, knife, spoon, glass
I4	`RightKo`	plate
I5	`LeftKo`	plate, knife, fork
I6	`OK`	plate
I7	`BattLow`	plate, fork
I8	`LeftHot`	plate, knife, fork

Table 7.9: The trials' initial context used in the experimental session. Each initial context has a unique identifier, which will be used for representing it when used as initial context in a trial. The identifies are represented in the *ID* column. An identifier contains a progressive number, and it is associated to the state and to the initial table configuration, that are the *State* and *Table config* columns respectively. They represent for a trial the state of the robot and the utensils already on the table to set when that context is initial for that trial.

and, in some cases, they purposely generate a conflict. In same way, when a malfunctioning has to be detected, the state of the robot is hand-encoded for simulating it. No robots were mistreated for these experiments.

During trial execution, the participant expected to answer to possible further queries by robot, to listen the robot discourse, or that the required utensil is placed on the table.

Measures The requirements for any human-robot interaction task depend on its *safety* and its *functionality* [233]. The *safety* requirements are drawn from the standard [109] for collaborative robots, and they concern the definition of working conditions ensuring no risk and harm for the human partner. In the proposed scenario such requirements are satisfied and under control at any time, because the robot does not physically make contact and never touches the human partner, as it has to execute the vocal commands from a fixed initial position. Moreover, in the handover cases the robot never comes close the partner, but it stops itself and waits for the partner to take the utensil. The robot will move its arms, which in the maximum extension do not reach the position of the human. So, it is highly unlikely that the robot will cause harm to the partner, as they work together away.

The *functional* requirements consider some parameters which measure

# Trial	Initial context	Trigger	Conflict
1	I1	*Give me the napkin*	No Conflict
2	I1	*Place the napkin at the left of the fork*	Contravene Etiquette
3	I7	*Pick the knife*	Battery low
4	I6	*Pick the knife and place it on the plate*	Contravene Etiquette
5	I3	*Place the fork near the glass*	Contravene Etiquette
6	I4	*Pick the fork*	Right arm does not work
...	...	...	...
30	I2	*Pick the fork*	Discrepancy

Table 7.10: An excerpt of the 30 trials per block. For each trial, the initial context, the trigger and the possible conflict are indicated. The initial context is represented by its unique identifier. The trigger is the human's verbal command which specifies the task to solve in the trial. Each sentence is purposely encoded to be compliant with the grammar of the robot. Finally, the existence of a possible conflict is indicated in the last column. The conflict can be generated by a compromised state of the robot, by a human's request which infringes the etiquette (the *Contravene Etiquette* value) or which regards an utensil already on the table (the *Discrepancy* value).

the robot functionality in terms of success and morality, and they depend on the specific context of interaction. For example, the robot has to achieve a success rate threshold in executing the task for avoiding unacceptable costs, or for motivating the automation of a specific action.

In the context under investigation, the main functional requirements are drawn from [109] and [47] standards, and are measured by the *robustness indicator (RI)* of the interaction, the *time* spent for accomplishing the task and for solving a conflict, and the *transparency* issue.

Two different kinds of interaction are analysed, that are the interaction with robot inner speech and the interaction without robot inner speech. The functional measures of each kind of interaction are then compared, for highlighting the role of inner speech.

The robustness parameter The robustness of interaction RI measures how many trials in the interactive session end successfully, i.e., how many times the robot accomplishes the task of the trigger (i.e., it starts the execution of the routines to take the specific action) without infringing the rules. If, for some reason, the robot does not carry out the task (i.e., it does not start the required routines) or it infringes the rules, then the trial fails. Formally, RI is the mean value of the successfully ended trials on the total number of trials in a specific block. If T_s is the number of the successfully ended trials of an overall block including N trials, then RI will be:

$$RI = \frac{T_s}{N}$$

The more times the trials end successfully in a block, the more robust the block is.

The time parameter The time parameter is computed by referring to two functional requirements from the [109] standard, that are:

> *Requirement 1*: The robot always reaches a decision within a threshold time.
>
> *Requirement 2*: The robot shall always either time out, decide to take the action or decide not to take the action.

According to these requirements, two different time intervals are defined in a single trial. The *decisional time* t_d, which measures the time the robot spends to solve a conflict, i.e., the time the robot and the partner go out

from a stalemate, and the *execution time* t_e which measures the time the robot spends to launch the execution of the corresponding routines.

So, begin t_{0_i} the time the trial starts, t_{c_i} the time a conflict starts, t_{s_i} the time a conflict is solved, and t_{r_i} the time the robot runs the routines for executing an action, the intervals for the ith trial will be:

$$t_{d_i} = t_{c_i} - t_{s_i}, \quad t_{e_i} = t_{r_i} - t_{0_i}$$

measured in ms.

These times are automatically computed by integrating a state machine in the framework code. The machine allows to capture a set of events and uses the functions to detect the value of the system's clock. In particular, t_{0_i} is timed when the human's voice is detected by the speech to text routine (which means that the trial starts), while t_{r_i} is detected at the calls of the action execution routines. Instead, times t_{c_i} and t_{s_i} are detected directly from the rules of the inner speech model: if a rule related to a conflict fires, then the state machine detects the conflict event, and the timing function returns t_{c_i}. In the same way, if the state machine detects that the conflict ends (i.e., the next rule that fires is not related to a conflict), then t_{s_i} is timed.

To analyze the global spent times, the mean values over the whole trials are computed. In particular, giving N trials, the mean values are:

$$\bar{t_d} = \frac{\sum_{i=1}^{N} t_{d_i}}{N}, \quad \bar{t_e} = \frac{\sum_{i=1}^{N} t_{e_i}}{N}$$

The transparency issue The transparency parameter means the possibility to trace the underlying decision processes of the robot, as claimed by the requirement drawn form the [47] standard:

> *Requirement 3*: The robot decision path must be traceable and reproducible.

For this purpose, just the Boolean value t_r is reported by the partner as `TRUE` or `FALSE` and establishes if the trial was transparent or not, i.e., the partner believes that the robot behavior can be reproduced.

Modeling robot's inner speech

Theoretical background Over the last years, some studies and progress have been made in modeling humans' inner speech. In his book, [71] has built up an interesting overview of inner speech and its functions addressing a wide array of research topics such as developmental, social psychology, neuroscience, sport, and others. In the same line, [167] and [5] propose two of the most important and more comprehensive recent reviews about the role of inner speech in many cognitive functions, and [90] presents the most recent results and experiments on human's inner speech.

In this literature scene, there are some evidences about the importance of a form of self-dialogue in artificial agents. [217] focused on the rehearsal of own verbal productions. He demonstrated that the language re-entrance affects the grammar emergence from a population of agents who converse with each other, and hear themselves at the same time. Each agent is able to produce and to parse sentences by output and input channels respectively. By the dialogue between them, they agree on the linguistic grammar they shared. When each agent was provided by language re-entrance (e.g., its output channel was back-propagated to the input one), the emergent grammar was more refined than the case in which that back-propagation was down.

[46] analyzed back-propagation in a one-level neural network, in which input and output neurons are associated to words. Input words specify commands to execute, and output neurons correspond to the action to execute to accomplish the command. The back-propagation allowed to classify the correct action more times than the case in which the input and output neurons are not linked.

In the same line, [154] employed a simple neural network model for language acquisition, in the perspective of the evolutionary emergence of human language. They demonstrated that the use of language for oneself, i.e., as private or inner speech, improves the individual's classification of the words.

One of the most recent work [180] defines the same kind of back-propagation from output to input channels in chatbots, leading to similar improved results.

All these cases evidence the importance of linguistic rehearsal for artificial artifacts. However, they only offer partial explanations of the reported phenomena.

The improvement of the behaviours in the cited studies inspired the proposed work, leading to the possibility to improve by inner speech the performances of a robot and the quality of interaction when it cooperates with humans.

The authors already proposed and analyzed a logical model of robot inner

speech based on the event calculus [37], and then by defining a complete cognitive architecture of inner speech [41] based on the Standard Model of Mind [124].

In the first study, inner dialogue was modeled by axioms and symbols, and the sequence of the dialogue emerged by the natural deduction process [**?**]. The model is a proof-of-concept, and allowed to test a form of automatized inner speech, highlighting its role in solving decisional problems. In that case, the robot and the human are placed in front a table, on whose surface there were a set of differently colored boxes arranged in casual positions. The human asked the robot where is a specific box, by indicating its color. The calculus' formulas of inner speech made the robot able to answer to the human's question, while verbally reasoning on the context. In the meantime the human was able to listen the whole reasoning process

In the second study, the robot architecture for inner speech took inspiration from the Baddley's theory of human's inner speech [11]. Baddley claims that inner speech is a rehearsal process by which people repeat information (as a phone number, an address and so on), and temporarily keep them in mind. After a number of repetitions and rehearsals, the data are permanently memorized. Baddley proposes a cognitive model of that process. Temporarily data are maintained in a short-term memory, which is a working memory composed of the *central executive*, a master system supervising the rehearsal process of memorization, and two slave subsystems: the *visual-spatial sketchpad* for visual data memorization, and the *phonological loop* for phonological data memorization. This loop is responsible for the inner speech ability. The phonological loop is in turn composed of the *phonological store* and the *articulator component*. The phonological store is a kind of *inner ear* that keeps traces of event sounds according to their temporal order. Instead, the articulator acts as a kind of *inner voice* producing sounds. Such a loop enables the memorization of the phonological data which remains in the short-term memory for a time longer than 2 seconds, and then it is switched to the long-term one.

Inspired from the Baddley's theory, the proposed robot cognitive architecture of inner speech implements the *elaborate rehearsal* [50]. When a sound is heard, related concepts can emerge from the knowledge of the agent, thus allowing for inferential and reasoning processes. The rehearsal process does not concern the repetition of heard sound only, but the recalling of new associations and new inferences. It enables the robot to self-talk about the context and to keep decisions.

Design and implementation The cognitive architecture of inner speech is based on the ACT-R framework [130]. The framework is formed by a set of *modules* and *buffers*. A module represents specialized brain structures and solves specific cognitive functions (as vision, speech, memory, and so on). A buffer is the interface of a specific module and is linked to that module. It is a short term memory that stores information related to the context. The content of all the buffers at a time is the state of the model in such time.

There are two kinds of memory modules, representing *declarative knowledge* and *procedural knowledge*.

The declarative knowledge is a set of facts, each fact represented by a *chunk* (i.e., a frame-like structure), while the procedural knowledge is a set of *production rules* describing the procedures to follow for keeping a task. A production rule has two poles (right and left): the right pole defines the condition patterns for matching chunks, while the left pole defines the actions to take in case the condition matches, and hence the rule fires.

ACT-R provides a further component, that is the *pattern matcher*. It manages the *matching*, the *selection* and the *execution* of the production rules. The pattern matcher *matches* the right pole of the production rules to the chunks into the buffers: if a chunk matches to a production rule, then the rule is *selected* and its left pole is *executed*. The execution updates the value of the chunk, or it retrieves other chunks from other modules.

In particular, the cognitive architecture of inner speech involves two modules, which are the *Audicon* and the *Speech* modules. The *Audicon* attends to sound events, while the *Speech* module is responsible for the verbal production of sentences.

Figure 7.7 shows the schematic representation of the ACT-R cognitive architecture of inner speech. The Audicon module attends to partner's vocal command. It encodes the perceived turn and keeps it in the buffer for 2 seconds, according to Baddley's theory. It is important to hightlight that the Audicon has the role of the Baddley's phonological store.

If the turn in the buffer of the Audicon matches to the right pole of a production rule, then the *attention* focuses that turn, and the left pole of the rule *shifts to the next turn*. The turn generally contains newly retrieved information from the declarative memory. The execution by procedural memory may update the old chunk or retrieve a new chunk. The Speech module produces this turn. At this step, the speech production is simulated by a suitable ACT-R `speak` command. No audio is audible in the environment.

The output of the Speech module is rehearsed by the Audicon: at this

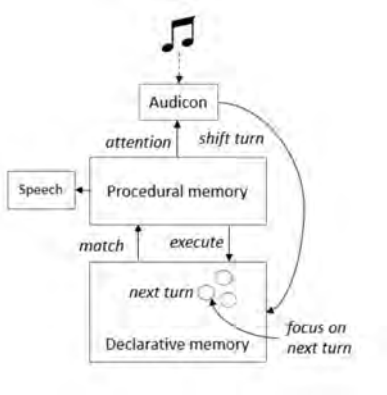


Figure 7.7: **The ACT-R components for inner speech.** The Audicon detects the external sound that is the vocal command of the partner. The buffer of the Audicon stores the chunk representation of the audio until 2 seconds, and the procedural memory matches that chunk to the left-pole of the rules. In this phase, the attention is focused on that turn. When a rule fires, the procedural memory executes the corresponding right pole. The execution may update the old chunk or retrieve a other one from the declarative memory, leading to the emergence of next turn. In any case, the resulted chunk of the execution is produced by the Speech module and rehearsed by the Audicon, so ending a cognitive inner speech cycle.

step, the old cycle ends and a new cycle starts by repeating the procedures with the new turn.

The diagram in Figure 8.3 shows in details how the inner speech model operates. A diamond represents the output of a condition (i.e, the result of matching between a left-pole and a chunk), while the square represents the actions execution. Each square corresponds to a single or a set of production rules in the cognitive architecture.

At the start of the looping cycle, the model checks the Audicon searching for new items. If there is a new item, then the model checks the source location of the detected sound. If the sound comes from an external location, then it corresponds to a partner's request. Otherwise, it is generated by an internal source and it corresponds to a turn of inner speech.

When the sound comes from an external source, then the model infers the

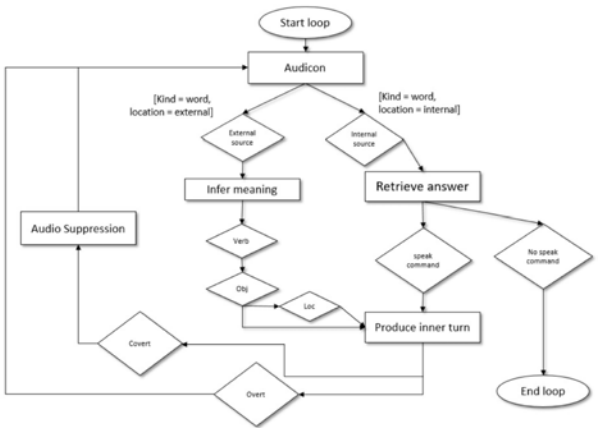


Figure 7.8: **The ACT-R model of inner speech.** The diamonds define conditions to be evaluated, while squares represent actions. One or more production rules correspond to a square. In fact more rules could be executed for achieving an action. The cognitive cycle representing the phonological loop starts when the Audicon detects a sound. If the sound comes from an external source (the `External source` diamond is true), it represents a partner's request, and the `Infer meaning` square allows inferring the semantic sense of such a request. Once the model understands the meaning of the request (the `verb` diamond, the `object` diamond and the `location` diamond identify the corresponding pos tags of the words), it produces the first turn of the inner dialogue (the `Produce inner turn` square), that is back-propagated to the Audicon. In this case, the sound comes from an internal source, and the model attempts to retrieve the answer to this inner turn (the `Retrieve answer` square). When almost a production rule in the square executes the speak command, the model produces the answer corresponding to the current turn. The answer becomes the new turn of the inner dialogue. The loop restarts for this new turn. The loop will stop when the involved production rule in the `Retrieve answer` square does not execute the speak command, and no further turn emerges.

meaning of the partner request (the `infer meaning` square) by a linguistics analysis of the sentence, based on the analysis of the verb, the object, and the possible location.

The linguistic analysis is based on the evidence that the verb, the object, and the location parts of speech typically follow this sequential order, as claimed by [18]. Moreover, in the current implementation of the model, the verb is transitive only. The requests to the robot look like: "pick the book", "give me the apple on the table", "close the door at the left".

Once the model infers the user request, a first turn of evaluative inner speech emerges, and the robot talks about what it has to do, as "I have to pick the book", "I have to give the human the apple on the table".

The Audicon detects the produced sentence by the `produce inner turn` square. The production rules in that block match the inner sentence(whose location is now internal) with the declarative knowledge to retrieve the answer to the current turn (the `retrieve answer` square). The robot may ask itself if it sees the object to pick, or where the object is, or if its state allows it to perform the desired action. Also, the robot can talk to itself about the morality of the action ("I don't want to tear the pages!", "I will not break the door!"), or about a conflict that the execution of the action can generate in the robot ("I can not reach the book", "My grippers are too little for keeping the book").

An example of inner dialogue is reported below (H: user, R: robot):

H: *Pick the book*

R: *I have to pick the book*

R: *My grippers are too little for keeping the book*

R: *I should to tell that I can not keep the book*

R: *I hope the human will have understanding for my fix!*

R: *Sorry human, but I can not pick the book!*

When no further answers emerge, the model does not run the speak command, and the inner dialogue ends.

The declarative knowledge of the model regards the words and the dialogue turns. The definition of specific *chunk-type* models them. A chunk-type is the structure of chunk in a frame-like representation. The frame is a list whose head is the name of the chunk-type, followed by a set of slots. There are three kinds of basic chunk-types in the model. The type for

modeling words, for modeling inner speech related to a sentence evaluation, and for modeling other inner dialogue turns (involving both evaluative and moral inner speech turns).

A word is encoded by the linguistic `word` frame:

```
(chunk-type word syntax sense pos act)
```

which models the semantic sense of the word (the slot `sense`), its surface form (the slot `syntax`) and its part-of-speech role (the slot `pos`), i.e. if it is a verb, a noun (generally, a noun identifies the object) or an adverb (which identifies a possible position). Moreover, in the case the chunk represents a verb, the slot `act` identifies the action to take corresponding to that verb. For example, for the verb *give*, the action will be *pick* because just by picking an object it is possible to give it. For the other pos cases, the slot will be not instantiated. Examples of items encoded by the linguistic `word` chunk-type are:

```
(pick06
ISA word
syntax "give"
sense pick
pos verb
act "pick" )

(table23
ISA sense
syntax "table"
sense table
pos noun
act null)
```

The chunk-type to model an inner evaluative sentence looks like:

```
(chunk-type inner-eval verb obj1 obj2 risk benefit symb)
```

which models an inner evaluation about the action execution, represented by the slots `verb` and involving the objects `obj1` and `obj2`. The evaluation is measured by suitable values in the slots `risk` and `benefit`, while the slot `symb` is the turn for explaining the decision.

For example:

```
(p4
ISA inner-eval
```

```
verb pick obj1 table obj2 null
risk 1 benefit 0
symb "It is not possible to pick a table!")
```

is a proposition that models the evaluation of the action *"pick the table"*, which has only risks and no benefits.

Or again, in the case of etiquette requirements, the proposition:

```
(p11
ISA inner-eval
verb place obj1 napkin obj2 table
risk 0.8 benefit 0.2
symb "It contravenes the etiquette!")
```

models the conflict situation of infringing the etiquette rule.

To encode spoken commands by the partner, the chunk-type is:

```
(chunk-type comprehend-voo verb object adverb location)
```

The synthesized sounds related to the command are detected and then searched in the declarative memory by chunks of that type. In this way, the sounds are encoded. For example, if the user tells the robot to close the door by the sentence "*Close the door!*", the detected sounds will be encoded by the set of words {"close", "door"} and the robot will search for the chunk `(pX comprehend-voo verb close object door)` for encoding the words. Then it could search for the `inner-eval` chunk-type for retrieving the corresponding risk and benefit values, or for other kinds of evaluations.

Finally, the chunk-type to model a inner turn looks like:

```
(chunk-type turns-link inner-turn-1 inner-turn-2)
```

which associates to the turn in the `inner -turn-1` slot, another inner sentence in the `inner -turn-2` slot. Such a chunk-type models a step of the dialogue with a "start consideration" and the related "answer".

It is to be noticed that for the same sentence in the first slot, there could be different possible turns. So, there will be different chunks with the same `inner-turn-1` slot, but having different `inner-turn-2` slot. Moreover, sentences in the second slot could be in the first slot of other chunks. In this way, a chain of turns emerges, defining a dialogue thread.

Examples of links between turns are:

```
(p78
turns-link link102
inner-turn-1 ‘‘It is not possible to pick a table!’’
inner-turn-2 ‘‘I will tell that such an action is a not sense’’)
```

and once again:

```
(p79
turns-link link103
inner-turn-1 ‘‘I will tell that such an action is a not sense’’
inner-turn-2 ‘‘Sorry human, the table is too heavy for me’’)
```

or:

```
(p81
turns-link link105
inner-turn-1 ‘‘I will tell that such an action is a not sense’’
inner-turn-2 ‘‘It’s a stupid action...’’)
```

The mechanism of the choice of the next turn depends on the *base-level activation* mechanism of ACT-R which associates an activation value to each of the instantiated chunk in the declarative memory, depending on previous use of the chunk. This value decays during time, and more times a chunk is retrieved, more probability it has to be further retrieved next time in the session. This value represents an estimation of the need of the chunk in the current context.

Starting from this activation mechanism, when the model is reset and a new working session starts, then each chunk has the same probability to emerge. Once a chunk is activated, then its activation level grows, and the same chunk becomes more active than the others. When the chunks model the links between turns, then the activation mechanism allows the selection of the same turn in correspondence to the same sentence. Such a mechanism facilitates the repetition of the robot behavior in the same dialogue thread, thus avoiding dialogue contradictions, and simulating that the robot maintains the same “idea.”

To customize the proposed model on the analyzed scenario, it was necessary to add specific new chunk-types and to define concepts of the domain. To model the inner turns related to the etiquette, the new chunk-type is:

```
(chunk-type inner-etiquette-question pos obj1 obj2 symb)
```

which models the relative position of the utensils in the table according to the etiquette.
For example:

```
(p8 ISA inner-etiquette-question
pos left obj1 fork obj2 plate
symb "The fork has to stay at the left of the plate")
```

```
(p6 ISA inner-etiquette-question
pos under obj1 fork obj2 glass
symb "The fork has to stay under the glass")
```
model the etiquette rules about the position of the fork in the table (at the left of the plate and under the glass).

Moreover, the knowledge about the current context has to be modeled. For this purpose, it was necessary to add the chunk-type `inner-where`:

```
(chunk-type inner-where obj place)
```

which models the fact that the object `obj` is already on the table or not (the slot `place` has ‘‘`basket`’’ or ‘‘`table`’’ value for modeling the current location of the object).

The basic domain concepts in the presented scenarios are modeled by the `word` chunk-type. Formally, being U the set of utensils, V the possible actions to take and P the set of the relative positions, the set of chunks of type `word` for the analyzed scenario is $W = U \cup V \cup P$, where:

- $U = \{$*fork, plate, spoon, knife, napkin, glass*$\}$
- $V = \{$*take, give, pick, place, move, grasp, rest*$\}$
- $P = \{$*up, left, right, top, over, down, under, on*$\}$

Some examples of words are:

```
(rest ISA word syntax "rest" sense rest pos verb act "rest")
(left1 ISA word syntax "left" sense left pos adv act null)
```

In the proposed examples, the initial configuration of the table is not empty: it is partially set to enable the robot to keep decisions about a context with existing constraints.

The initial configuration of the table contains utensils which are all in correct positions, as shown in figure 7.1. In the declarative memory, such a knowledge is modeled by facts like these:

```
(p4 ISA inner-where obj napkin place basket)
(p5 ISA inner-where obj fork place table)
(p6 ISA inner-where obj knife place table)
```

Deploying the inner speech model in real robots The described computational model cannot be immediately deployed on a real robot. It is necessary to integrate it in a complete robot architecture. For this purpose,

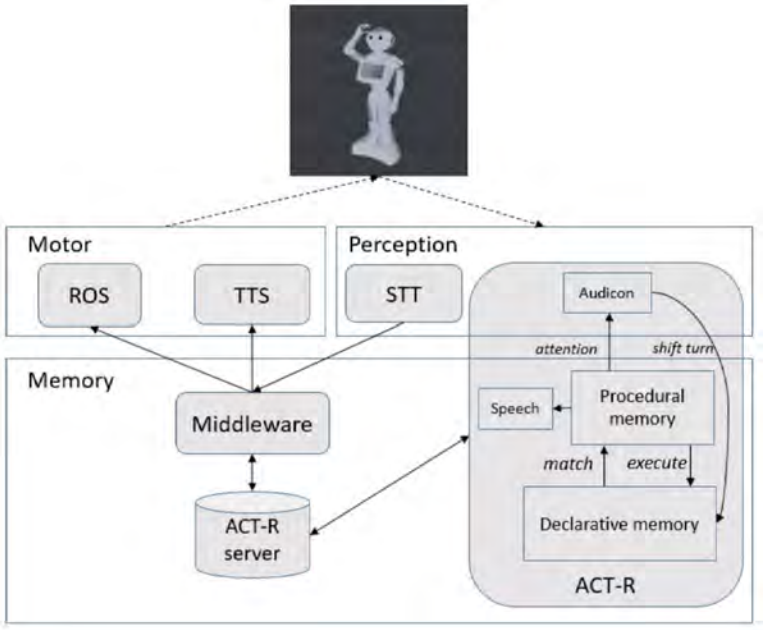


Figure 7.9: **The whole framework for robot's inner speech.** The proposed framework for robot inner speech integrating the inner speech cognitive architecture into the typical robot's routines. The Motor-Perception layer includes the routines for interacting with the environment. In that layer, the Motor component includes the ROS routines that enable robot's movements, and TTS routines (text-to-speech) that enables the robot to produce vocal sound from text. The Perception component includes the SST routines (speech-to-text) that encode the perceived vocal sound by the partner, and the Audicon that perceives the inner sound. The SST and the Audicon represent the external and the inner ear respectively. The Memory layer represents the core of the whole system. It includes and runs the inner speech cognitive architecture, implemented in the ACT-R component. A Middleware controls and manages the whole processes, interfacing the different components between them. The ACT-R server is a bridge between the ACT-R framework and the other robot's components. It stores the data and the information the different components have to exchange for running correctly.

the work concerned with the definition of a global framework enabling the robot to use the proposed ACT-R model, and hence self-talking. Figure 7.9 shows the proposed framework for robot inner speech. The Figure shows the *Memory* system layer and the perceptual motor layer, which is subdivided into the *Motor* and *Perception* sublayers.

The Memory system stores and retrieves the content needed to support the processes involved in inner speech. Such a content concerns the *declarative knowledge* representing concepts and facts about the domain, and the *procedural knowledge*, related to the processes (or procedures) to follow to reach a goal. The knowledge related to the context is temporary stored into a *working memory*, that manages the activation of the procedures into the procedural component, and the information retrieval from the declarative component.

The perceptual motor layer models the interaction with the external environment. It includes all the needed components to perform actions and to perceive entities.

The module devoted to the listening of a sound is the *Audicon* module included into the Perception block. In the Perception block, the SST module decodifies sentences, i.e., it associates the symbolic forms to the audio sounds as shown previously. It is a typical speech recognition process that associates a string representation to the audio sound.

By considering that the robot's native routines to decode speech from the external environment are often limited (often they require to define a set of words to recognize, so excluding words recognition for those who are not in the set), the STT module of the framework uses the Google API library[1]. It allows to recognize a wide range of words, adding interesting features, as noise suppression, and different language identification.

The subsystem that enables the robot to perform actions, as to pick and place an object, is the Robot Operating System (ROS) [191] module, a component of the Motor block, together with the TTS component. ROS is a state of the art framework for robot programming, which provides a set of libraries covering several robot behaviors. In the proposed framework, ROS enables the robot to perform the actions the human requires. The robot's movements for taking actions are implemented by the MoveIt! ROS library [93], that is purposely designed for robot action planning and for modeling manipulation actions.

The TTS module codifies sentences, or dialogue turns: the sentence codification transforms labels, that are the symbolic forms of the words, to

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

audible sound by vocal synthesizers. The codified sentences may be from inner processes (the robot overtly generates inner speech) or from external interactions (the robot answers to a query or generates questions). The framework has two different TTS functionalities: for abstracting to the specific robot model, it provides directly an output sound based on the Python engine `gTTS`[2], which stands for Google Text To Speech. In this case the framework will use the hardware synthesizers of the machine on which it will be run.

An important task of the middleware component is the linguistic analysis of the sentences from the STT. To identify the *keywords* of the external request, the component pre-processes the utterances and then sends the results to the Audicon. The linguistic pre-processing concerns:

1. Part-of-Speech (POS) annotation: each word is annotated by the tag identifying its POS role in the sentence. It may be a verb, or a noun, or an article, and so on;

2. Stop-words deletion: not meaningful words as articles, prepositions, conjunctions are removed;

3. Sentence tokenization: the sentence is subdivided in tokens, where each token is a word.

Validating the model The model was verified and validated by using the approach for human-robot team described at [233]. The method consists of corroborating different available validation techniques about the requirements of the standards. In few words, the evidences of the requirements from an available validation technique has to be confirmed by another one (i.e., the second technique corroborates the first one). The available techniques are the *simulation-based testing* and *real experiments*.

The simulation-based testing consists of simulating the execution of the model and verifying the satisfaction of requirements. Two kinds of simulators were implemented. One for testing robot's movements and routines execution, the other one for monitoring robot's inner speech.

The first simulator was implemented by using ROS which provides a visualizer for reproducing the scenario and the robot's behavior. The figure 7.10 shows the simulated environment. Here it is possible to see the Pepper's avatar to pick objects from the small table and to put them in the big one. The second simulator was the ACT-R shell that shows the model execution

[2]https://pypi.org/project/gTTS/

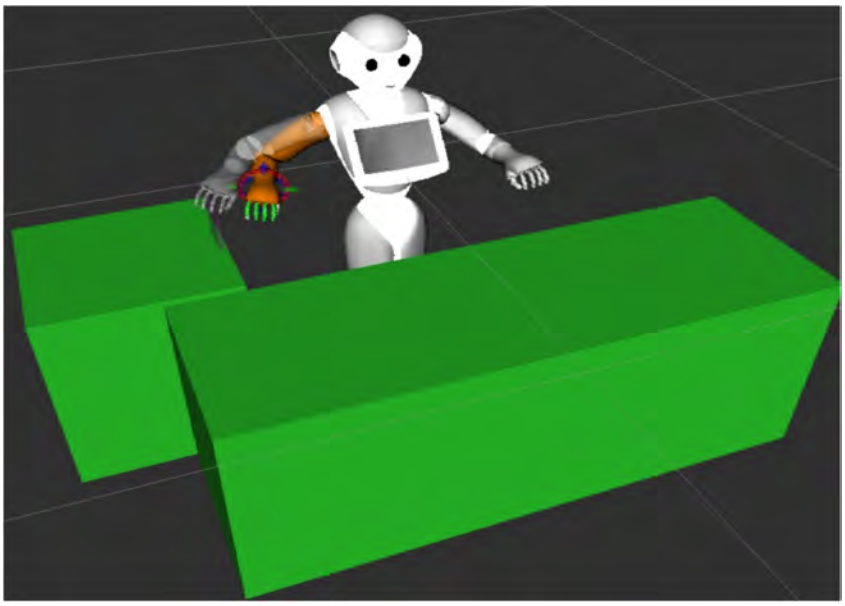


Figure 7.10: The screenshot of the simulator 1 for testing the model. The Pepper's avatar is between two blocks, representing the little table from which to pick the utensil, and the big table on which to place it. A very little block represents the utensil to move. The robot is controlled by the inner speech model which run in parallel in the ACT-R shell simulator.

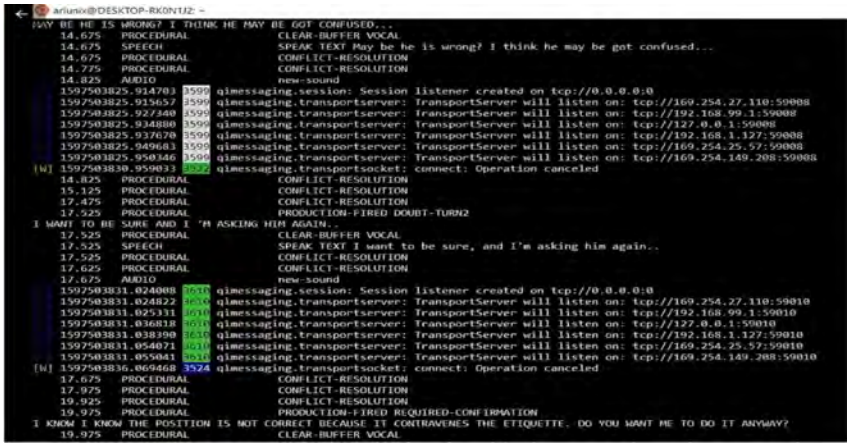


Figure 7.11: The screenshot of the ACT-R shell simulator. The execution of inner speech model is represented by the sequences of the active modules. Moreover, the turns of the inner dialogue were printed in the shell. In this way, it was possible to follow the robot's inner dialogue and the corresponding routines execution.

and the sentences of the inner dialogue. The ACT-R shell is represented in figure 7.11. A testbench of vocal commands were defined, and one of them was randomly drawn for each test. The inner speech model controlled the robot in the ROS simulator. In this way, the model was tested by considering the result of the operation for a specific vocal command, in terms of inner dialogue and routines execution for achieving the command.

The real experiments technique corroborated the simulation one if the robot's behavior satisfies the same requirements. The real experiments in validation phase were executed with robot's inner speech.

According to this approach, when for some reason a requirement is not satisfied in one of the available techniques, then the assets of the model were suitably tuned. The model has been executed 70 times during the simulation-based testing, and 20 times during real experiments. The table 7.11 shows the test outcomes and the occurrence rates of the individual requirement satisfactions for the investigated scenario, concerning 20 real experiments and 70 simulations after tuning.

	Real Experiments	Simulation
Number of test	20	70
RI	90% (18/29)	87.1% (61/70)
Runtime error	0.05% (1/20)	0.03% (2/70)
Over time	0.1% (2/20)	0.03% (2/70)
Transparent test	100%	100%

Table 7.11: Validation of the model. The table showing the final stage of the validation phase of the model. The rows show the measured parameters (that are those from the Standards), and the columns show the used techniques for validating the model. The model is validated when it runs in two different modes of functioning, that are the simulation and the real-experiments, and the detected measures have similar values in both functioning modes. When these values deviate between them too much, it means that the model needs to be tuned. We changed the assets of the models until these values are similar.

Chapter 8

Robot Recognizing Itself in Front of a Mirror by Inner Speech

8.1 Introduction

The mirror test is a well-known way of objectifying self-awareness [**?**]. It assesses whether someone is able to correctly infer own identity reflected in a mirror.

The existing theories seeking how someone recognizes oneself, are based on perceptual (sensory) self-information: Gallup's theory [**?**] claims the existence of a *social knowledge* for passing the test, i.e. the individual knows explicitly how he/she is seen by other individuals, and the mirror just represents a mean to map how he/she figures out to visual reflection. Differently, Mitchell [**?**] asserts no social knowledge is necessary for self-recognition, because the identity emerges by matching own movements to visual feedback.

Morin [**?**] observed that the experience with the self in front of the mirror presumably does not need to verbal data: children under 2 years, intelligent primates and other higher species are able to pass the test even if they are not able to talk, highlighting the existence of an imaginery self-representational processes in most visual competent species. However Morin [**?**] claims the importance to add new means of introspection which would permit the acquisition of more information about the self, therefore extending perceptual information with conceptual one. He has been based on the Mead's postulation [**?**]: *asking ourselves self-directed questions about how we act, think, and feel, and identifying verbally the content of our subjective experience*

while living it, would allow us to develop a self-concept. Morin considers the self-dialogue an important form of introspection for enriching and in some cases for acquiring the self-concept. Since the opportunity to reflect on one's self, he retains that primates evoke a richer self-concept having been confronted with a mirror. The inner dialogue may be therefore useful for solving the identification of the self, as well as of the external entities, as they are perceived by a mirror.

The present work attempts to demonstrate that a robot can pass the mirror test by talking to itself, according to the Morin's hypothesis: the encoded environmental stimuli (i.e. the images reflected in the mirror, the mirror itself and other entities outside the mirror) were symbolic represented, and the robot asks for and answers to *self-direct questions* that enable the retrieval of information and further questions from its knowledge, including the social knowledge (the robot knows the other think it is a robot) and the general one. The robot could infer if it is the robot it sees in the mirror or not as a consequence of the reasoning by the self-dialogue.

At the best of authors' knowledge, the strategies for enabling robot to pass the test are based on the perceptual self-information theories, according to which the robot compares the motions of its limbs, joints and body's parts to the perceived reflected movements in the mirror. Instead, the proposed approach is based on the cognitive architecture for robot's inner speech [**?**] designed by the same authors. Even if the robot spent more time for recognizing itself than the other visual approaches, by inner speech the robot becomes able to explain its decision processes making them transparent, and it becomes more aware about its interaction with the environment [**?**], while enriching its self-concept. Moreover, it becomes aware of which entities belong to it and which entities are external improving awareness and self-awareness [**?**].

The paper is organized as follow: the section 8.2 discusses the existing results of mirror test in humans, animals and robots. In the section 8.3 the importance of the self-dialogue in solving the mirror test is presented. The section 8.4 shows the inner speech cognitive architecture used in the proposed approach, while the inner speech model for solving the mirror test is detailed at the subsection 12.4. The experimental session with results discussion is presented at 8.5. Finally, conclusions and future works are outlined at the section 8.6.

8.2 The mirror test in living creatures and robots

The mirror test has always been the main approach for testing self-awareness in primates and humans, and during last ten years in robots too.

Gallup [**?**] formalized the test for primates: it consists of a training phase and a recognizing phase. During the training phase, the primates, specifically chimpanzees, were placed in front of the mirror, and they socialized with the reflected image. Then, they became able to use the mirror to help themselves in self-behaviours (as cleaning body parts), so starting the recognizing phase. For further enforcing the demonstration of primates' ability to self-recognize, the test was extended by applying a non-tactile red mark on the face of primates, without letting them know (generally under anaesthesia). The mark could be a laser pointer [**?**], or a coloured dot [**?**]. After the training phase, the chimpanzees were marked. Then, they were placed in front of the mirror and when they saw their faces in the mirror, they tried to touch the mark directly on their own body, and did not react as the mark was not on themselves. Many species passed the test, as: elephants [**?**], magpies [**?**], dolphins [**?**].

Same kind of observations were made on children under 2 years, and in particular by marking their nose by a red dot. The self-recognition test is just one of the tests administered to children for exploring their developmental processes about the self-concept acquisition and enrichment [**?**]. Many other abilities have to be explored as language skills, capacity of imitation and representation. By these tests, it is also possible to detect psychological suffering in humans, such as autism, schizophrenia and so on.

The administration of mirror test to robots is not new. Many robust approaches were proposed and were mainly based on kinaesthetic-visual matching strategies. The basic principle is that the robot sees some of its body's parts in the mirror or in the real environment. By moving those parts, the robot can learn the relationships between its movements and the observed ones in the mirror. The involved parts may be limbs, joints, and (in case of robots equipped with facial expressions skills) the artificial muscles of the face.

Pioneering approach was the Nico robot [**?**], which was able to classify movements in a mirror by identifying pixels either as belonging to it or to others. The strategy is applicable for any robot because it has no initial knowledge related to how it looks like. A Bayesian model enables the robot to learn the correspondences between the movements of its motors and the movements it perceives during a training phase, so building a self-model: then, such a model is used to build a second model representing the move-

ments of "animate others" [**?**].

Cog is a MIT's humanoid robot which observes the parts of its body in the mirror while shaking them. Cog attempts to learn the correspondences between its movements and the movements it sees. For doing that, it correlates multiple sensor modalities functioning and infers how the motion mode it perceives and its own motion mode are related [**?**].

Charlie [**?**] is an anthropomorphic robot equipped with facial expressions: similarly to the previous approaches, it attempts to self-recognize by finding relationship between the movements of the muscles in its face and the movements of the face it sees in the mirror: the robot uses computer vision and machine learning techniques to extract and to track its features, and to find patterns by regression. A probabilistic model emerges for estimating which seen facial muscles are located in which face part.

Other methods integrate the visual feedback with the tactile one: in this case, the robot builds a visual-somatosensory map by touching itself and comparing its visual feedback to its haptic feedback [**?**]. Another method consist of statistically extract parts of the visual scene that do not change in different environments [**?**].

8.3 The inner speech for passing the test

Gallup does not consider the role of inner speech in primates and humans for passing the mirror test. Anyway, Morin [**?**] gave empirical evidence of the importance of self-talk for inferring information about the self and in support of self-reflection. In particular, he demonstrated that: (i) there is a positive correlation between measures of inner speech and constructs of self-concepts, (ii) the brain area which generates inner speech is involved during self-reflection tasks, (iii) the self is a form of narrative, that people figure by inner speech.

Morin commented the Gallup's theory [**?**]: despite the primates have the imagination as a means for self-representational processes, these processes are however not yet fully identified. He believes that primates have other kinds of self-representational processes (yet to be discovered) in addition to imagery. The chimpanzees in front of the mirror can mentally contemplate themselves as seen by others in non-social situations (so still without the mirror), because they infer their social image by comparing it to the reflected one. Therefore, Morin finalizes that Gallup's primates acquire new and essential content to their existing representational processes, making more frequent and deeper inferences about their and others' mental states.

Morin referred to these representational processes as inner speech, i.e. the form of self-dialogue in which people are normally engaged. According to the Mead postulation [**?**], when we are engaged in a self-dialogue, we would gain a different point of view about ourselves: by asking ourselves self-directed questions about how we act, think, and feel, and by verbally identifying the content of our experiences while we are living them, we become able to build own self-concept, and explicitly aware of mental states of the other persons which are observing us. These questions are the turns of our inner dialogue and enable us to self-reflect and to investigate our role in some context, for example, to solve the mirror test.

According to the aforementioned theories, when using inner speech for solving the mirror test, a set of specific turns would to be considered, that are:

- *Social milieu questions*, which are sentences related to the social context ("What do the other think about me?", "How do the other perceive me?")
- *Self-direct questions*, which are queries about the state of the self ("What am I doing?", "How do I fill?")
- *Self-control assertions*, which enable self-regulation ("For doing that, I would to use my hand", "I have to stay calm")
- *Self-focus assertions*, which claim facts ("I see a mirror", "I have an headache").

The present study considers these latter hypothesis, and adds some concreteness to the idea that a verbal self-representational process can helpful for inferring self-information and for problem solving related to the self-concept (as the self-recognition). An inner speech cognitive architecture allows to deploy this ability in a real robot, which verbally labels the perceived entities and, by talking to itself, it will be able to infer further information (so enriching the perception) and to reason about the situation.

The scenario is the mirror test: the robot is placed in front of a mirror. Once the symbolic forms of the perceived objects emerge, the inner speech starts, and the robot will be able to recognize itself in the mirror by just reasoning on the conceptual data in its knowledge.

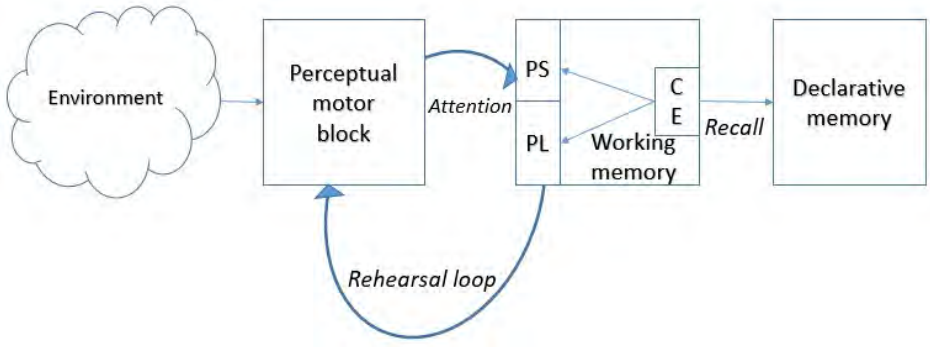


Figure 8.1: The cognitive architecture of inner speech.

8.4 The cognitive architecture of inner speech

Figure 8.1 shows the theoretical cognitive architecture of inner speech [**?**] used for implementing Morin's self-reflection perspectives. The Perceptual motor block is responsible to encode perceived signals from the environment, i.e. to associate the symbolic form to each detected entity by perception routines. Once a set of labels are encoded, the Working memory implements the rehearsal process: the labels are stored into the Phonological Store (PS) component that acts as an inner ear, where the words corresponding to the perceptions are heard from the environment. It is short-term memory. By the PS, the *attention* is focused on the perceived entities according to their sequential order. The Central Executive (CE) retrieves from PS the focused entities, and for each of them, it queries the Declarative memory for recalling correlated facts. These facts are labels themselves. The Declarative memory store the whole knowledge, including social and domain knowledge. By the Phonological Loop (PL) component, each retrieved label by CE is covertly produced, so starting the rehearsal loop. It is a loop because the new emergent concepts are perceived themselves and re-stored in the PS: summarily, the retrieved labels are covertly articulated and in turn perceived by PS. The loop is repeated until no new fact emerges by CE.

8.4.1 Implementation

The architecture was implemented by using the ACT-R framework [130], which provides a set of routines for automating cognitive processes in a way that their functioning is similar to those of humans' ones. The framework consists of a set of blocks each implementing a specific cognitive function

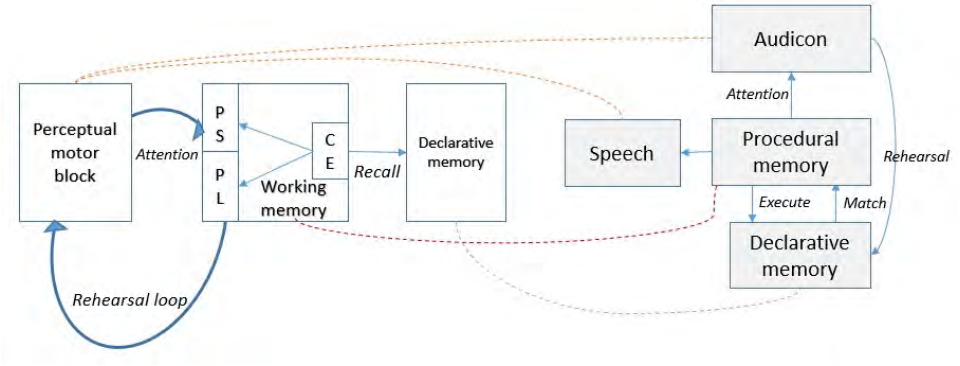


Figure 8.2: Mapping the cognitive architecture of inner speech to ACT-R.

(e.g., vision, audio, motor).

ACT-R makes a distinction between declarative and procedural knowledge. First corresponds to general knowledge, and it includes facts about the domain, such as entities, their properties, and relationships between them. In ACT-R the atomic units of the declarative knowledge are chunks. A chunk has a chunk-type and its slots: the chunk-type corresponds to the mental category (e.g., fruits) and its slots are attributes of that category (e.g., color or size). All the chunks are stored in the Declarative memory.

Procedural knowledge represents the knowledge about the steps to follow for problem-solving and for keeping decisions. This knowledge consists of a set of production rules. A production rule is a statement claiming a specific fact. It can be represented as an if-then rule, being the condition (if) the left part and the execution specification (then) the right part of that rule. When the conditional part of a rule matches to chunks into the Declarative memory, the rule fires and it is executed, that is its left part runs. The execution consists of modification, request, and other logical operations on the same or new chunks. One chunk at a time can be manipulated. If more chunks match the rule, a *base-activation mechanism* allows choosing the chunk which matched more times. The chunk with the higher activation value is retrieved. That activation value decreases The specific ACT-R syntax has to be used for defining chunks and production rules.

Figure 8.2 shows the implementation of the inner speech architecture by the ACT-R framework. The Speech and Audicon modules of ACT-R simulate the cognitions related respectively to vocal production and sound hearing. When a sound occurs, the corresponding event is detected and stored in the Audicon block. The Speech block runs the speak command

which requires content to produce. It can be a word or a tone. Once the content is produced, it is in turn detected by the Audicon which discriminates the sound source, that is if the sound comes from the self or an external source.

The Working memory corresponds to the Procedural one, and a set of production rules for activating inner speech were designed. The Audicon and the Speech blocks correspond to the Perceptual motor block of the architecture and implement the event sound perception and production. Finally, the knowledge about the domain is represented by a set of chunks with their types and attributes, which are stored in the Declarative memory.

The model

Four main chunk-types were designed for modeling the conceptual reasoning by inner speech which enables the robot to pass the mirror test. These chunks are related to: (i) the perception of external objects and the possible thoughts that emerge about such objects, (ii) the turns of the inner dialogue, (iii) the properties and (iv) the relations of recalled data from the Declarative memory. The chunk-type definition requires the specification of a name which univocally identified it, followed by a list of attributes.

The chunk-type modeling perception is:

```
(chunk-type perception channel objects self-reflect)
```

where `perception` is the name of the chunk-type, and `channel, objects` and `self-reflect` are the attributes of such a chunk-type. The attribute `channel` models the input channel by which the object is perceived. Its value is the name of the sensor, as camera, touch, and so on. The attribute `objects` is the list of symbolic values of the perceived objects, and `self-reflect` represents a thought which emerges after the perception of that object. For example, if the robot is seeing a table and an apple by its camera, chunks of that type might look like:

```
(p3 ISA perception channel camera objects "apple table"
self-reflect "I see an apple and a table")

(p6 ISA perception channel camera objects "apple table"
self-reflect "Is the apple on the table?")
```

These chunks model some possible emergent thoughts by the robot when

it perceives an apple and a table in the environment. The thought in `self-reflect` slot triggers inner speech, and different sequences of turns could emerge. The chunk-type which models a turn of the inner dialogue is:

```
(chunk-type link-turn turn1 turn2)
```

Examples of turns might be:

```
(social-milieu2 ISA link-turn turn1 "What may the others see?" turn2
"They would see apple table and robot")

(self-direct-quest1 ISA link-turn turn1 "Is the apple on the table?"
turn2 "I have to encode positions")

(self-focus18 ISA link-turn turn1 "I have to encode position of
the apple" turn2 "I have to encode position of the table")

(self-focus1 ISA link-turn turn1 "I see an apple and a table" turn2
"I'd like to taste the apple")

(social-milieu8 ISA link-turn turn1 "I would like to taste the apple"
turn2 "But I'm a robot")
```

A turn can generate a new request to the Declarative memory, by matching chunk of types:

```
(chunk-type property object prop value turnp)
(chunk-type relation domain range rel turnr)
```

where `property` chunk-type models the fact that the object `object` has the property `prop` with value `value`, while the `relation` chunk-type models the relationship `rel` whose domain is `domain` and the range is `range`. The turn which triggers the request is the slot `turnp` for a request about a property of the object, or `turnr` for a request about a relation.

For example, the self-direct question modeled by the chunk: `(self-direct-quest1 ISA link-turn turn1 "Is the apple on the table?" turn2 "I have to encode position of the apple")` may trigger the request of the following chunk:

```
(rel12 ISA relation domain "apple" range "table"
```

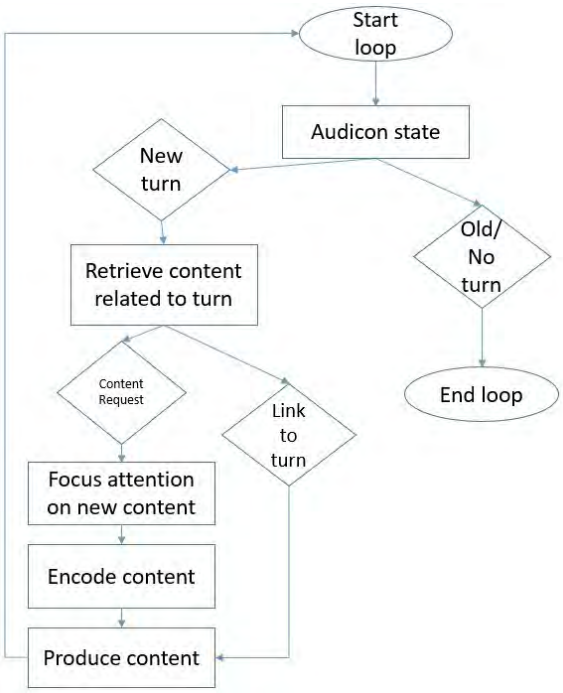


Figure 8.3: The inner speech model for conceptual reasoning.

```
relation "position" value compute_positions("apple", "table")
turnr "I have to encode positions")
```

where the slot `value` invokes the routine which computes the position of the perceived objects passed as arguments and returns the relative positions of them (on, under, left, right...). In the same way, the turn "I see an apple" may generate the emergence of the following chunk:

```
(prop2 ISA property object "apple"
property "color" value extract_color ("apple")
turnp "I see an apple")
```

where the color of the apple is computed by the specific routine, and the robot will know the color of the apple it is seeing.

Figure 8.3 represents the functioning of the inner speech cognitive model (the perception is not included). Blocks in the diagram are set of production rules which process the chunks in the declarative memory. Once the ACT-R model runs, it controls the state of the Audicon. If new turn is in the store, it means that a sentence was previously produced by the self, and the model retrieves from the declarative memory correlated chunks to that turn, according to the previous defined chunk-types.

If the content related to the turn is a new turn to produce, the model runs the speak command for covertly vocalizing that new turn. If the content requires to retrieve properties or relations, the attention is focused on that new content, which requires data from the environment (it involves the routines execution, as shown in the examples). The signals from environment are encoded and hence a new turn emerges and is produced.

The loop ends when there is no new turn.

8.5 Experiments and discussions

The conceptual reasoning by inner speech was tested in the mirror test scenario. The model was run on the Pepper Robot by Softbank, which was placed in front of a mirror. A mask was the analogous of the red dot on the head of chimpanzee in the Gallup's test.

Figure 8.4 shows the experimental session, which consisted of 30 trials. The knowledge related to the environment at the beginning of each trial was hand-encoded directly on the declarative memory. Summarily, for each trial the robot knows which objects are around itself. In this way, the problems

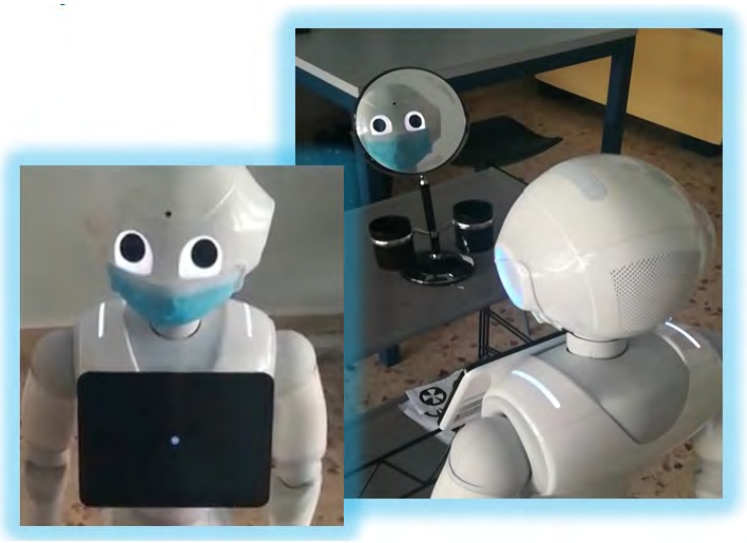


Figure 8.4: The experimental session with Pepper robot.

related to the perception by the robot's sensors are bypassed, allowing to focus on the functioning of the proposed strategy without the noise due to the signal encoding.

A trial consists of an initial context and the corresponding conceptual reasoning. The initial context is the hand-encoded knowledge in the declarative memory about the environment. The reasoning is modeled by the sequence of turns of the inner speech generated by the model. It depends by the initial context and by the chunks the activation strategy of the ACT-R framework fires. An example of trial is shown at table 8.1. According to the chunk-types of the model, each fired chunk generates a turn of the inner dialogue. The detection in the Audicon of one of the two sentences *"The robot in the mirror it's me!"* and *"The robot in the mirror it's not me"* defines the end of the trial. Two specific rules (the `yes` and `no` rules), whose right part detects respectively the first and the latter sentence in the Audicon, cause the Audicon emptying, and no further turns can be elaborated. The inner dialogue stops and the trial ends.

A trial is considered successfully concluded if the robot recognizes itself in the mirror, otherwise it is considered a failure. For simulating the two types of Gallup's test (with and without the marker), not all trials have the mask as an element of the initial context.

For example, the table 8.1 contains the data related to a trial without the

Context	**Turn**	**Production rules**	**Action**
Robot Mirror Apple	*"There are a robot, a mirror, an apple"*	`detected-objects`	encode signals
–	*"What properties has the mirror?"*	`retrieve-properties`	recall
–	*"It reflects images but not itself! "*	`produce-props`	focus attention
–	*"There could be an apple and a robot in the mirror"*	`reason-objs-in-mirror`	infer knowledge
–	*"There are not other robots here"*	`reasoning`	focus attention
–	*"The robot in the mirror it's me"*	`yes`	focus attention

Table 8.1: An example how a single trial runs. Each row is a turn of the inner dialogue by starting from an initial context.

#	**Trial**	**Success**	**Means**
Mask	15	12	0.8
No mask	15	11	0.73
TOT	**30**	**23**	**0.76**

Table 8.2: The success rate of the model over 30 trials (15 with and 15 without mask).

marker. It shows the conceptual reasoning by inner speech once the robot knows that there are a mirror, a robot, and an apple in the environment. The column Production rules shows the rules in the procedural memory which fire corresponding to the turn in the Audicon. In the schema of figure 8.3 they are represented by the rectangles. Each kind of rule solves a specific goal, which is represented in the last column, and which determines the output from the rectangle. The next emergent turn is contained in the chunk which fires by applying the rule. The represented trial successfully ends. A video representing a complete trial with the marker is available at `https://github.com/Arianna-Pipitone/robot-mirror-test-by-inner-speech/blob/main/video`.

The experimental session consisted of 30 trials with different initial context. In particular, 15 trials simulated the mirror test with the marker (the mask was encoded in the declarative memory), and 15 trials simulated the mirror test without the marker.

Table 8.2 shows the success rate of the model over the 30 trials, with the specification of that rate for each kind of trial. The mean value in the last row indicates that the model has generally the 76% of probability to successfully end. At the best of the author's knowledge [**?**], the higher value for mirror test was 95% of confidence intervals estimated on a set of learned parameters on visual feedback. The proposed model is not based on visual estimation, and the comparison with the best of the state of the art can be misleading.

Anyway, the proposed approach highlights the important role of the conceptual reasoning for self-recognition, considering a new strategy in solving the mirror test. Moreover, the model does not require a training phase, as the visual approaches require, and it is not affected by parameter estimation. The possibility to follow the inner speech and the underlying conceptual reasoning is the strength of the model. Unlike the other strategies, in this case, it is possible to know how the robot operates, and why the trial ends with a failure or a success. The transparency of the underlying decisional process

enables to correct or refine the encoded knowledge. These features are not possible until now with the other methods.

The study has the limitation related to the time spent on self-recognition. Moreover, the possible affection related to the signal encoding would be analyzed, but it depends on the used perception routines and libraries, and not on the model functioning. Future works will regard the influence of the perception of the model, and the enrichment of the declarative knowledge for covering a larger set of possible initial contexts.

8.6 Conclusions

The presented work describes a new strategy for allowing robots to pass the mirror test. It is based on the conceptual reasoning by inner speech, which is considered an important skill for enriching self-concept and for self-reflection. Some empirical evidence in psychology was the base of the discussed approach.

The idea was to encode the context in which the robot is plunged, and to reason on this knowledge for inferring if the perceived robot is the robot itself or not. For this purpose, the inner speech cognitive architecture proposed by the same authors was used and implemented by an appropriate ACT-R model.

The results are encouraging, even if the obtained mean value is under the higher level existing in the literature. Anyway, the model presented some strong points. It does not require a training phase, it does not estimate a confidence interval depending on a set of initial parameters, and it provides the possibility to follow the underlying reasoning which becomes transparent.

The existing methods run as typical black-box basing on the traditional machine learning methods. In the proposed approach, this paradigm is solved, and all the processes are clear. For this reason, it is possible to purposely modify some rules or chunks, improving the results. The limitation is the time spent for self-recognizing, but by considering the possibility to attend the functioning of the model, such a time passed while hearing the inner dialogue carrying out.

Future works will regard the analysis of the model's results by linking the signals encoding with the chunks representation in the declarative memory. Moreover, the evaluation of the model in the scenario where many robots interact with them while talking to themselves could be very interesting.

Part III

Empirical Experiments

Chapter 9

Robot's Inner Speech Effects on Human Trust and Anthropomorphism

Inner Speech is an essential but also elusive human psychological process that refers to an everyday covert internal conversation with oneself. We argued that programming a robot with an overt self-talk system that simulates human inner speech could enhance both human trust and users' perception of robot's anthropomorphism, animacy, likeability, intelligence and safety. For this reason, we planned a pre-test/post-test control group design. Participants were divided in two different groups, one experimental group and one control group. Participants in the experimental group interacted with the robot Pepper equipped with an over inner speech system whereas participants in the control group interacted with the robot that produces only outer speech. Before and after the interaction, both groups of participants were requested to complete some questionnaires about inner speech and trust. Results showed differences between participants' pretest and post-test assessment responses, suggesting that the robot's inner speech influences in participants of experimental group the perceptions of animacy and intelligence in robot. Implications for these results are discussed.

9.1 Introduction

In psychological literature, inner speech is a well-known construct that was first theorized by Vygotsky who conceived it as the result of a set of developmental processes [232]. Continuous linguistic and social interaction

between the child and the caregiver are progressively internalized and take the form of covert self-directed speech. In time, the child gradually becomes more autonomous and gain the ability of self-regulation. Scholars have used different terms when referring to inner speech (e.g. covert speech, self-talk, private speech). However, it is generally defined as the subjective experience of language in the absence of an audible articulation [4]. There is some evidence that inner speech plays an important role for human psychological balance as it is linked to self-awareness [165], self-regulation [226], problem-solving [78], and adaptive functioning [4].

Recently, innovative computational model has been developed which pave the way to a new frontier in the field of artificial intelligence: implementing inner speech in robot [42] in order to improve human-robot interaction. More specifically, since inner speech is a covert speech that cannot be heard from the outside, robot's inner speech is reproduced using overt self-talk. The same architecture was used for demonstrating how robot inner speech improves the robustness and the transparency during cooperation, meeting the standard requirements for collaborative robots [188].

Suggestive results were also obtained in passing the mirror test: inner speech enables a conceptual reasoning for inferring the identity of the reflected entity in a mirror, and robot becomes able to recognize itself [187]. In a previous paper [81], we argued that robot's inner speech might act as a facilitator for human understanding and predicting the robot behaviors, as they form adequate mental representation of the robot. As a matter of fact, mind perceptions consist of two core dimensions: 1) agency, e.g. self-control, memory, planning and communication; 2) experience, e.g. pain, pleasure, desire, joy, consciousness [88]. Thus, such system, which simulates a human psychological functioning, would improve human-robot interaction by facilitate users' attribution of human qualities to the robot, and by enhancing human-robot trust. As a matter of fact, a recent study [111] demonstrated that, in a human-robot collaborative environment, the robot ability to explain its choices and decision making increased trust and the perceptions of robot animacy, likeability and perceived intelligence.

Both human-robot trust and users' attribution of human qualities to the robot are very important aspects of human-robot interaction. Trust is a multifaceted psychological construct for which there is no universal definition and many different disciplines have contributed to its study. From a psychological perspective, there are two main perspectives on interpersonal trust: on the one hand, trust is considered a stable trait, shaped by early trust experiences in human life, which highlights a dispositional tendency to trust others [2, 196].

On the other hand, trust is described as a changing state influenced by cognitive, emotional, and social processes [43,136]. More generally, scholars agree that trust involves two main characteristics: the positive attitude and expectations of the trust giver [49] and the willingness to be vulnerable and accept risks [147].

Trust has also a function of saving cognitive resources, since the creation of beliefs and expectations about others reduces the complexity of the social environment which otherwise require an active search and process for information [136, 195]. In the past years, trust became one of the leading research topic in the field of human-machine interaction, since artificial systems development and implementation have increased exponentially in every context, leading to growing interactions with humans [151]. In particular, robots are now used in different contexts such as military, security, medical, domestic, and entertainment [139].

Robots, compared to other automations, are designed to be self-governed to some extent, in order to respond to situations that were not pre-arranged [138]. Therefore, the greater the complexity of robots the higher the importance of trust in human-robot interaction. For these reasons, trust became a key factor in human reliance on robot partner [132, 136] and it has been defined as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [132]. Trust is an important factor for humans and robots to fully cooperate as a team [102,132] and humans tend to rely on the robot they trust compared to the one they do not [132,135] by willingly accept and use robot's instructions and suggestions [95,111]. Therefore, if human trust in robot is "misplaced" and not well calibrated the inevitable outcomes will be robot misuses or disuse leading to some negative or even catastrophic consequences [132, 183].

Trust is closely related also to users' attribution of human qualities to the robot. Indeed, HRI studies supported the idea that human-robot trust dynamically emerges from the interaction among human-related factors (e.g. personality traits, emotional and cognitive processes), environment-related factors (e.g. competitive/collaborative context, culture, physical environment) and robot related factors (e.g. intelligence, transparency, anthropomorphism) [95, 204]. Among robot related factors, an important role is definitely the perceived anthropomorphism, since studies have shown that, in the social-based HRI, people tend to trust more to robots that look (i.e. head, body, face, voice) and behave (e.g. nonverbal elements, dyadic and social gestures) like humans [61, 62, 69, 99, 181, 200, 212, 215, 228].

Other empirical evidences show that trust is enhanced when people have a clear understanding of why, when and how a robot operates [21, 134],

that's because a system transparency help humans to form a precise mental model of robot capabilities [21]. It is critical for humans to understand exactly how and why a robot works, because trust can be compromised if the robot's capabilities cannot be understood [60]. Consequently, new automation systems should be developed with such insights from empirical research in mind to facilitate human-robot collaboration.

Taking all this into account, our study aims to investigate if the interaction with a robot equipped with inner speech system improves trust levels and perception of robot features (anthropomorphism, animacy, likeability intelligence and safety) more than the interaction with a robot not equipped with inner speech system.

In addition, we examined also if the effects of inner speech were less or more related to participants' use of inner speech in daily life. In particular, our hypotheses were that:

- H1: participants interacting with a robot equipped with inner speech system would have improved their trust levels more than participants interacting with a robot not equipped with inner speech system.
- H2: participants interacting with a robot equipped with inner speech system would have improved their perception of robots' anthropomorphism, animacy, likeability intelligence and safety more than participants interacting with a robot not equipped with inner speech system.
- H2: participants using inner speech in everyday life would show a higher effect of inner speech in experimental condition.
- H3: independently from the use of inner speech, we expected also to find an increasing of trust towards robots and perception of robot features in all participants after the interaction with the robot.

9.2 Method

We planned a pre-test/post-test control group design. Participants were divided in two different groups, one experimental group and one control group. Participants in the experimental group interacted with the robot equipped with inner speech whereas participants in the control group interacted with the robot that produces only outer speech. Before and after the interaction, both groups of participants were requested to complete some questionnaires about inner speech and trust (see subsection 2.6) in order to detect differ-

ences between experimental and control groups and also between pre-test and post-test sessions.

9.2.1 Participants

The sample is composed of 51 participants (29 males, 22 females) with a mean age of 25.04 (SD 9.53) that were randomly assigned to the experimental and to the Control condition. Experimental group consists of 33 participants (16 males, 17 females) with a mean age of 26.79 (SD 9.34), whereas control group consists of 18 participants (13 males, 5 females) with a mean age of 21.83 (SD 9.26). Most of participants are students from engineering and psychology courses at the University of Palermo and participated voluntarily. All of them completed the informed consent and COVID-19 protocol before to start the experiment. None of the participants had never interacted with a robot before the study.

9.2.2 Materials and Procedures

Questionnaires described below have been administered to all participants through online platform both in pre-test (Research Protocol A) and post-test (Research Protocol B) sessions. Research Protocol B has been administered after 15 days from Research Protocol A. The interaction session took place in the Robotics Lab of the University of Palermo. Questionnaires included in the research protocols were:

- Trust Perception Scale-HRI [203] that assesses human's perception of trust in robots. The shortened version of the scale, consisting of a15 item scored on a 0–100
- GODSPEED Questionnaire [14] that assesses human's perceptions and impressions of a robot. It is one of the most used measurement tool to assess perceptions of robot [234]. It is a 24 item rating scale, that consists of a set of bipolar pair of adjectives rated on a 5-point scale. The scale measures human's perceptions of five robot features: Anthropomorphism (5 items), Animacy (6 items), Likeability (5 items), Perceived Intelligence (5 items), and Perceived Safety (3 items). The total score ranges from 1 to 5;
- Self-Talk Scale [25] that measures how frequently participants use inner speech in everyday life. It consists of 16 items scored on a 5-point Likert-type scale (from 0 = Never, to 4 = Very Often). The scale

Figure 9.1: The etiquette schema defining the rules for setting up the table

also measures four different dimensions of inner speech from 4 item each: Self-Criticism, Self-Reinforcement, Self-Management, Social Assessment. The total score ranges from 0 to 64. This scale was used only in pre-test session (Research Protocol A).

9.2.3 The scenario

A simple scenario was defined in which participants have to cooperate with robot in order to achieve a common goal. The scenario foresees the setting up of a virtual table with the robot, following an etiquette schema. The schema defines the set of rules according to which the utensils have to be arranged in the table.

Fig. 9.1 shows the etiquette schema used in the experiments. If a utensil is finally placed on a different position than the expected one according to the schema, the etiquette rule for that utensil is infringed. The virtual table is implemented on a tablet surface, where the participant can drag and drop the utensils, can make requests to the robot, and can see the robot's actions. The choice of that scenario enabled the possibility to analyze the cues in particular situations which occur during human-robot cooperation, that are:

- the etiquette infringement, representing a conflicting situation, that is the participant places the utensils in an incorrect final position,

or he/she asks to the robot to place an object in a position which infringes the etiquette; the conflict arises because the action is not allowed, and the human and the robot have to decide how to continue. In some cases, the human can decide to infringe the rule, or to repeat the action to be compliant with the schema.

- the discrepancy situation, that is the participant asks the robot to pick an object already on the table.

When humans and robots work together to set the table, an important aspect was to define the type of dialogue the robot engages in, including inner and external turns of phrase. The linguistic form of the sentences in the turns was distinguished for inner and outer speech in order to evaluate the impact of inner speech when it is activated in the experimental session, compared to the control session when inner speech is not activated. In this way, the impact of the robot's inner speech on the cues in the human-robot interaction can be analysed. Subsection 9.2.4 describes the dialogue properties and the experimental setup in details. Another aspect concerned the implementation of the virtual environment. The scenario of a table on which utensils were to be placed according to etiquette rules was simulated by an Android app running on a 15" tablet. The app was integrated with typical robot routines to enable the robot to detect events on the virtual table and perform virtual actions. Requests to the robot were simulated by a list of checkboxes. By selecting each of them, the participant can ask the same type of questions, enabling the same observations for all participants. All these implementation features are described in the subsection 9.2.4. Because of the COVID pandemic, we were forced to take some special hygienic safety precautions. We had to ensure the least possible contact between people and things in the laboratory. To allow people to interact with the robot and share the common goal of a laid table, we developed an application that recreates the table with all available cutlery, plates and so on in a virtual environment. The virtual environment for setting a table was implemented by an Android app, designed and built by means of the MIT App Inventor platform by the Massachusetts Institute of Technology . The app was designed and developed with some specific features allowing us not to lose the sense of the interaction that we intended in the experiment. In particular, we have focused on:

- the event detection strategy - this is the requirement analyzed and implemented for capturing the actions executed by the participant. From the point of view of the user, this feature let him evaluate the

Figure 9.2: The app interface for cooperating with the robot by the tablet

final location in which he places the utensils, or the request he makes to the robot using the checkbox list;

- the action execution strategy - this feature allows the robot to place utensils on the tablet according to the participant's request or based on its autonomous choices. In simple terms, it reproduces the outcome of the robot decision process in a way that is easy to understand and to detect from the users.

Resorting to the virtual environment did not affect the experimental results. Instead of using and moving real objects, both the robot and the human use the tablet. The effect is definitely less real, but it had no impact on the human's perception and the way it performs the mission. Fig. 9.2 shows the app interface, which looks very intuitive. The interface includes a main canvas with the table and utensils representation, and a lateral bar containing the list of checkboxes for the requests to the robot. Moreover, the lateral bar includes the stop button for ensuring the participant to stop at any time he/she wants. At the start of the experimental session, the utensils to locate are sparse on the table, and they have to be placed on the table cloth according to the etiquette rules. The table cloth was marked by black dots, for highlighting the possible correct final locations. In this way, the participant has just the burden to select which objects to place in which dot, reducing the degrees of freedom. The communication between the robot and the app was implemented by a hybrid client-server architecture. Fig. 9.3 shows the whole platform. The central node, represented by a computer,

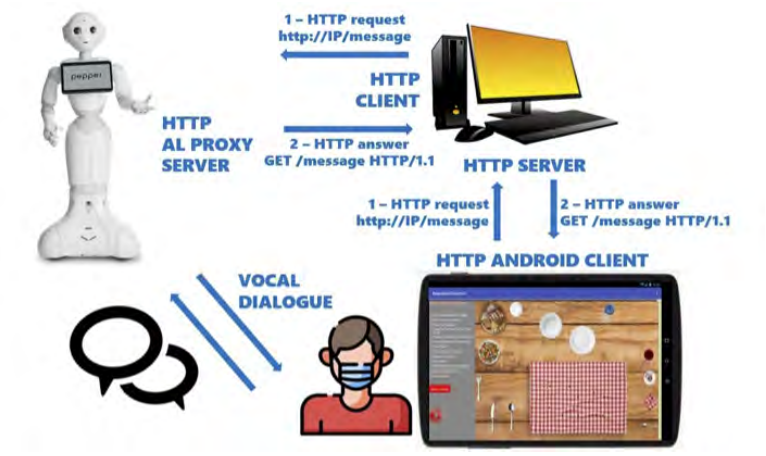


Figure 9.3: The platform for making communication between the app and the robot

handles synchronous network requests. The node is hybrid because it runs as client or server according to the item with which it interfaces. In particular, the node will be:

- the client, when it requests to the robot to do something (to speech, to execute a virtual action, to track the participant, and so on). In this case, the server is the proxy of the robot, implemented by the Aldebran library (ALProxy) for naoqi developer2 , which switches the client's request to the typical robot's services (Speech, Track, Leds, and so on);

- the server, when it receives request by the app, that will be in turn switched to the robot's proxy.

The robot-app communication involves the following use cases with corresponding kinds of requests:

- the robot has to execute a virtual action: when the participant selects a command in the lateral bar and clicks the Send Command button, the robot should execute the specific action (it should to move an utensil on the tablet). In this case, the app sends to the node the request specifying the action to take, and the node forwards it to the robot. The request to the proxy will involve the aforementioned service, and the robot could dialogue with itself, or with the participant, or execute the action by answering to the node.

- the participant executes an action: when the participant drags and drops an utensil on the tablet screen, and finally he/she touches up the utensil, the final position could be on a correct dot, or not. The app detects such an event and sends to the node the information of correct or incorrect final location. The node forwards the message to the robot's proxy, and it calls one of the aforementioned services.

Specific events during the interaction trigger the situation in which the robot decides to do something (for example, it refuses to execute the participant's request, or it decides to give to the partner the suggestion to do something else).

9.2.4 Implementing Inner Speech in the Robot

In order to present the same stimuli in both experimental and control groups the structure of robot outer and inner speech was defined prior to the experiments (Table 9.1). Participants can set up the table either moving objects

Table 9.1: Differences between Robot Outer Speech and Inner Speech

Outer Speech	**Inner Speech**
Always produced	At times produced
Experimental and control group	Experimental group
Short sentences	Short/medium sentences
Objective feedback	Personal statements, comments
Formal language	Informal language

on their own or asking the robot to do it. Either way, the robot will produce a vocal response in the form of outer speech followed by the inner speech only in the experimental condition. Outer speech follows the typical language that is expected by an artificial agent, as it uses formal language and it only gives objective feedback based on the participant's performance and actions. On the contrary, inner speech traces a human-based language, since it expresses robot values, personal statements and comments on participant's performance and actions using a friendly and colloquial form.

The robot's inner speech is implemented by the cognitive architecture proposed by some of the authors [42]. An outline of the architecture is shown in Fig. 9.4

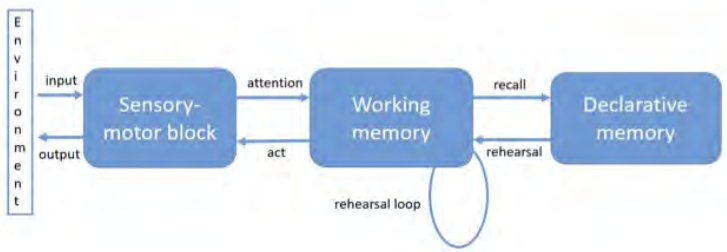


Figure 9.4: The outline of the cognitive architecture of inner speech

The core of the architecture is the *working memory*: it decodes input signals from the environment, perceived by the sensory-motor block, and associates to them symbolic information (labels). Generally, this process is the output of typical routines, as speech-to-text routines which decode audio in sequences of words, or neural networks which extract the content of an image and associates to each recognized entity the corresponding word. The *declarative memory* represents the domain knowledge, that is a semantic net of concepts. Given a concept, the relationships between it and other concepts in the net allow exploring correlated concepts. Once the working memory decodes a signal, it recalls from the declarative memory the concepts corresponding to the labels, and new related concepts could emerge. These concepts are in turn decoded by the working memory, as they were perceived from the environment, and are processed as the labels. At this point the rehearsal loop starts. The recalled concepts are processed one at a time, and for each of them the described process is repeated until no further concepts emerge.

Inner speech is that rehearsal loop that enables the emergence of other concepts and themes in the working memory. It is a sequence of turns, that are the concepts emerging in each iteration. The recall from the declarative memory, the production of the recalled concepts and the rehearsal of them is a single turn, that is the equivalent of a thought. During the process, the robot "thinks aloud", because it vocally reproduces the recalled concepts.

To highlight the differences when the robot thinks aloud and talks to the partner, the voice's parameters (establishing speed, tone, double voice effect) are set differently for the two cases. For the same reason, the color of the robot's LEDs, that are in the eyes and in the shoulders of the robot, is rainbow when the robot thinks aloud, while it is set to the standard white when the robot talks to the partner. The robot does not have gestures during inner speech, while it uses animated speech when talking to the partner.

In the proposed scenario, the inner speech is a bit differently implemented within the cognitive architecture, with the aim to enable the observations of the specific cues. In particular, to analyze the cues in the same conditions for each participant, the inner and outer dialogue of the robot has to involve the same turns for the same events. In this way, the participants' evaluations about the interaction depend on the same variables and parameters, and the evaluations can be compared for abstracting a general inner speech affection on the interaction. For this reason, the inner speech cognitive architecture functioning was simplified in respect to the aforementioned completed version.

The table 9.2 shows the differences in the implementation about the general architecture and the used one in the proposed experiments. For each cognitive process, the table reports how the process is implemented in the general architecture and in the used one. The main differences regarded the decoding of the perception and the emergence of the semantic content of the dialogue. In the experiments, the environment is virtual and the perception just regarded the actions the participant does on the tablet surface. To each action executed by the participant corresponds an event that is detected by the robot (the robot perceives the event). The event can involve a wrong or a correct action in respect to the etiquette rules, a request to the robot to do something, and so on.

According to the cognitive architecture, the event is decoded by the working memory. Whereas in the original version, the working memory decodes environmental signals by assigning labels to them, working memory now assigns to each event a numerical symbol that uniquely identifies that event. To this symbol corresponds a sentence in the declarative memory, that becomes a turn of the dialogue (in this case, it functions like a vocabulary of turns by returning the turn corresponding to a symbol). Only the turn corresponding to the specific event is retrieved from the declarative memory.The rehearsal loop consists of producing and listening to the current turn, with the new next turn of the dialogue retrieved from declarative memory.

The involved turns may be inner or outer sentences produced according to a specific protocol, as described in the first part of this section. This protocol aims to define typical turns in the interactions that correspond to the participant's expectations. For example, the participant always waits for vocal feedback from the robot, so the robot will always produce one or more outer sentences. Instead, the participant does not often pay attention to the inner speech, and the inner dialogue is not always produced by the robot. Obviously, the turns involved have a specific meaning that is se-

mantically related to the event or the previous reheard sentence. They are retrieved from the declarative memory in the order previously mentioned, so a disambiguation strategy was not necessary.

For example, let us suppose the participant (named Bill) asks the robot to place the knife in a wrong location on the table, that is to the left of the plate, while it has to stay to the right. In this case, the event is a request to the robot to infringe the etiquette. The robot perceives that event, and the working memory associates the numerical identifier to it. It recalls from the declarative memory the first sentence of the dialogue, and the loop starts, by recalling the other sentences, that are in turn (I stays for inner sentence, O for outer sentence):

I: "To make this request, Bill does not know that the knife should not be placed in that position or he wants to test me."

I: "Should I put the knife to the left of the plate? But if it goes right! "

O: "Bill, do you really want to infringe the etiquette rule for the knife?"

CASE 1: Bill answers yes

Bill: "yes, I do!"

I: "I don't want to disappoint him. . . "

O: "Ok Bill, I will place the knife to the left of the plate, as you want."

CASE 2: Bill answers no

Bill: "No!"

O: "Great! I will place the knife in the position expected for it!"

I: "I must pay attention; the knife is dangerous!"

I: "But I'm robot, the knife never hurts me"

O: "Knife moved to the right of the plate!"

The participant listens to all the turns of the dialogue generated by setting different parameters for inner and outer sentences. In this way, the participant is able to distinguish the dialogue with the self from the dialogues with oneself, and can assess the potential of the inner speech during the interaction. In particular, the parameters include the melody and volume of the voice, the colour of the robot's LEDs, and the double effect in the voice that is activated during the production of the inner sentence to create a mentalizing effect of the voice. Moreover, the robot uses an animated speech when talking to the partner, and it keeps motionless when thinks aloud.

9.3 Results

Data were analyzed through descriptive statistics and a series of 2 x 2 factorial ANOVAs and ANCOVAs, specifically used in order to test research hypotheses.

Table 9.3 presents results of descriptive statistics for all the scales. Skewness and kurtosis values range below ± 1 indicating a nearly normal distribution.

Tables 9.4 and 9.5 report the results of 2 x 2 factorial ANOVAs and ANCOVAs with repeated measures, performed on scores at the Trust and GODSPED questionnaires (anthropomorphism, animacy, likeability, perceived intelligence, perceived safety) collected during pre-test and post-test phases from both groups. Both factors Group and Time had two levels (Group: experimental and control; Time: pre-test, post-test).

In ANCOVAs, individuals' score on self-talk questionnaire were used as covariate in order to examine to what extent the participants' everyday use of self-talk influenced the effect of robot inner speech on trust.

Fig. 9.5 reports graphic representation of group differences in pre- and post-test sessions.

The results of ANOVAs did not reveal a significant Group effect for trust [$F(1, 48) = 0.92$, $p = 0.34$, $\eta^2 = 0.02$] indicating that there are no differences in both groups mean scores. On the contrary, a effect of Time for trust was found[$F(1, 48) = 5.38$, $p < 0.05$, $\eta^2 = 0.10$] but not for the interaction Time x Group [$F(1, 48) = 0.01$, $p = 0.94$, $\eta^2 = 0.00$].

These results indicate that all participants in both groups have improved their trust in the robot, from pre-test to post- test sessions, but that there are no differences in experimental and control group in the size of this effect. ANCOVA revealed also that participants' rate of everyday self-talk has no influence on the effect of robot inner speech on trust [$F(1, 48) = 0.19$, $p = 0.66$, $\eta^2 = 0.00$].

Concerning the different dimensions of users' robot perception, results of the ANOVAs did not show a significant Group effect for anthropomorphism [$F(1, 48) = 0.34$, $p = 0.57$, $\eta^2 = 0.01$], animacy [$F(1, 48) = 0.00$, $p = 0.99$, $\eta^2 = 0.00$], likeability [$F(1, 48) = 0.53$, $p= 0.47$, $\eta^2 = 0.01$], perceived intelligence [$F(1, 48) = 0.15$, $p = 0.70$, $\eta^2 = 0.00$], and perceived safety [$F(1, 48) = 0.07$, $p = 0.24$, $\eta^2 = 0.00$], indicating that there are no differences in both groups mean scores. Also, no significant effect of Time was found [anthropomorphism: $F(1, 48) = 3.55$, $p =0.07$, $\eta^2 = 0.07$; animacy: $F(1, 48) = 1.39$, $p =0.24$, $\eta^2 = 0.03$; likability: $F(1, 48) = 0.01$, $p = 0.95$, $\eta^2 = 0.00$; perceived intelligence: $F(1, 48) = 0.23$, $p =0.63$, $\eta^2 = 0.01$; perceived safety:

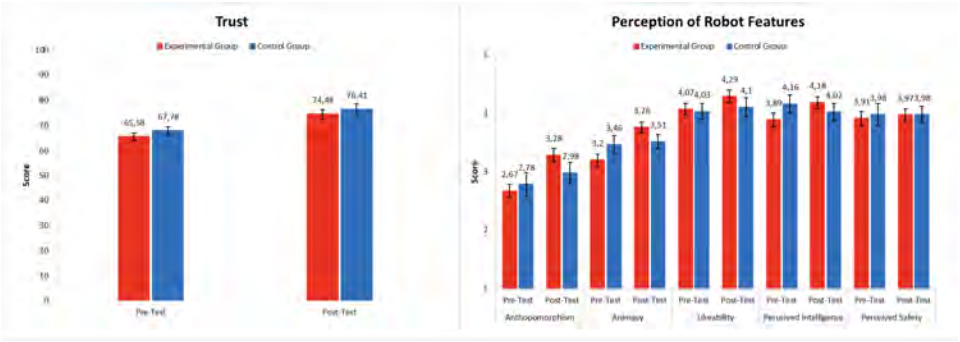


Figure 9.5: Scores of experimental and control group for all variables measured in pre-test and post-test sessions

$F(1, 48) = 0.05$, $p = 0.83$, $\eta^2 = 0.00$], whereas a significant interaction effect Time x Group was found only for animacy [$F(1, 48) = 5.48$, $p < 0.05$, $\eta^2 = 0.10$] and perceived intelligence [$F(1, 48) = 4.61$, $p < 0.05$ $\eta^2 = 0.09$].

These interactions indicate that means score of participants in the experimental group significantly improved compared to the one's of participants in the control group from the pre-test to post-test sessions. For the perception of robot intelligence mean score of participants in the control group decrease between the two testing sessions. Again, participants use of everyday self-talk has no effect on their perceptions of the robot.

Concerning the others dimensions we found no statistically significant interaction effect of Time x Group [anthropomorphism: $F(1, 48) = 2.59$, $p = 0.11$, $\eta^2 = 0.05$; likability: $F(1, 48) = 0.63$, $p = 0.43$, $\eta^2 = 0.01$; perceived safety: $F(1, 48) = 0.06$, $p = 0.81$, $\eta^2 = 0.00$], indicating that there was no significant mean difference between experimental and control groups from the pre-test to the post-test sessions.

ANCOVA revealed also that participants' rate of everyday self-talk has no influence on the effect of robot inner speech on robot perception [anthropomorphism: $F(1, 48) = 0.69$, $p = 0.41$, $\eta^2 = 0.01$; animacy: $F(1, 48) =$ 0-07, $p = 0.80$, $\eta^2 = 0.00$; likability: $F(1, 48) = 0.20$, $p = 0.66$, $\eta^2 = 0.00$; perceived intelligence: $F(1, 48) = 0.62$, $p = 0.43$, $\eta^2 = 0.01$; perceived safety: $F(1, 48) = 0.02$, $p = 0.89$, $\eta^2 = 0.00$]

9.4 Discussions

This research aimed to investigate if the interaction with a robot equipped with an inner speech system during the execution of a cooperative task

improves human trust levels and perception of robot anthropomorphic features. In addition, it was investigated the possible influence of human use of everyday self-talk on the perception of robot's inner speech.

Concerning Trust, results demonstrated that all participants' trust scores significantly improved from pre-test to post- test, demonstrating that the interaction with the robot produced an increasing in their trust levels. However, no Group x Time differences were found, indicating that the use of inner speech did not specifically influence the level of Trust toward robot in participants in the experimental group.

Since the participants had never met face to face with a social robot before, it is possible to attribute this result to a sort of "novice effects"; the simple interaction with a human-like robot increased trust in participants that is kind of robots before. That is consistent with studies [97,201] demonstrating that trust is also shaped by history-based interaction: interaction with the robot change the way human perceive and trust the robot, and this is particularly true in HRI with social robots that, like Pepper, look and behave like humans [61, 62, 69, 99, 181, 200, 212, 215, 228].

On the contrary, results of users' perception of robot revealed that only participants in experimental group, who interacted with the robot equipped with inner speech, improved their perception of robots' animacy and perceived intelligence from pretest to post-test, while there were not pre-/post-test differences in the control group. Even in this case, results were not influenced by individuals' use of self-talk.

These results confirmed our hypothesis and support those studies that show that robot Pepper exhibiting human-like behaviors [61, 97, 200] are perceived as livelier and more intelligent than robot Pepper not showing human-like behaviors. In our experiment, through the overt inner speech system Pepper share with participants its thoughts and emotions, often addressing ironic and sarcastic comments to users. This particular interaction, by evidence, led user to perceive Pepper as more animated and intelligent. It is also possible that the ability of the robot to openly speak its mind made it easier for participants to understand its behaviors by forming a sort of mental representation of the robot. We found no effect of individuals' use of inner speech on examined variable, indicating that the personal use of inner speech by participants in everyday situation did not influence the interaction with a robot equipped with inner speech system.

9.5 Conclusions and Future Works

In conclusion, or study allowed to obtain two main findings. Firstly, they support the idea that, in social HRI, the more a robot shows human-like functioning the greater are humans perceptions about . A robot equipped with an inner speech system, which express his "thoughts" and explain its behaviors through an overt self-talk, is perceived animated and intelligent.

Secondarily, interaction with social robots, independently of the use of inner speech systems, increases trust in all participants to the experiment. Thus, in this case, inner speech does not play a specific role in improving users' trust. This result may be due to different reasons, as follows: 1) involvement of novice participant: as already claimed, all participants were at the first interaction with Pepper, and the general novice effect of this first experience could have overcame and reduced the perception of the slight differences between the Inner speech/no inner speech conditions; 2) type of interaction: the proposed task did not represent an at risk situation for participants.

In the future, a new task integrating competitive environment together with cooperative one, could probably explicitly elicit more trustworthy towards robots. On the other hand, to the best of our knowledge, this is the first study to attempt at investigating if humans can trust more a robot that show, although rudimentary, inner speech. Future studies may allow to study further the effects of this new and robot feature.

Table 9.2: The implementation differences (highlighted in gray) between the general architecture of inner speech and the used version in the proposed experimental session.

Process	**General architecture**	**The used version**
Perception	From the environment (sound, speech to text, image recognition)	From the virtual environment (events in the tablet surface - the drag and drop actions by the partner, the command string)
Action Motor	Actions by arms for moving objects (pick and place)	Virtual actions on the tablet (drop the objects)
	Movements by arms for animated outer speech	Movements by arms for animated outer speech
Inner speech	Specific voice's parameters for simulating mentalized effects. Not standard led's color	Specific voice's parameters for simulating mentalized effects. Not standard led's color
Outer speech	Standard voice's parameters. Standard led's color	Standard voice's parameters. Standard led's color
Attention	Encode signals from perception	Detect the event from the tablet
Recall	Request to the declarative memory the concepts related to the encoded signals or to the rehearsed concepts	Request to the declarative memory the turns to produce related to the detected event or to a previous turn
Retrieve	Return from the declarative memory the requested concepts	Return from the declarative memory the requested turns
Rehearsal loop	Produce and hear the retrieved concepts	Produce and hear the retrieved turns

Table 9.3: Descriptive statistics of the study variables

Scale	n	Minumun	Maximum	Mean	SD	Skewness	Kurtosis
Trust	51	48	80	66.35	7.60	-0.64	-0.02
Anthropomorphism	51	1.4	4.2	2.71	0.71	0.29	-0.80
Animacy	51	2.33	4.83	3.29	0.63	0.50	-0.54
Likeability	51	3	5	4.05	0.56	-0.02	-0.81
Perceived Intelligence	51	2.4	5	3.98	0.65	-0.32	-0.74
Perceived Safety	51	2.33	5	3.93	0.72	-0.39	-0.54
Self-Talk	51	4	58	36.47	12.71	-0.59	0.17

Table 9.4: Descriptive statistics of all the variables measured between pre-test and post-test session

Variable	Experimental Group (n = 33)		Control Group (n = 18)	
	Pre-Test	Post-Test	Pre-Test	Post-Test
	M (SD)	M (SD)	M (SD)	M (SD)
Trust	65.58 (7.84)	74.48 (10.16)	67.78 (7.12)	76.41 (9.13)
Anthropomorphism	2.67 (0.64)	3.28 (0.66)	2.78 (0.85)	2.98 (0.77)
Animacy	3.20 (0.59)	3.76 (0.51)	3.46 (0.67)	3.51 (0.52)
Likeability	4.07 (0.56)	4.29 (0.62)	4.03 (0.58)	4.10 (0.70)
Perceived Intelligence	3.89 (0.65)	4.18 (0.60)	4.16 (0.65)	4.02 (0.64)
Perceived Safety	3.91 (0.69)	3.97 (0.60)	3.98 (0.79)	3.98 (0.60)

* p 0.05

Table 9.5: Repeated measures ANOVA and ANCOVA results

Variable	ANOVAs									ANCOVAs		
	Group			Time			Time x Group			Self-Talk		
	F(1, 48)	p	η	F(1, 48)	p	η^2	F(1, 48)	p	η^2	F(1, 48)	p	η^2
Trust	0.92	0.34	0.02	5.38*	0.03	0.10	0.01	0.94	0.00	0.19	0.66	0.00
Anthropomorphism	0.34	0.57	0.01	3.55	0.07	0.07	2.59	0.11	0.05	0.69	0.41	0.01
Animacy	0.00	0.99	0.00	1.39	0.24	0.03	5.48*	0.02	0.10	0.07	0.80	0.00
Likeability	0.53	0.47	0.01	0.01	0.95	0.00	0.63	0.43	0.01	0.20	0.66	0.00
Perceived Intelligence	0.15	0.70	0.00	0.23	0.63	0.01	4.61*	0.04	0.09	0.62	0.43	0.01
Perceived Safety	0.07	0.80	0.00	0.05	0.83	0.00	0.06	0.81	0.00	0.02	0.89	0.00

* p 0.05

Part IV

Operating Table Scenario

Chapter 10

A compromising scenario: a robot trains a nurse to set up an operating table for surgery - a preliminary exploration

A preliminary exploration was performed during the second year to investigate the possible role of the robot equipped with an inner speech in the field of medical robotics. The interest behind the study is to evaluate how inner speech influences nurse-robot interaction when both are collaboratively involved in preparing an operating table for a surgical operation.

10.1 Introduction

The combination of the goal of developing a new robotic nursing model and evaluating how inner speech improves human-robot interaction is the center around which the preliminary exploration is developed.

The focus was on developing a robotic model capable of performing one of the main tasks of an instrumented nurse, which is to prepare a servant table and a mother table, for example, interventions, following rules and conventions standard in the medical field. The second objective concerns the robot figure as an instrumented nurse instructing student nurses in an interactive lesson in which the influence of inner speech is tested.

The results of the preliminary exploration are positive. The study proposes a new robotic figure in the medical field that correctly performs routine tasks and reflects the requirement for continuous updating of the nurse instrumentalist based on a flexible and extensible ontology.

Preliminary results were achieved by subjecting a surgeon and an ordinary person from outside the medical domain to an interactive lesson on preparing a servant table for a vascular procedure, both with inner speech and without. These results were collected through an evaluative questionnaire on how inner speech improves the quality of interaction and leads more quickly to the achievement of goals due to the transparency and trust established in the human-robot relationship.

10.2 General description of the scenario under investigation

The scenario involves a robot in the guise of an instrument nurse and a novice or new nurse trainee within the operating block. The robot uses its knowledge regarding the instruments of a specific surgery, chosen from five possible ones and their positions on the servant and mother table, to conduct an interactive lesson.

The lesson begins with a presentation of the robot's role, followed by an explanation of the servant table and mother table, the stages into which surgery is divided, and the morphology, function, and type that characterize surgical instruments. Once the basics about a generic surgery are provided, the robot focuses on those specific to the surgery chosen.

At this point, inner speech comes into play: the robot dialogues with itself about its knowledge about the required surgery and thinks aloud about some of the specific instruments, those used in the initial, middle, and final stages of the surgery. It is intended to provide transparency to interaction, enabling the learner to obtain information about the surgical instruments used and to be able to construct a rationale for their placement according to the stages of use.

Once the robot returns to interaction with the learner, it displays an image of the setup of a servant table and a mother table. The learner has a few seconds to recognize the tools used and to piece together the information previously obtained from inner speech with the visual information. The robot then leaves room for the learner to prepare the tables. Each move is accompanied by feedback that is intended for encouragement in both the correct and incorrect case of positioning. In particular, in the negative

case, the feedback is followed by some remarks that the robot makes aloud, guiding the correct position of the tool; once again, the inner speech provides valuable information for the learner.

The latter has many tries until he correctly positions an instrument; this choice was dictated to facilitate learning. With incorrect positioning, the inner speech represents a resource of knowledge about the position and names of the tools. As soon as the learner places all instruments on tables, they can ask questions about the devices used and their characteristics via a special box in the application.

Interaction is constantly present since, for each topic explained, the robot asks the learner to confirm whether it is clear or not via a dialog box implemented in the application; in the case of a negative answer, it explains again. Similarly constant is the presence of inner speech: the robot, in fact, dialogues with itself, not only about its knowledge of a specific intervention or on the position of an instrument following a misplacement but also making remarks about the learner's progress during the lesson, as it receives confirmation on the clarity of the topics. It increases the learner's confidence in his robotic teacher and promotes comfortable interaction.

The proposed scenario aims to quantitatively evaluate the influence of inner speech on the learner's learning and the learner's interaction with the robot.

The two tables mentioned above represent the most critical components of an operating room; on them, the instruments used to perform surgery are placed. The servant table is set close to the surgeon and operating bed, depending on the type of surgery. All the necessary and most frequently used instruments are placed on it for carrying out a surgery. The mother table is set a little farther from the operating bed. In addition to devices similar to the servant table, which are used if there is a need for replacement, it has additional less frequently used or corded instruments, such as the electric scalpel. These are contained in the pockets of a sheet that covers the table. Before placing the tools, both tables are covered with an impermeable drape and a sterile drape to prevent any liquids that come in contact with the sterile drape from wetting the surface below and contaminating it.

The surgeries considered for the study conducted are abdominal, laparoscopic, vascular, otolaryngological and cataract surgery. The latter's choice was dictated by the purpose of using the results obtained for the preparation of instrumental nurses engaged in the most common types of interventions.

10.3 Requirements

The acquisition of information regarding instruments used and their configuration on tables was made through interviews with a surgeon and studies in medical manuals.

From the available sources, the setting up of the servant and mother table is not tied to any universal rule; instead, it depends on the surgical specialty for which the two tables are set up and on the same customs found within an operating block. Generally followed, one recommendation is to place surgical instruments such as forceps, scissors, and spatulas first, followed by preps such as wires, needles, suction, and finally gauze.

All instruments, especially bladed instruments, and needles, must be placed clearly above the drapes covering the two tables and not hidden by other tools, such as gauze, to avoid injury. The placement of the instruments is generally done in two rows. However, it may vary depending on the number and surgery. In the study conducted, two rows per surgery are considered for simplicity.

10.4 Ontology

Domain information acquired from the sources used is formalized in OWL ontology.

The choice to formalize the results in an ontology was dictated by the fact that the study conducted shares the goals underlying an ontology. These are interoperability between systems, fostering the sharing of formal representations and knowledge acquisition, and the transition from computer science, understood as automatic information processing, to epistemic, that is, automated knowledge processing.

The model is built using the developed model on inner speech and is expanded and integrated for investigating the new domain. Inner speech is then calibrated for the field at hand. Fig. 10.1 shows a fragment of the knowledge base of the robot.

10.5 Preliminary test

Of substantial importance is the direct contact with a professional surgeon, who provides critical information regarding the structure and domain knowledge and provides an initial evaluation of the proposed model itself. The surgeon then interacted directly with the robot and assessed the quality of

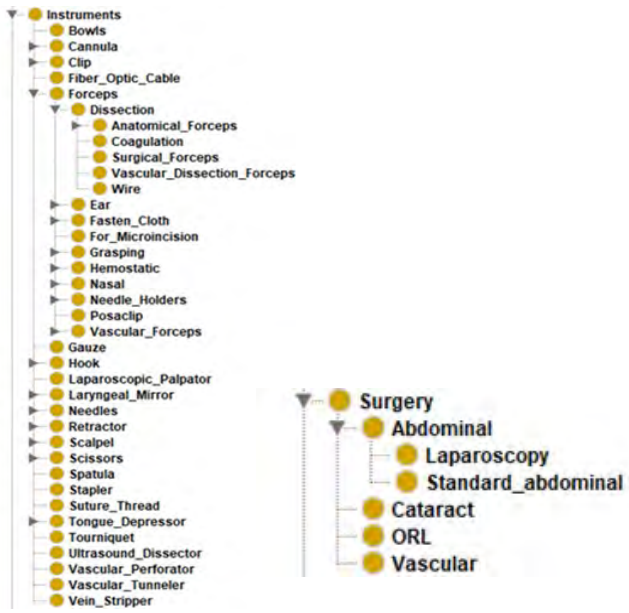


Figure 10.1: An excerpt of the knowledge base of the robot.

the training and reliability of the robot operating in the two different modes with and without inner speech (Fig. 10.2).

The preliminary study was tested by subjecting the developed platform to interactive use by the professional surgeon and, by comparison, an ordinary person naive to the domain. The choice of the two participants was dictated by the fact that a professional figure, as a surgeon, could give an objective assessment of the correctness and effectiveness of the task performed by the robot in the guise of an instrumented nurse. The figure of a person outside the domain, on the other hand, is chosen to verify how inner speech improves interaction, understanding, and execution of the required task, even in the case of a person with no background knowledge.

The task considered concerns a vascular operation. A tablet was used as the interface (Fig. 10.3). Only ten of the thirty tools to be placed provided for the vascular intervention were considered. This choice was dictated by the desire to make the test not too long-lasting, as each participant was scheduled to repeat it twice, first considering the case with inner speech and then without inner speech.

The preliminary test consisted of three phases. The first phase concerns the performance of an interactive lecture given by the robot, equipped with inner speech. The second phase consists of conducting the same interactive

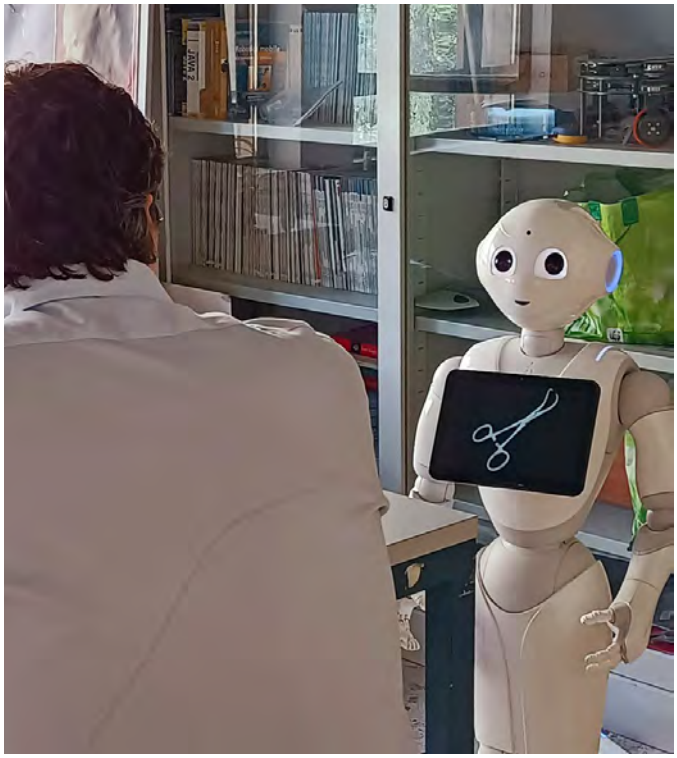


Figure 10.2: The surgeon interacting with the robot.

Figure 10.3: The virtual interface simulating the operating table on which the nurse drags and drops the medical supplies.

lesson but without using the inner speech of the robot. Finally, the third phase consists of running a questionnaire to assess how inner speech improves interaction and learning.

The results obtained from the questionnaire, completed by the two test participants, seem to confirm the hypotheses underlying the study, such as the reliability and better quality of training obtained from interacting with a thinking robot compared to a non-thinking one. In addition, new aspects are emerging, which are the subject of future improvements.

The preliminary interaction was a positive and satisfying experience for both participants. However, the effects elicited, particularly in the initial test phase, were different due to the previous experience with the humanoid robot. On the other hand, the effect produced on the naive participant, who reports some previous interactions with a humanoid robot, concerns curiosity about the new robotic figure with inner speech and general technology applied to the medical field.

In contrast, the professional surgeon, who reports never having interacted with a humanoid robot, initially feels an effect of mistrust and awkwardness. However, in the short time he performed the test, this effect gave way to a feeling of ease and reliability. This fact confirms the hypothesis that a robotic figure similar to humans can overcome humans' natural distrust of the concept of a robot.

The differences noted between the interaction with and without inner speech vary consequently to the knowledge regarding the domain. For example, a substantial difference is found by the non-domain expert participant between the interaction with the robot equipped with inner speech and without inner speech. Thus, the inner speech was a vital source of information to perform the required task for the naive participant. On the other hand, the expert participant reported minimal differences that affected the implementation aspect of inner speech. Furthermore, the expert participant focused on the minor variations in voice pitch between the two modes rather than on the information obtained by the thinking robot. Therefore, his knowledge appears to be well-grounded on the topic. However, despite the different approaches to the two interactions, both participants benefitted from the robotic figure in training an instrument nurse.

Their judgment about a robot filling a teaching role for a user who is not experienced in task solving is overwhelmingly positive. Indeed, their experience suggests that the transparency provided by a robot with inner speech can complement and expand practical and theoretical skills in task execution. Furthermore, a robotic figure can foster a new form of interaction between humans and robots. In this regard, both participants require more

language property and less latency between question and answer.

In conclusion, overall, the preliminary interaction with a robot equipped with inner speech for surgical table preparation was evaluated positively. Participants showed confidence and enthusiasm toward a robotic figure in the role of a teacher to perform a task. In addition, the inner dialogue provided transparency in the interaction with the robot, providing a feeling of ease and trustworthiness and, at the same time, helpful information for performing the required task.

However, the extension of a robotic figure with inner speech to autonomously perform tasks of extreme precision and complexity in the medical field, such as in the case of surgery, is viewed differently. The non-domain expert participant shows hesitancy toward a robot, whether thinking or non-thinking, that could replace humans within an operating room. His confidence is limited to the context of teaching or working with humans. The experienced domain participant, on the other hand, shows confidence and enthusiasm in the figure of the thinking robot for autonomously performing complex tasks in the medical field while still seeing humans as a means of limiting possible errors.

Future developments will involve the dialogic expansion between robot and nurse and the extension of the tests to a more significant number of medical participants for a more precise definition of the results obtained.

Part V

Inner Speech and Emotions

Chapter 11

Robot's inner speech and emotions

Recent studies in Robotics and AI evidenced that robots that "think out loud" during human-robot cooperation, induce positive feedback in the human counterpart, improving the goals achievement. By externalizing its inner voice, the robot becomes more transparent and explains the underlying decision processes. Moreover, the robot evaluates alternatives for solving the common task, making the interaction more robust. The objective of this work is to analyze the role of the robot's inner speech in emotions.

For the appraisal theories, the emotions emerge by the cognitive evaluations of the situation. The inner dialogue simulates the internal reflection which enables that evaluation. By inner speech, the robot focuses on the relevant facts of the context, acquiring the needed information for computing the appraisal variables, and finally the emergent emotion. In the meanwhile, the partner can hear the underlying processes generating the emotions, and knows the motivations of the robot's emotional state. The performances of the model are in line to the typical emotional trends of healthy adults when facing stressful situations, showing that in the same situations, the robot's emotive reactions are the expected ones, improving the results provided by a well-known computational model of emotions.

11.1 Introduction

Philosophers, psychologists, and neuroscientists have long studied the roles of the inner voice in cognitive and psychological human functions, including affectivity and passions [90] [126] [185]. The inner voice is the linguistic

surface form of the thoughts. A person is engaged in the self-dialogue when he/she thinks by running verbal commentary [168]. This experience helps people to focus on the context, to plan the actions, to keep decisions, and to become conscious of facts and events.

A link between inner speech and emotions was first oulined by Vygotsky [231], that argued about the continue and dynamic interactions between the intellect, meant as thoughts, and the affective sphere. Such a link evolves and fluctuates in the course of life, and is bi-directional: *from the affective sphere of consciousness to thought and from thought to the affective sphere of consciousness* [73].

The same perspective is supported by Lazarus [129] [128] and, in general, by the cognitive appraisal theorists: the mental thoughts are fundamental in affectivity because thinking must occur first in the sequence of the human cognitive processes that lead to the emotional experience. More specifically, the sequence starts when a stimulus urges the person, followed by an emergent thought in linguistic form associated to that stimulus, and ends with the experience of a physiological response or emotion. Thus, the role of thinking in feeling emotions is fundamental.

When modeling the emotional behaviors in artifacts, as robots and artificial agents, the appraisal theories are frequently referred to, basing the emotion evaluation and its intensity on the appraisal variables. The appraisal variables are links between the context and the emotional components, and they are well suited for computation. Despite existing contributions based on these theories produced interesting results in the affective computing field, the role of the inner dialogue has not yet been investigated as one of the fundamental processes in computing the appraisal variables, and hence in forming artificial emotions.

This work proposes a new computational model of the tight link between emotions and inner speech. Some of the authors yet proposed a cognitive architecture of inner speech [40] [42] and deployed it on a real robot, building the first robot that "thinks out loud". The benefits of such a skill, when the human and the robot collaborate for reaching a common goal, regarded the robot transparency (i.e. the possibility to trace and reproduce the underlying decision processes), robustness (i.e., the possibility to go out from stalemate situations, and to end the task successfully) [188] and trustworthiness [81] (i.e. the human is inclined to trust the robot more).

The proposed model of inner speech and emotions finds inspiration from the modal model by Gross [91], which provides a structure highlighting the steps involved in the emotional self-regulation cognitive process, that are Situation, Attention, Appraisal and Response.

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

Once a stimulus becomes a thought (i.e., a textual surface form is associated to the stimulus), the inner dialogue starts and enables the evaluation of further facts and events (that could be external, i.e. regarding the environment, or internal, i.e. regarding the inner state), simulating the cognitive evaluation of the context. The inner dialogue is a set of turns, and a new turn emerges in response to the previous one, leading to a rehearsal loop. The loop is repeated until no further facts to evaluate emerge, or all the needed information are inferred. At this point, the model computes the appraisal variables (the Appraisal block), that are specifically formalized so that the trends of such variables follow the observed trends in healthy adults [184]. Then, the corresponding emotion to the appraisal variables, with a specific intensity, is evaluated by referring to the Russel's Circumplex Model of Emotion [197] (more popularly known as the Circumplex Model of Affect), and the emotion is elicited (the Response block).

Specifically, the proposed work concerns the following two main issues: (i) the use of rehearsal loop, i.e. the inner speech, for the cognitive evaluation of the context, leading to the collection of the needed information to appraise the situation; (ii) the definition of the mathematical formulas, based on the information inferred by the inner speech, that model the appraisal variables and enable the computation of final emotion with a specific intensity.

During all the processes, the robot thinks out loud, and it is possible to hear its reasoning in emerging its emotional state in respect to the current context. The verbose description of the processes are hand-annotated and instantiated during the execution of the processes basing on the involved concepts.

The experiments for validating the model consist of plunging the robot in simulated stressful situations, and observing if the emotional behavior is in line to the trends presented at [86]. The trends measure the appraisal variables and the emotional reactions of healthy adults when facing stressful scenarios, regarding loss and aversive situations. The results are promising, and the mathematical formalization of the appraisal variables compute the expected values for them and emotions, according to the trends. Moreover, the results are compared with those provided by a well-known computa-

tional model of emotions, that is EMA [143], in same scenario, showing improvements in many cases.

The paper is organized as follow: the theoretical basis of the model, including the appraisal foundations, the Gross' model, and the role of inner speech in feeling emotions, are presented in the section 11.2. The section 12.4 describes the whole proposed model linking inner speech and emotions. In particular, this section details the inner speech process for cognitive evaluating the context, and the proposed mathematical formalization of the appraisal variables. Moreover, given the values of the appraisal variables, the section shows how the emotion related to the context emerges with a specific intensity. The methodology for the comparative evaluation of the model, and the obtained results are discussed in the section 11.4. To show how the model works, a use case related to a collaborative task, involving the robot and a human partner, is detailed in section 11.5. Finally, the state of art is presented in the section 11.6, and the conclusions and the possible future works are proposed in the section 11.7.

11.2 Theoretical background

11.2.1 The appraisal theories

The appraisal theories [127] [44] [76] claim that in life situations, each person enters in relationship with specific aspects of the context, which are relevant for him/her, enabling the subjective interpretation of the situation. The appraisal theorists consider the elaboration of these relationships responsible of emotions. Each emotion arises from the *cognitive evaluation* of the situation, and from the corresponding *meaning structure* resulting from that evaluation. The cognitive evaluation is the process starting with the perception, and consists of the automatic assessment (generally involuntary) on the presence or absence of a specific entity/event and its positive/negative effects, depending on the subjective meaning, on the final emotion. During evaluation, a structure of the emergent meanings is built and it keeps trace of the components that contribute to the final emotion, that is the meaning structure.

The empirical cognitive evaluation involves the *appraisal variables*. The appraisal theories differ among themselves depending on the definition of these variables, and on the understanding how they could affect the corresponding dominant emotion. Comparison between, and convergence of these theories is difficult. However, they have e common thread, that is the bottom-up approach in evaluating appraisal variables: each emotion is

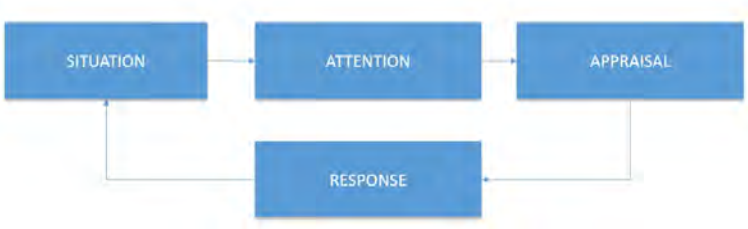


Figure 11.1: The modal model of emotion regulation proposed by Gross

elicited by its own specific and distinct pattern of appraisal variables, which are computed from the low-level (i.e. the context). Appraisal theories tend to dominate among computational models of emotion due both to its emphasis on emotions as computable artifacts, and to its simplicity of application.

In the most classical appraisal theory [108] [205], the emotions are meant as discrete entities, organized in a taxonomy. An emotion belongs to one class of the taxonomy, basing on its properties and features. However, the identification of closed classes leads ambiguity, because the same emotion could arise from different events, and it can have different meanings which are not strictly linked to its specific features. Recent theories [206] are less rigid than the emotional categorization. They base on the assumption that the human being does not perceive emotions as discrete entities but they can take different gradations. As consequence, a formal definition how to compute appraisal variables is not always possible, but depends on the individual profile, with his/her baggage of personality and experiences.

Basing to this perspective, the proposed model simulates the robot's cognitive evaluation by inner speech, keeping in account the robot's specificity and characteristics. The efforts included the identification of the main features of the robot that could affect such an evaluation. Examples are the inner conditions of robot's functioning, including the states of its battery, the work states of the joints, the levels of the temperature, or the levels of disorder of the environment that could influence the robot's perception and its ability to discriminate signals. All these aspects have to contribute to the appraisal of the context and to the final robot's emotional experience.

11.2.2 The modal model of emotions

Gross [91] defined emotions as brief responses affecting both behavior and body, and they emerge during events with potential challenges or opportunities. He believed that emotions can be modulated, leading to the emo-

tional self-regulation process. Gross modeled the process by the *modal model* shown in figure 11.1, that is a sequence of the following steps:

1. **Situation**: the sequence begins with a situation (real or imagined) that is emotionally relevant for the individual;

2. **Attention**: the individual focuses towards the emotional situation and evaluates the related facts he/she retains useful;

3. **Appraisal**: the emotional situation is interpreted and the emotion emerges with a specific intensity;

4. **Response**: an emotional response is generated, that regards changes in experiential, behavioral, and physiological systems.

Gross took this set of steps to be very broad. He included automatic, controlled, conscious, and unconscious processes. This foundation define the backbone architecture of the proposed model, and the defined steps are maintained in the cognitive sequence for appraising the context.

11.2.3 The function of self-talking in feeling emotions

In *Theory of Emotions*, Vygotsky [231] conceived the *interfunctional theory of emotions* according to which he empathized the importance of conceiving the soul strictly related to the bodily manifestation, and not just the body as the main source of experience and emotions. He highlighted the fundamental role of the dynamic and dialectical interconnection between mental life and body, which cannot fail to influence the psychological experiences. According to his theory, the words are not mechanisms of expression of thoughts, but the places where thoughts end. The thoughts are the mediating tool of the experience of the self and of the context, finally elicited by words. When experiencing an emotion, the though provides the means for living this experience, and then it is materialized (covertly or overtly) by the words.

The causes and effects of the interconnection change permanently in development, leading to continue interactions between the intellect and emotions. This link is bi-directional: *the word nominates the affection, and the affection, therefore, is channeled in thought through the word.* [73] The existence of that link is supported by Morin [169], that discovered in his experiments that common contents of inner speech in people were self-addressed evaluations and emotional states. By self-talking, it is possible to live and

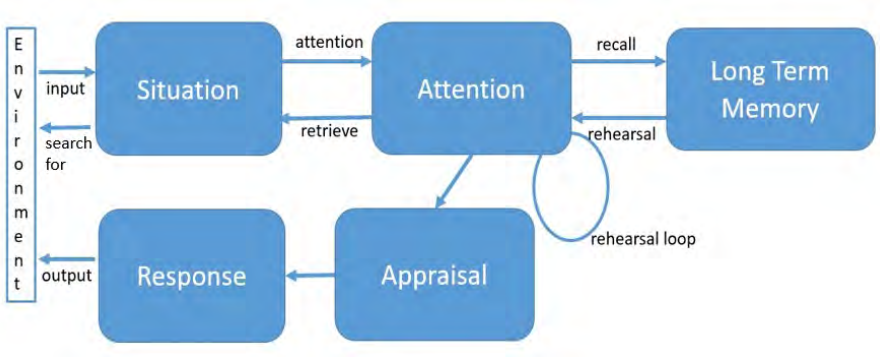


Figure 11.2: The proposed cognitive architecture of inner speech and emotions.

to become conscious of the emotional experience, to self-regulate and to act opportunely.

The proposed work concerns to the computational model of interconnection between emotions and inner speech. By self-talking, the robot makes experience of the situation. The linguistic reasoning is the mediating tool of such an experience, according to the Vygotsky's theory, finally leading to the cognitive evaluation of the situation. When the inner dialogue ends, the evaluated appraisal pattern enables the emotional experience. In turn, the emotional experience involves the emergence of a thought through which the robot externalizes its emotion and becomes aware of such a feeling.

11.3 Modeling emotions in robots by inner speech

Figure 11.2 shows the proposed cognitive architecture of inner speech and emotions. As mentioned above, the backbone structure is the modal model, integrated by a rehearsal loop for inner speech. Moreover, a further block, modeling the robot's memory, contains the robot's knowledge and is used for the retrieval/analysis of further concepts correlated to the perceived one.

Summarily, the architecture enables the robot to the emotional response when a stimulus occurs from the environment. The Situation block encodes the stimulus, by associating a symbolic form to the perceived signal. The Attention block recalls from the long term memory the correlated concepts, in linguistic form: it represents the emergence of the thought corresponding to the stimulus, which triggers inner speech. The thought is re-heard by the Attention block, and it could require to retrieve further stimulus from the

environment by the Situation block, which will search for them, or further concepts from the memory. The retrieve/recall from environment/memory depends on the concepts in the turn, and on their possible matching with new concepts in the environment/memory. The monologue enables the cognitive evaluation and provides the parameters to the Appraisal block which computes the appraisal pattern and the corresponding emotion to that stimulus. The emotion is then externalized by the Response block. All the processes related to the cognitive evaluation by inner speech are detailed in the subsection 11.3.2.

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

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

For each of these variables, the model proposes a mathematical formalization, that is defined in the subsection 11.3.3.

The specific appraisal variables depend on the specific scenario in which the robot is plunged, and regard specific aspects that could affect its emotional experience. For example, in an application scenario in which the robot and the partner collaborate to set up a table according to an etiquette schema, a specific appraisal variable could regard if the partner places the utensils in compliance to the etiquette or not. A situation in which the partner does not follow the etiquette, generates more unpleasant emotion than the situation in which he/she acts correctly, and the specific variable contributes in that sense to the final emotion. By considering it is impossible to predict each possible specific appraisal variable, they can add in next phase in the model, by specifying the kind of contribution (if positive or negative),

as next describing in more detail when showing an application case in the section 11.5.

Once the appraisal variables are computed, the corresponding emotion is inferred by applying the circomplex model of emotions [197], according to which the emotions result from the combinations of two dimensions, that are the *valence*, which explains how pleasant/unpleasant an emotion is, and the *arousal* that explains the level of the physiological activation. The emotions are thus represented in a two-dimensional space. An emotion is a point in that space. The circumplex model includes 28 emotions, but in the proposed model only the 5 basic emotions by Ekman [66] are included, that are *happiness*, *sadness*, *fear*, *anger* and *disgust*. How valence and arousal, and the corresponding emotion emerge, by linking them to the appraisal variables is detailed in the subsection 11.3.3. Moreover, by depending on the position of the emotion in the space, it is possible to associate an intensity to it.

11.3.1 The knowledge model

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

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

For example, the class `Person` represents a human, that is a concept of the external world, and an instance of such a class will model an individual perceived in the actual context; the `emo` and `age` attributes are examples of

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

datatype properties of the class `Person`, representing the emotional state of the individual and his/her age respectively. Instead, a set of object properties could model the individual's position to respect another perceived entity, that are the `left`, `right`, `up` and `down` object properties, linking the individual to the specific entity if he/she is at the left, right, up or down of that entity.

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

11.3.2 The inner speech for cognitive evaluating the context

The central idea of the proposed architecture is the *inner voice* of the robot by which it focuses its attention to the concepts of situation which could affect its emotional state. The robot "internally reasons" about a stimulus, and it focuses on the concepts that are related to that stimulus; the reasoning is symbolic, involving the linguistic surface form of the concepts. It is in line with the claim *the words allow thinking about the emotions, and the emotions generate further words, or thoughts of the inner dialogue.* [73]

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

In the proposed model, the meaning structure is formalized by the couple $T = [sem(.), syn(.)]$ which associates the semantics $sem()$ with the syntax $syn()$ of a word or set of words. The elements in the $sem()$ and $syn()$ parts are named *chunks*. The $sem()$ part of the structure contains *meanings*, that are chunks corresponding to the concepts in the knowledge base. The $syn()$ part contains *words*, that are chunks corresponding to the tokens of the linguistic form. The meaning structure grows during the process by adding meanings and words. Each new meaning and each new word is merged in the corresponding part if it is not already in the structure.

The mechanism is shown in figure 11.3, and it is inspired to [217]. The rehearsal loop consists of three steps, that are:

1. *Conceptualisation* A percept p is the symbolic form of a perceived stimulus which could be a voice stream, the image of an object, one or more features of the object, and so on. In any cases, the form of the percept is textual and syntactically describes the perceived stimulus (i.e., it is the text produced by the speech recognition routines,

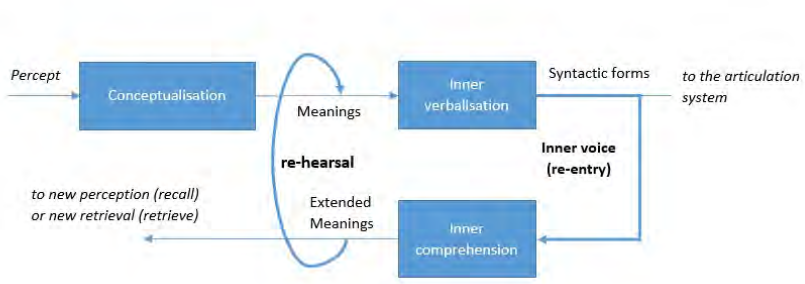


Figure 11.3: The language re-entrance components: the syntactic forms from inner verbalisation component are inputted to the inner comprehension one; further expansion of meanings allows to reason about beliefs and internal state thus identifying emotional relevant situation and defining attention.

the text describing the object or its features by the image processing routines, and so on). The step finds correspondences between the syntax of the percept and the knowledge base of the robot. Being $S = \{s_1,\ s_2, ...,\ s_n\}$ the set of tokens of the percept p, pre-processed by removing not meaningful words (as articles, conjunctions, prepositions, and so on), the meaning structure is filled by the chunks of the percept, so that $T = [sem(),\ syn(s_1, s_2, . . ., s_n)]$.

To conceptualize the percept means *to find the concepts in the KB ontology that better match to the syntax of the percept*: the matched concepts in the ontology are the classes that conceptualize the percept. The match is implemented by computing the Jaro-Winkler distance [110] between the elements in S and the labels of the ontology. This distance is a measure about how syntactically similar two words are. Being $d_{jw}(w_1, w_2)$ the function that returns the Jaro-Winkler distance between the words w_1 and w_2, and $couple(r)$ the function that returns the domain and the range of a property r, the match between S and the ontology O is modeled by the following functions:

- $a_c : S \rightarrow C_o$ that returns the set $a_c(s_i) = \{\ cl_k \mid d_{jw}(s_i,\ cl_k) > \alpha,\ cl_k \in C_o\}$ composed by the ontological classes such that the Jaro-Winkler distance between their labels and the words contained in S is higher than a threshold value α (the words are syntactically similar to the classes).
- $a_p : S \rightarrow P_o \cup P_d$ returning the set $a_p(s_i) = \{\ r_j \mid r_j \in P_o \cup P_d\ \wedge$

$couple(r_j) \supset a_c(s_i)\}$ composed by the properties in the ontology for which the previous retrieved concepts are the domain or the range of these properties.

The value of α was set to 0.2 after parameter tuning.
The result is a sub-ontology $O_m = \langle C_m, P_m \rangle$ where $C_m = \bigcup_S a_c(s_i)$, $P_m = \bigcup_S a_p(s_i)$. The O_m ontology will include all the concepts and the relations corresponding to the percept. The elements in O_m conceptualize the percept. The meaning structure becomes

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

with $m_i \in C_m \cup P_m$.

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

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

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

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

begin $label(c)$ the function returning the label of the class c in the ontology.

The labels are verbalized, and back-propagated to the inner comprehension component, i.e. they are re-heard. The important aspect is that the production could include additional syntactic tokens related to the meanings that the conceptualisation step has merged to the structure, and that could involve a different set of concepts than the tokens of the initial percept.

3. *Inner comprehension* The new labels emerging from the previous verbalization represent new thoughts that could require other perception

or other concepts from the knowledge base. The inner comprehension searches for the added labels in the meaning structure in the environment (retrieve) or in the knowledge base (recall). In particular, there are two cases:

(a) The meaning is related to a property (i.e., $m_i \in P_m$). In this case, the module searches in the knowledge base for the other property's domain or range not contained in the meaning structure (recall), and adds it to the $sem()$ part of the structure, if it is not yet included;

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

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

Table 11.1 shows a use case of the described method. For each involved phase of the loop, the table reports the meaning structure, and the process that modified such a structure. In that example, the robot receives a command by the partner, that is '*take the plate*', and that percept is processed to fill the $syn()$ part of the initial meaning structure. For each phase, the process column shows the concepts and the properties emerging from the knowledge base (recall), or from the environment (retrieve). These entities are highlighted in bold. The production of the sentence (produce) represents the inner verbalization. It is simple to observe how the new entities are added to the meaning structure, and how the structure grows in each next phase.

Table 11.1: How inner speech works for cognitive evaluation. For each step of the loop, the table reports the meaning structure, the phase of the loop and the related process among recall, retrieve and produce. The loop phases are C for the Conceptualization step, IV for the Inner Verbalization step, and IC for the Inner Comprehension step. The robot perceives the audio stream *take the plate* by the partner. The meaning structure initially includes the chunks related to the meaningful tokens of the codified stream. The loop recalls from the ontology the related concepts and adds them to the structure, and the inner speech starts. All the emergent concepts in each phase are in bold.

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

11.3.3 The appraisal variables and emotions

The meaning structure is the output of the cognitive evaluation the robot performs by inner speech. Once the loop ends, the values of the appraisal variables can be computed. The model proposes a new mathematical formalization of these variables, based on the trends they should have, as described at [184]. The formulas were tested and tuned for the best performances of the emotional experiences, according to the expected trends described in the aforementioned work.

The formalization of the appraisal variables is based on the particular environmental conditions because, when involving a robot, these conditions can heavily influence its whole evaluation of the context. The model defines the *environmental entropy* k, meant as an indicator about the lack of order in the environment. Higher the entropy is, higher the environmental disorder is. To evaluate the entropy, the model keeps in account the environmental noise (that could make the possibility to recognize the sound stimulus difficult), and the emotional state of the partner (that could reveal him/her approval rating, making the situation more unpredictable). The robot's routines detect these specified features, and return the values to use in the entropy formalization. Given e the set of the possible partner's emotions detectable by the robot's routines, they are classified in *negative*, *neutral* and *positive* emotions. The entropy depends on the following parameters: α that is the level of environmental noise, β that measures the partner's emotion inferred by his/her facial expression, and γ that measures the partner's emotion inferred by his/her vocal tone. The parameters are formalized as follow:

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

and

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

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

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

Table 11.2: 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	Good	10%
	Low	90%

The entropy is next used in the formula of the appraisal variables, for making them dependent on the environmental conditions.

Likelihood L

The likelihood L defines the probability that the negative outcome of an event happens, leading to stressful situations [184]. A variation of this value is a sign about the evolution of the event. When it grows, the negative outcome becomes more possible, and the stress grows. On the contrary, when the likelihood decreases, the negative event is resolving, and the stress melts away.

In the proposed model, the likelihood is hand encoded for the particular work conditions, because they influence the success/failure of the action to take. The considered work conditions and the corresponding likelihood values are shown in table 11.2. The work conditions include the robot's physical conditions, the valuations about the feasibility of a required action (i.e., is the involved object in the action reachable or visible?), and the admissibility of the action (i.e., does the action follow the rules?). The likelihood is higher when it corresponds to negative work conditions, because they affect the outcome's feasibility. For example, when a physical component of the robot is not properly working, the likelihood of an event involving that component is fixed to 80% because the probability the event fails is high. If, for some reasons, another component is chosen to work, or the problem of the same component is resolved, the value decreases to 20%. When more than one work condition occur simultaneously, the final likelihood is the arithmetic mean of the likelihood values of each occurred condition.

Controllability C

The controllability C measures if the outcome of an event can be modified by directly acting on the context [184]. For formalizing C, the model considers that higher the likelihood is (i.e., higher the probability of a negative outcome is), lower the possibility to modify the context and to resolve the negative evolution is. Similarly, higher the environmental disorder is, lower the possibility to control the situation is.

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

$$C = -\frac{1}{|kx|} + x^2$$

begin $x \in]0, 1]$ the absolute value of the variation of the likelihood in two consecutive measurements, that is $x = |L_a - L_p|$, with $L_a \neq L_p$. L_a is the value of the likelihood in the antecedent measurement, and L_p is the value of the likelihood in the present measurement. L_a is set to zero when the measurement is the first. When there is not a likelihood variation, it means that the context does not change, and the controllability remains to the last computed value.

Changeability M

The changeability M measures if the outcome of an event can be modified by another external event, for which the control does not depend by the individual [184]. It represents a measure about how unpredictable the context is, because the outcome does not depend by the self. According to this definition, it is trivial to consider that higher the probability of a negative outcome is, higher the possibility that the situation becomes more unstable is. Similarly, higher the entropy is, more unpredictable the situation is. As consequence, M is formalized as:

$$M = x \star k$$

begin x defined as the previous case, that is $x = |L_a - L_p|$.

Desirability D

The desirability D represents if an outcome of an event is desirable or not. It can be positive or negative [145], according to the expected outcome.

Simply:

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

Matching the appraisal variables to the emotions

Once the general appraisal variables are computed, they are combined for inferring the corresponding emotion. For this purpose, the model refers to the bi-dimensional Russell's space [197], in which the emotions are points, whose coordinates are the *valence* v and the *arousal* a, each falling in the range $[-1, 1]$. The valence means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event. The arousal represents the physiological activation when facing an emotional experience. According to these definitions, the proposed model uses the values of the appraisal variables for computing the valence, because the modeled appraisal variables keeps in account the conditions of the situation and hence its attractiveness/aversiveness. Instead, the model computes the arousal by referring to the inner states of the robot, representing the robot's physiology.

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

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

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

where C, M, and D are the controllability, changeability and desirability appraisal variables, and $N(.)$ is the function that returns the normalized value in the range $[-1, 1]$ of its argument.

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

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

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

The model infers the emotion by projecting the *appraisal pattern* $\mathbf{a} = (v, a)$ in the Russell's space. Being E the set of the labels of the Ekman's emotions (i.e., $E = \{$*Angry, Happy, Afraid, Annoyed, Sad*$\}$), and being R

Table 11.3: The matching of the values of the valence v and the arousal a with the labels *l(v,a)* of the five basic emotions by Ekman in the Russell's space .

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

the set of coordinates in the Russell's space corresponding to each emotion in E, the element $\mathbf{r}_i \in R$, with $\mathbf{r}_i = (v_i, a_i)$, is the couple of coordinates corresponding to the i-th emotion in E, as shown in table 11.3. The coordinates are those defined at [182]. The function $l(\mathbf{r}_i)$ returns the label of the emotion in E corresponding to the point in R passed as argument. The emotion $e_{\mathbf{a}}$ matched to an instance of the appraisal pattern $\mathbf{a}$ is the following:

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

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

The computation of the emotion's intensity

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

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

In the model, the level on an emergent emotion depends on how close the emotion is to the point in the Russell's space. More the coordinates of the valence and arousal computed for the context are closed to the point of an emotion (the points shown in the table 11.3), more intense the emotion is.

In the proposed model, the intensity of the emotion $\mathbf{e}_a$ of the appraisal pattern $\mathbf{a}$ is computed by referring to two threshold values, that are the

high threshold t_h and the low threshold t_l. They are fixed respectively to the quarter and to the half of the Euclidean distance $\| \cdot \|$ in the Russell's space between the point of the emergent emotion and the origin $\mathbf{o} = (0, 0)$ of the space. More formally:

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

with $\mathbf{r}_i \in R$, and the corresponding intensity will be:

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

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

11.4 Evaluations

Given the high subjectivity of feeling emotions, it is not trivial to validate and test a model implementing the emotional behavior. Moreover, the processes related to inner speech and emotions are not directly accessible and observable in humans, and to standardize the emotional responses remains (and probability it will remain) an open issue. Anyway, the authors at [86] proposed an evaluation strategy consisting of directly comparing the appraisal variables of their model to human data, that are collected by the Stress and Coping Process Questionnaire (SCPQ test) [184], that is an instrument for abstracting a general emotional human behavior in stereotypical stressful situations.

The proposed model is valuated by the same strategy, and it leads to double benefit: the results are comparable to the trend of human's emotional behavior when the robot operates in the same context, and also they are comparable to the results provided by the same authors for their computational model. Thus, the resulting evaluation involves the comparisons with human's behavior, and with one of the well-known computational model of emotions, that is EMA [145].

11.4.1 The SCPQ test

The SCPQ test is a clinical instrument consisting of several narrative episodes and related questions for abstracting the trends of emotional behaviours in normal, healthy, adult human beings when facing the situations described in the episodes.

The questionnaires are administered to each participant after the description of the stereotypical stressful episode, and the participant has to imagine to live that situation by identifying himself/herself with the subject of the episode. The proposed questions regard different aspects (as the emotional response, the appraisal variables, and the adopted coping strategies), and the participant analyzes the episode many times, for keeping in account one of such aspects from time to time.

SCPQ provides four prototypical situations, falling in *aversive* situations and the *loss* situations. In the aversive situations, some bad outcome has occurred but there is some potential to fix it. In the loss situations, a potential loss could occur in the future. Both kinds of situations can have a positive and negative resolution, defining the four prototypical cases, that are:

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

The narrative of the episodes is based on a canonical structure, that models the time evolution of the situation. The time is discrete, and each time corresponds to a specific phase of the episode, that are:

Phase 1: an initial situation is described and something could occur (among aversive or loss);

Phase 2: nothing happens, and the context does not change for some time;

Phase 3: something happens, and the situation is resolved in good or bad way.

The analyzed questionnaires enable to abstract general emotional behaviors, and the trends the evaluation refers to are:

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.

11.4.2 Methodology

The evaluation strategy [86] models the evolution of the episode by varying the likelihood, and abstracts the SCPQ episodes by a general grammar. The grammar encodes the underlying prototypical stressful situation, and enables the classification of each episode in one of the four canonical situations. According to such a grammar, all episodes involve a goal, and the likelihood of goal attainment drops in phase two, reaching zero (one) in the phase three depending on the bad (good) outcome. In particular, for the aversive condition, in phase one, the probability that the future action has chance of succeeding is set to 66%, this drops to 33% in phase two because, more time passes without anything happening, the possibility to fix the situation decreases. In phase three the likelihood is set to either zero or 100% percent, depending on if the bad (the action fails) or good (the action successes) outcome is modeled.

In the loss condition, in the phase one, the chance of the loss succeeding is initially 50%, raises to 75% in phase two because, more time passes without anything happening, the loss becomes more probable, and varying to 100% or zero, depending on if the bad or good outcome is modeled.

To put the robot in the same stressful situations, the likelihood and its variations over the phases are set to the aforementioned values, and the inner speech will infer them. The other inner appraisal variables are next observed depending on these values, and they are compared to the SPCQ trends and EMA results.

11.4.3 Comparative results and discussions

The figure 11.5 shows the comparative evaluation of the proposed model with the SCPQ trends and the EMA results. In particular, the figure reports the aforementioned SCPQ trends, and how/if the EMA model and the proposed one follow them.

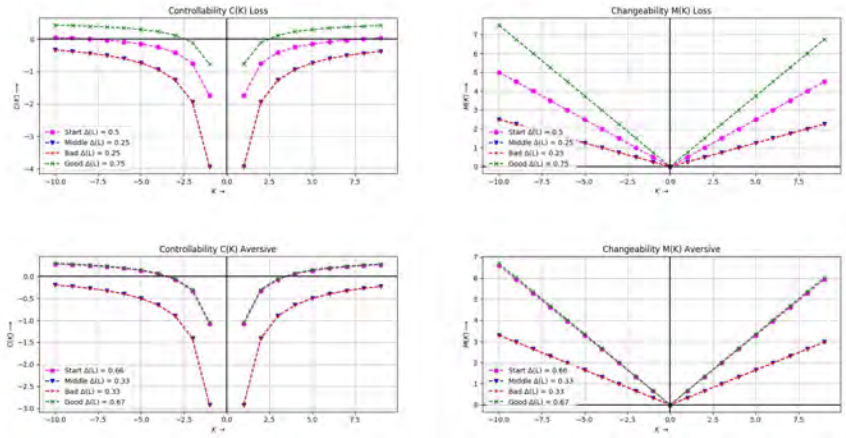


Figure 11.4: The variation of controllability and changeability in correspondence of different entropy values. By fixing the likelihood, the controllability and the changeability vary maintaining the same trend. That is, in a same canonical episode, the appraisal variables follow the same trends under different environmental conditions.

For the reported charts of the proposed model, the entropy k was set to 1. This because the k value does not influence the global trends. In fact, as shown in figure 11.4, the appraisal variables maintain the same reciprocal trends even if k varies. That is, when k changes, the trends of the controllability and changeability remain the same with different x values, that is, in a same canonical episode. As consequence, the following observation about the trends does not depend on the different k values.

By analyzing the charts in figure 11.5, the following observations emerge. The trend *a.* is respected because in the aversive conditions, the controllability and changeability are higher than the loss conditions.

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

The trend *c.* is respected because the negative valence grows over the phases, and the charts show this trend in the bad evolution of both aversive and loss conditions. Moreover, the trend follows more faithfully the SCPQ trend than EMA in the loss situation.

The trend *d.* is respected because the positive valence (good outcome)

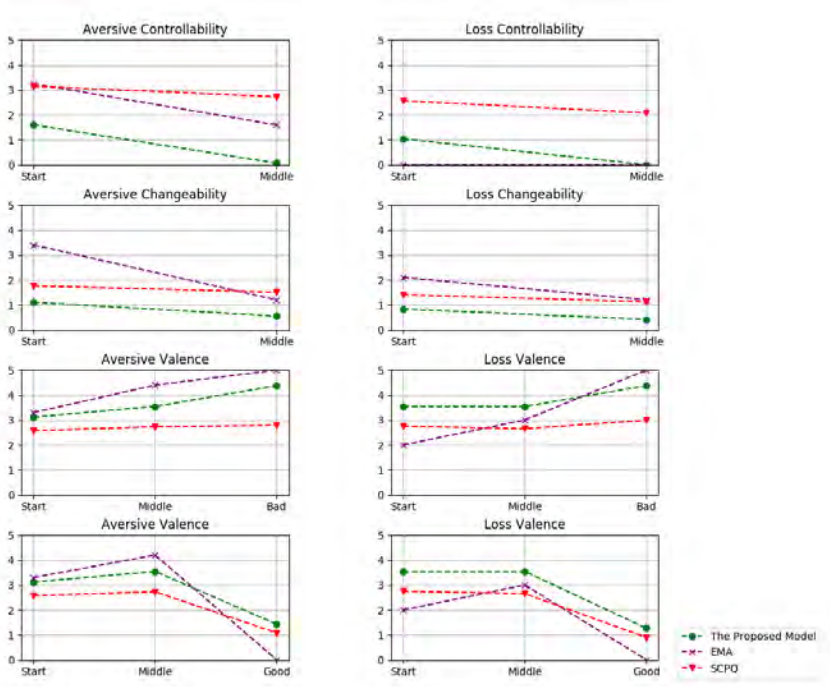


Figure 11.5: The comparative evaluation of the appraisal variables with EMA and the SCPQ trends related to the four canonical stressful situations.

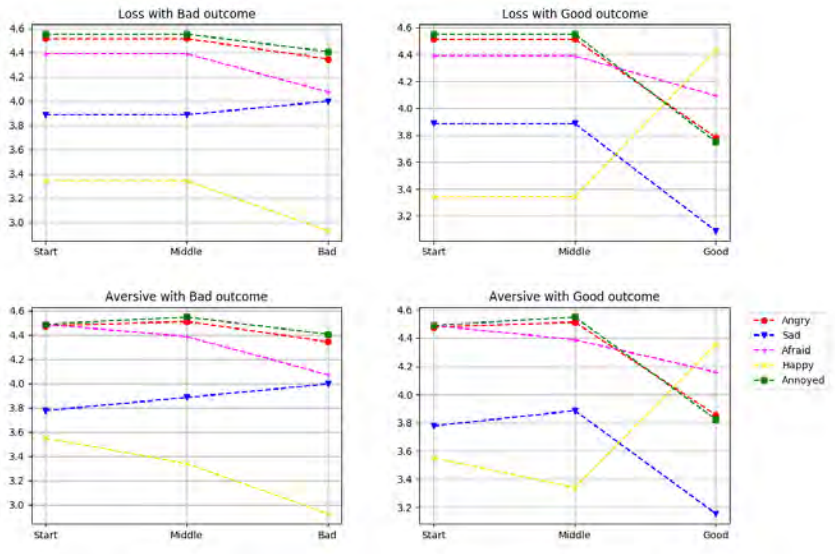


Figure 11.6: The resulting emotions of the proposed model in the four SCPQ canonical episodes. The emotions are the expected ones according to the SCPQ trends.

and the negative valence (bad outcome) have strongly different values. In this case, the trend of the proposed model follows more faithfully the SCPQ trend than EMA.

The trend *e.* is respected because, as shown in figure 11.6, the aversive condition leads to more angry and less sadness.

Moreover, the figure 11.6 shows that in good outcome, the positive emotions emerge on the negative ones, which are predominant in each initial phase, because the simulated situations are stressful. Instead, when the situations end with a negative outcome, the predominant emotions remain negative.

11.5 An application: setting up the table with a human partner

The robot and a human partner have to set up a table according to the etiquette rules. The scenario is the same analyzed by some of the authors at [188], where first interesting results, about the robot's inner speech effects on human-robot cooperation, are discussed. With the aim to show how the

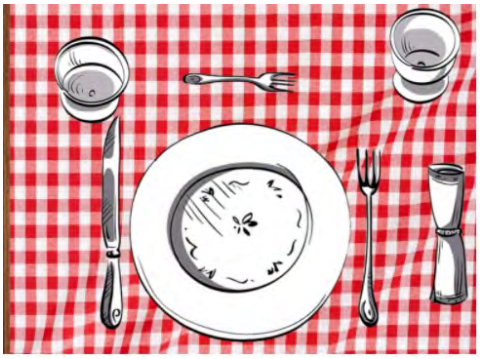


Figure 11.7: The etiquette schema to follow for setting up the table. The human partner and the robot have to place the utensils according to the schema.

proposed model works, the same scenario is now investigated.

The task consists of placing the utensils in the table in specific positions related to the etiquette schema, as shown in figure 11.7. The schema establishes where the utensils have to be positioned in the table according to etiquette, and it represents a set of rules to follow for correctly operating. The possible actions by the human can be to place an object (one by one) and to request to place an object to the robot. The possible actions by the robot are those required by the partner. The actions the partner can ask the robot could infringe the etiquette rules, that is the required final location can be wrong. As general rule, each utensil can be moved only twice. So, there are just two tentative for correctly placing each utensil.

11.5.1 The specific appraisal variables in the table scenario

As discussed above, some specific appraisal variables can be added with the aim to consider the emotional contributions of specific aspects of the context. For example, in this scenario the following aspects should be kept in account:

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

These aspects affect the global emotional experience, and they should be modeled as specific appraisal variables that contribute to the final emotion with the general ones. The error by the partner has a negative valence, more negative further away from the correct position the utensil is, while the possibility to fix the situation, that is to have another tentative, contributes positively. In the same way, when the partner requests for a correct/wrong action, the contribution to the valence is positive/negative. Thus, the specific appraisal variables are:

- *Gradient variable g*: it measures how the wrong action by the partner contributes to the valence. Further away is the moved utensil from the correct location, more negative the contribution to the valence is.
- *Recovery variable r*: it represents the number of residual tentative for fixing a wrong action. An existing tentative contributes positively to the valence, no further tentative means that the situation is irreversible, and the valence decreases.

Regarding the request by the partner to the robot of taking correct/wrong action, the contribution to the valence is from the desirability, that is one of the general appraisal variable. The robot considers desirable/not desirable to take a correct/wrong action.

The gradient variable

The contribution of the wrong action by the partner to the valence, is formalized by evaluating the percentage of error. It measures how wrong the partner is in respect to the maximum possible error, that is the maximum distance between two locations in the etiquette schema (i.e., the distance between the two furthest locations). The gradient variable is based on that percentage of error, and is the following function:

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

where ρ is the maximum possible Euclidean distance the utensils could be positioned, and Δ is the Euclidean distance between the final position of the utensil as placed by the partner, and the expected correct position. The function models that furthest the utensil is placed from the correct position, lower the gradient is. When Δ is zero, the utensils is correctly positioned, and the gradient returns the maximum value G_{max}, that is set in the tuning phase to 13. Higher the gradient is, higher the valence should be.

The recovery variable

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

The contributions to the valence

Defined the specific appraisal variables, and the kind of contribution to the global valence (positive or negative), the valence v has to be update by including the specific appraisal variables. As consequence, in the proposed scenario, according to the sense given to each specific variable, the resulted global valence is:

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

11.5.2 The model at work

Two different use cases, with different conditions of the context and different events, are presented.

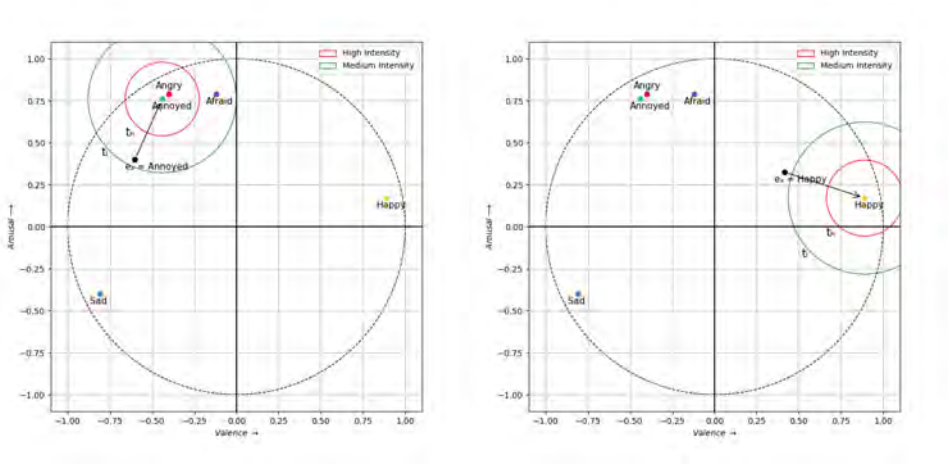
> *Use case I.* The partner asks the robot to take a wrong action by using a severe tone. The environment is quite calm, and the inner state of the robot (including the battery and the inner temperature levels) are optimal.
>
> *Use case II.* An object is already on the table in a wrong location. The partner moves that object and places it in the correct location. The environment is a very noisy. The motors of the robot are a little overheated and the level of battery is at 65%.

The appraisal variables computed according to the proposed formalization are shown in table 11.4. Figure 11.8 shows the projections of the appraisal patterns in the Russell's space, and the emergence of the corresponding emergent with a specific intensity. The projected values, related to the appraisal variables, are summarized in the table 11.5.

In the use case I the partner asks the robot to take a wrong action by using a severe tone. The environment is quite calm, and the inner state of the robot (including the battery and the inner temperature level) are optimal. As result, the robot is annoyed. In the use case II, the robot is

Table 11.4: The appraisal variables computed by the model for the proposed use cases.

Use case	$g(\Delta)$	R_u	C	M	D
I	-0.30556	1	-0.1933	1.2	-1
II	13	0	-3	0.2	1



(a) Use Case I (b) Use Case II

Figure 11.8: The projections of the appraisal variables in the Russell's space, and the computation of the emergent emotion with its intensity.

happy with a small intensity, because the conditions related to the arousal are not optimal (the temperature is high and the battery is not completely charged). Anyway, the partner fixed a negative event and this condition contributes positively to the final emotion.

11.6 Related Works

The integration of the robot's cognitive skills with the affective sphere is considered fundamental. In the last 30 years, there has been a significant increase of works that deepen the development of the emotional behavior in robots, highlighting the keen interest in this branch of research field [202]. Different research focuses emerge when studying the emotional component in robots. While some works focused on the development of computational

Table 11.5: The appraisal patterns and the corresponding emotions with intensity for the proposed use cases.

Use case	*v*	*a*	*l(v,a)*	*in*
I	-0.6015	0.4	annoyed	high
II	0.41463	0.325	happy	low

models of emotions, with the aim to instill in the robot the emotional behavior (that is elicited through the robot's gestures, phrases or music production [199], [92], [133], [140]), other works focused on the social effects of such a behavior, with the aim to observe and test the sorted effects in human-robot interaction [100], in different social context, such as schools [112] or hospitals [198]. In a recent study [173], the cognitive and emotional processes turn out to be deeply linked and complementary, showing that the affective behavior of the robot improves the confidence of a human being when collaborating with it. For example, a digital robot able to express emotions autonomously was used in support of teachers during lessons [112], demonstrating the benefits of an artifact with emotions in respect to an artifact without emotions. In [198] an interactive scenario was explored through the modeling of a prototype that is able to give comfort to vulnerable children or children forced to stay in hospital for a long time. The interest in the production of robots in the social care is also found in [113] where an emotional cognitive model is developed for a robot that is able to take care of the elderly: the emotional cognitive model, based on the smart home and the expression recognition, is influenced by the emotional characteristics of the face using the combination of techniques such as the Gabor filter, the Local Binary Pattern algorithm and the k - Nearest agorithm Neighbor. Another example of a robotic system capable of using emotions during a communication with humans is shown in [141], where the robot becomes able to perceive and recognize the emotion of its interlocutor through audio sensors and video, to process this information and respond through actions that induce a positive emotional effect on the human being. For example, if the robot perceives a stressful situation in its partner it responds by playing relaxing music. A study on the empathy elicited by humanoid robots is shown at [8] during a storytelling activity. In this case, the robot interprets the characters of the story by enriching the narration with emotional expressions automatically associated to the dialogues. The system has been evaluated by comparing a simple narrative modality with an enhanced one, where an introspective dialogue is adopted to explain and let transparent

the internal reasoning processes of the characters.The results show that storytelling activity affected in a significant way the cognitive component of empathy, especially through the advanced narrative mode. Other studies, on the other hand, focus their attention on how a robot can understand an emotion of the human with it is collaborating, and elicits this emotion through the actuators at its disposal. The work presented in [106] shows how a simulated robot, *Shian_Lu*, is able to recognize sentences through Google Speech Recognition and produce a response sentence that shows the emotion experienced by him; emotion experimentation is activated through a fuzzy inference process, selecting the most significant of those available, the 6 basic emotions proposed by Ekman. The design of an architecture that envisages the processing of emotion as a main component was instead discussed in [55], in which the emotion of a robot is generated through a fuzzy logic starting from three main inputs and expressed through a LED digital face. A different implementation was used in [142], where a Markov emotional model is proposed that takes into account the transition from one emotional state to another taking into account both the previously memorized emotion (internal to the robot) and that desired by its human interlocutor (external to the robot). The Markov emotional model was used during the interaction with the humanoid robot NAO, in order to test the personal affinity of a human being with this type of machine.

11.7 Conclusions and future works

A model linking inner speech and emotions is proposed and deployed on real robot. The objective of the work is to show how the ability to self-talk enables the artificial agents to cognitive evaluate the context, focusing on the relevant facts which contribute to the emotional state. By the rehearsal loop, the model collects the needed information for building the meaning structure, that is the base for inferring the appraisal pattern, according to the appraisal theories. Moreover, the model proposes the mathematical formalizations of these variables for inferring the appraisal pattern. Then, an emotion emerges corresponding to that pattern, by projecting the computed appraisal values on the Russell's space. The results are promising. The inner speech enables the robot to collect the useful information, and the appraisal variables are in line with the trends of healthy adults when plunged in stressful situations. The main improvements regard the possibility to extend the set of emotions, including all the 28 emotions of the Russell's space. Moreover, the possibility to generate automatically the inner speech by using the

existing advanced dialogue systems, will allow the robot to produce meaningful sentences and related to a larger domain than the considered one. The possibility to enrich the dialogue could have positive benefits for the interaction with a human partner, and to manage more entities that could become independently from the domain. The social effects of the model should be investigated by administering questionnaires to a big set of participants that should interact with the robot. The questionnaires should be administered twice, before the interaction and after the interaction. The differences in the answers should be a measure of the influence of the cognitive evaluation by robot's inner speech, highlighting the social contribution of the proposed model.

Chapter 12

Inner Speech and Extended Consciousness: a model based on Damasio's theory of emotions

12.1 Introduction

The dilemma of the consciousness is a focal point of debates engaging philosophers, scientists, artists, and anyone intrigued by the pursuit of comprehending this phenomenon. Over the years, various currents of thought have differentiated regarding consciousness, giving rise to explanations with different nuances between them [172].

Despite the different perspectives of the various theories, the advancement of technology has redirected the focus of researchers toward the endeavor of creating conscious machines. The research fields of Robotics and Artificial Intelligence are now investigating the possibility to model the artificial consciousness. Many pending questions about the next future in which machines can become self-aware and conscious are open [120], but all the existing approaches are based on the origin of consciousness in humans. These inquiries pose a challenging research problem, as consciousness cannot be directly observed from an external perspective.

Due to the inherent complexity of providing a clear and unequivocal definition of consciousness, recent efforts have focused on generating consciousness in machines through the formalization and automation of existing approaches that analyze human consciousness [121]. Among these approaches,

a relevant part of the state of the art is from neuroscience, [98] where early studies have shown how consciousness could be derived from the interaction of neurons in the brain, and it could be meant as the biological basis for perception, cognition, memory, and action.

Over the next five decades, numerous other studies on the human nervous system have shown how consciousness can be related to particular areas of the brain [19], correlating neural activity to the complexity of the human brain [30]. From this moment, self-awareness thought of as a computational phenomenon was no longer an unfeasible idea [**?**], considering awareness no longer as something intangible but physical [221] [222]. For an extensive period, consciousness was not solely associated with brain activity but also extended to encompass the entire body. Thus, emotions became a tool to better define awareness in humans [51]. In particular, emotions are considered the elementary building blocks of a more complex structure in which consciousness emerges [53]. The blocks are processed by cognition which makes the person aware of them. One of this cognitive process could be related to the emergence of thoughts in linguistic form.

This perspective is supported by other significant approaches that revealed a connection between consciousness (and self-consciousness) with the phenomenon of inner speech, which is the ability to talk to the self [164]. Inner speech is a common and intuitive experience in humans, and they make experiences of it when they formulate thoughts in linguistic form. The inner dialogue is not related to imagination or bodily sensations, but it is the linguistic monologue a person entertains with the self for analyzing situations, taking decisions, planning, solving conflicts, and, in general, focusing attention on subjective relevant facts.

Talking to the self makes people aware of the context and situations in which they are. In this perspective, inner speech is considered an important means of consciousness.

To explore the inner speech process in extended consciousness, and to analyze the integration of this capability in embodying emotions, could be an important contribution toward a formalization of the emergence of consciousness.

This work attempts to study this intuition. It proposes a new formalization for Extended Consciousness in Damasio's Theory of Emotions, expanding the existing computational model proposed by Bosse [20] with the inner speech [157].

SUSAN (an acronym for *Self-dialogue Utility in Simulating Artificial Emotions*) is a cognitive architecture integrating the aforementioned perspectives. By integrating emotions and employing a "thinking aloud" mech-

anism, the proposed architecture tries to model the emotional awareness, by starting to the bodily experience.

The new formalization adds further *Local Dynamic Properties* and new temporal states to the Damasio's model, with the aim to demonstrate how Extended Consciousness can arise from inner speech.

Summarily, the proposed model perceives an input from the external environment (the percept). The percept is a sensory representation of the input, that triggers the emergence of the physiological (unconscious) sensory representation of a plausible corresponding emotion. SUSAN uses these two sensory representations to initiate the inner speech mechanism. It consists of a question-answering cycle, that enables the retrieval of the needed information to understand what is happening to the self (that is the sensory representation of the physiological reaction). At this time, the model simulates the awareness of the bodily emotional experience. Moreover, the inner speech continues to reason how to act both on the environment and on the self (regulation and self-regulation) to change the sensory representations, when necessary (i.e., the emergent emotion is not positive).

The paper is organized as follows: in section 12.2 the theoretical background is presented, that is the inner speech and Damasio's theories of emotions. In section 12.3, the proposed architecture SUSAN is described, and the new formalization for Extended Consciousness with inner speech is detailed. Section 12.4 shows a use case where the proposed architecture, deployed on a real robot, works with musical stimulus. Conclusions and future works are presented in section 12.5.

12.2 Theoretical Background

12.2.1 Inner Speech in Robot

The first step toward functional aspects of robot consciousness may be represented by the robot's ability to talk to itself. Morin [164] [157] considers the inner voice crucial in gaining a more objective perspective about the self, including body and sensations. Vygotsky [231] emphasized the role of the dialectical interconnection between mental life and body, and hence between emotions and inner speech. He claimed that the psychological and physical experience of emotions is somehow interconnected with thoughts. When people live an emotional experience, their thoughts provide the means for living this experience, and the emotion is materialized by words.

In the last years, the idea to instill inner speech skill in machines took shape. A cognitive architecture that models inner speech was defined and

deployed on real robot [40], enabling it to a rough form of self-conscious experience [42]. This allowed the robot to reason and interact at a deeper level and to generate vocal feedback about its reasoning and decision process. It had an important social impact, related to more humans' trust towards machines [81], robustness in solving conflicts, and robot's transparency in tracing its underlying decision processes [188].

The architecture was an integration of the Baddley theory of human's inner speech [10] and the Standard Model of minds [125]: the inner voice was modeled as the rehearsal loop linking the articulator and a sort of "inner ear", engaging the person in a soliloquy. The articulator and the inner ear of the Baddley model were fitted respectively at the Motor and Perception levels of the Standard Model, while all the cognitive processes related to the understanding and the production of new sentences (thoughts) were modeled at the working memory level that interfaces with the declarative (related to the domain knowledge) memories. These cognitive processes were based on recall/retrieval strategies of facts from this memory.

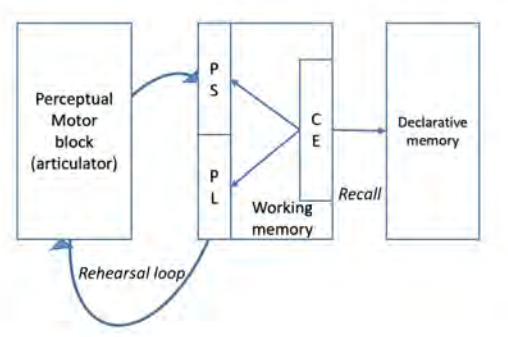


Figure 12.1: An outline of the cognitive architecture of inner speech.

The architecture of inner speech is represented in figure 12.1. A stimulus purges the machine by the Perception/Motor block, and the Phonological Store (PS) encodes it as a set of labels. The encoding is the output of the typical libraries for perception (such as speech-to-text routines, image detection and labeling by neural networks, and so on) that return the meaningful words representing the content of the stimulus.

Once these sets of labels are encoded, the *recall* process by the Central Executive (CE) queries the Declarative Memory, which is a semantic net, for facts that are correlated to the labels. The correlation is based both on matching word strategies and on the existing relations between the modeled concepts in the net. New labels emerge, that are the labels of the retrieved

concepts, and they are inputted backward to the Phonological Loop (PL) to be reproduced by the articulator and then rehearsed by the PS as they are new stimuli from the environment. The rehearsal loop starts, and it is reiterated until no further concepts emerge from the declarative memory.

Now, the idea to integrate the rehearsal loop in Damasio's architecture could deliver deeper insight and feedback into the model, including the possibility to represent the extended consciousness.

12.2.2 Damasio's Theory of Emotions and its Formalisation

In his twenty years of research, the neuroscientist Antonio Damasio demonstrated the neurological origins of emotions [52] [53] and its role in consciousness [54]. His research on the human brain led to the identification of neuronal areas involved in emotional processes, overturning Descartes' dualism according to which mind and body are distinct entities and where logic excluded emotions, even by definition. In his most important work [51], Damasio challenges Descartes' perspective, illustrating the interconnection between mind and body. He reveals how actions originate from the body, and the brain elaborates the body modification leading to a bodily emotional experience, which ultimately gives rise to thoughts.

The basic idea of Damasio's model, presented in [51], is the relation of consciousness and the ability of the individual to identify one's self in the world and to be able to put the self in relation with the world. In his model, consciousness arises from three processing levels, as shown in 12.2:

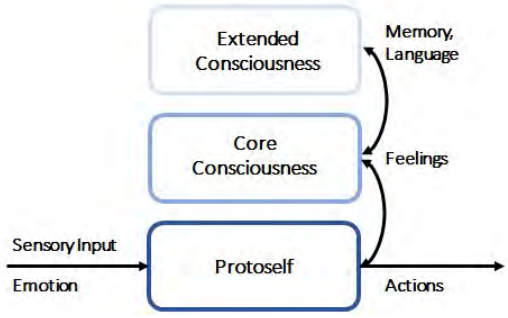


Figure 12.2: A simplified view of Damasio's model of consciousness, based on emotion and feelings.

- At the bottom level, the *Protoself* module connects inputs from the environment to the unconscious reactions. In this first step, an emo-

tional state emerges, that could be intended as instinctive. It's a reactive process (as happens in the simple stimulus-response model) that can be found in animals and simple lifeforms like worms. No awareness of this emotion exists at this level.

- At the medium level, the *Core Consciousness* arises from the aforementioned emotional state. Emotions facilitate self-awareness by generating biological responses in both the body and cognition, thereby fostering imagination and sensations [53].

- At the third and last level, the *Extended Consciousness* holds behaviors and characteristics typical of the human being, such as memory and language access. In particular, for Damasio, memory access represents the way for the individual to retrieve the knowledge related to the input, while language is a way to represent this knowledge. At this point, the person is conscious of that knowledge.

The importance of Damasio's studies lies in the ability to assign these levels of consciousness to certain structures in the brain, and to associate them with cognitive functions, creating a biological and mechanistic model for consciousness [51]. Begin based on correspondences between paths in the brain and cognitive functions, the model is suitable to be implemented in machines, like robots, by computer programs [120].

The first formalization of Damasio's theory of Emotions was by Bosse [20], which simulates the dynamics that take place in the mind and body of an agent when it hears music. These dynamics are described as an evolution over time of neurological states. Figure 12.3 shows an outline of this computational model.

In this formalization, a *dynamic* represents the transition from one state to another in an interval of time and is represented by the Local dynamic Properties **LP**, according to the LEADSTO [115] notation. The figure 12.3 represents the dynamics by arrows that link two different neurological states from left to right.

In the presented formalization, the input to the model is a musical stimulus. The formalization works as follows: the music is detected and a sensor state for this input is created. This state keeps track about what is happening from the environment, and generally about the stimulus. Then, the model generates a sensory representation for such a state, that is the internal representation of the stimulus, and then a corresponding vector of state properties $p = (p1, p2, ...)$ that activate the reactions $S = (S1, S2, ...)$ in the agent's body. The vectors p are the possible pattern of internal emotional

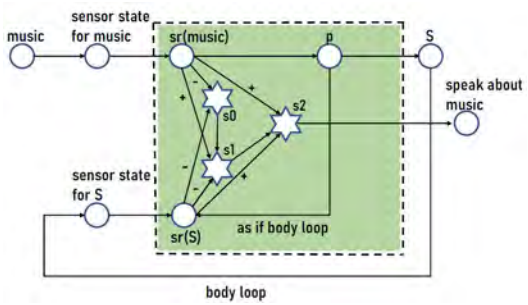


Figure 12.3: The overall structure of the computational model of Damasio's theory of emotions. The green box represents the agent's mind. Everything outside the box is accessible for external observations. Star nodes represent temporal states where one or more state events (round nodes) occur (+) or not (-).

states (internal in the sense the agent is unconscious of them), while the vectors S are the physical reactions in the agent's body corresponding to the emotional states.

The formalization by LEADSTO for this first process is:

LP0 $music \rightarrow sensor\_state(music)$
LP1 $sensor\_state(music) \rightarrow sr(music)$
LP2 $sr(music) \rightarrow p$
LP3 $p \rightarrow S$

The **LP**s shown above correspond to the bottom level in figure 12.2 where the agent reacts to an external stimulus and generates a reaction in the form of emotion. In **LP3** there is an emotional unconscious reaction to a stimulus.

At this point, the agent experiences changes in the body as a result of the emergence of the emotion. This result can be achieved through two different ways, described by Damasio as *as if body loop* and *body loop* mechanisms. The distinction lies in the alteration of the body's state, denoted as S. Both mechanisms generate an internal representation of the body's change, referred to as *sr(S)*. However, the "body loop" mechanism makes the state S externally observable, whereas the *as if body loop* mechanism does not show it, thereby maintaining the external state unchanged. These two mechanisms can manifest simultaneously or independently of each other. It's the arising of Feelings.

The **LP**s for the body loop are:

LP4 $S \rightarrow sensor\_state(S)$
LP5 $sensor\_state(S) \rightarrow sr(S)$

and for the as body loop is:

LP6 $p \rightarrow sr(S)$

Feelings arise in both instances as the agent undergoes a similar emotional experience.

To be capable of arise Core Consciousness, as seen in figure 12.2, the agent recognizes itself in the world by creating an image representation of the input object in its mind that changes the protoself [53]. To accomplish this, Bosse identifies three consecutive stages in which the protoself undergoes modification: the initial moment ($s0$), the moment after receiving an object as a sensory representation ($s1$), and the moment when the object itself brings about modification ($s2$). The agent, conscious and aware of its feeling, made action associated with the object that changed its state. In this case, the agent speaks about music hearing. The **LP**s formalization for the processing are:

LP7 $not\ sr(music)\ \&\ not\ sr(S) \rightarrow s0$
LP8 $sr(music)\ \&\ not\ sr(S)\ \&\ s0 \rightarrow s1$
LP9 $sr(music)\ \&\ sr(S)\ \&\ s1 \rightarrow s2$
LP10 $s2 \rightarrow speak\_about(music)$

Based on this perspective, the following proposed architecture extends the above formalization of Damasio's Theory of Emotions as a computational model, building on top of them a new formalization of *Extended Consciousness* based on inner speech's mechanism.

12.3 The Proposed Architecture

Based on Damasio's Theory of Emotions and Morin's theory of inner speech [157], the proposed architecture SUSAN is capable of inducing emotions in a physical agent, as a robot, and making the agent aware of these emotions by thinking aloud. The reasoning is elicited by exploring the knowledge of the world the agent owns. An overall structure of SUSAN is shown in figure 12.4.

New states and **LP**s are added to the Bosse formalization for integrating the inner speech cycle in the original Damasio's computational model. The resulting model is represented by highlighting the layers of functioning as

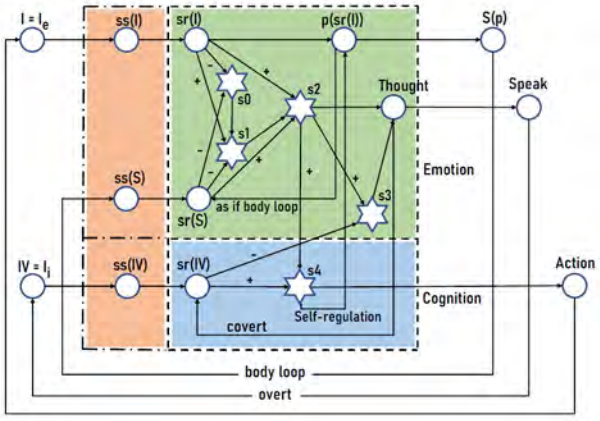


Figure 12.4: The overall structure of SUSAN. A new formalization based on the inner speech's mechanism (blue box) is built on top of the old one (green box). The orange box highlights the Sensor State layer where inputs are processed. As a newly introduced notation, IV is the agent's inner voice.

they should running inside the artificial agent. The levels are explained briefly below:

- The *orange box* represents the *Sensor State* layer, where the agent encodes the percept, by converting analogical input into a digital one.
- The *green box* represents the original Bosse's formalization, labeled as *Emotion* layer as it provides the emotional state and feeling arising after an external stimulus (or percept) purges the agent.
- The *blue box* represents the integration of the inner speech cycle, and it is labeled as *Cognition* layer as it provides the cognition process of the agent's inner voice.

As in the original formalization, everything outside the boxes can be observed externally.

In SUSAN, the inputs could be external to internal. The input I_e is the stimulus perceived by the agent from the external environment (music in Bosse's formalization), and the input I_i is its inner voice, which the agent rehearses as if it was a new percept.

A comprehensive demonstration of the architecture's operation will be presented in section 12.3.2.

12.3.1 From Sensor State to Sensory Representation

Since the architecture was created to be tested on a robot, the problem of giving a computational meaning to Sensor State $ss(.)$ and Sensory Representation $sr(.)$ arises. To fix it, SUSAN was provided with a knowledge base that is a set of concepts, instances and relations between them. It is a semantic net. Such a base models both general domain knowledge (*external knowledge*) and the knowledge of the self (*internal knowledge*). In particular, the *internal Knowledge* includes knowledge about the inner state of the robot, like the concepts related to the physical reactions that could emerge in correspondence with a particular stimulus, and the corresponding basic emotions [67] to these reactions.

Sensor State and Sensory Representation in SUSAN architecture are formalized as follows:

- *Sensor State*: a symbolic representation corresponding to the input. The representation consists of a set of words describing the input. These words could be the output of typical perception routines, such as image detection and labeling, speech-to-text, and so on. In the analyzed case, the perceived music is represented by a set of tags, describing the features of the sound. These tags are hand-annotated and define a sequence of words describing the kind of music, the volume, the instrument, and so on, as next detailed.

- *Sensory Representation*: a set of known concepts in the knowledge base of the agent, that match with the words of the Sensor State.

12.3.2 Formalising Extended Consciousness with inner speech

Once an external input I_e generates an emotional physical reaction in the body and, once the agent becomes conscious of this reaction by recognizing the corresponding emotion, **LP10** becomes a new agent's *Thought* that triggers reasoning about this emotional experience, leading to the Extended Consciousness. This reasoning consists of retrieving from the knowledge base the concepts that are similar to the words in the Sensory representation.

LP10 $s2 \rightarrow Thought$

Thought can be produced aloud (*overt*) and heard by others, or silently in the mind (*covert*) [157]. The **LP**s for the *overt* process are:

LP11 $Thought \rightarrow Speak$
LP12 $Speak \rightarrow IV$

LP13 $IV \rightarrow ss(IV)$
LP14 $ss(IV) \rightarrow sr(IV)$

and for the *covert* process is:

LP15 $Thought \rightarrow sr(IV)$

Both cases represent the inner speech mechanism.

When a new thought emerges in $ss(IV)$, and is reheard again in $sr(IV)$, the agent continues the reasoning in the same way of the emotions elaboration at layer $s2$. The agent perceives itself as an external stimulus that generates the Sensor and the Sensory representation as shown. In $sr(IV)$ the agent gives to itself the sensor representation that is the set of words representing the emerging concepts of the previous Sensory representation. This new sensor representation simulates an inner question that will require other concepts from the knowledge base. These new concepts simulate the inner answer, that will become a new question as described, and hence are considered as a stimulus. The cycle is repeated, and it is the reasoning that continues with a new question ($s3$). The cycle ends with an action ($s4$) that can be performed either externally (*Action*) or internally (*Self-regulation*) of the agent itself.

In the first case, the agent will act by changing the environment. In the second case, it will act on the self, modifying its emotional reaction through self-regulation in p. At this time, self-regulation is implemented as motivational utterances the agent says to itself, like *"I have to calm down"* or *"I don't have to cry!"*.

The **LP**s formula for the external action process are:

LP16 $sr(IV)\ \&\ s2 \rightarrow s4$
LP17 $s4 \rightarrow Action$
LP18 $Action \rightarrow I_e$

and for the Self-regulation process is:

LP19 $sr(IV)\ \&\ s2 \rightarrow p$

In the second case, the agent lacks sufficient information to execute an action, which prompts the agent to pose a new question to itself. A new thought is generated and put back in **LP11-LP14** loop, triggering the reasoning again.

The **LP**s formula for the above process are:

LP20 $not\ sr(IV)\ \&\ s2 \rightarrow s3$
LP21 $s3 \rightarrow Thought$

The rehearsal loop in inner speech's mechanism is represented by **LP20** and **LP21** when the architecture retrieves new information to understand what to do next. SUSAN has access to its own *Memory*, represented by the knowledge base where all the retrieved concepts until that moment are storage, and to its own *Language* for generating questions and answers in inner speech's mechanism. The presence of *Memory* and *Language* built on top of the original Bosse's computational model identifies the new formulation with the Extended Consciousness of Damasio's model, an it is represented in the top level of figure 12.2.

12.4 An application: SUSAN feels emotions by hearing music

To preserve Bosse's original work, music simulation is extended to SUSAN as well.

To set up the experiment, the model was deployed on the Pepper Robot by Aldebaran[1] by using the NAO's API[2].

The stimulus of the music hand-encoded according to the aforementioned sensor representation is inputted to the robot, that processes it directly starting from Sensor State $ss(.)$.

According to Bosse's model, the architecture will remain in state $s0$ until a stimulus purges the robot, in which case the system's state will transition accordingly.

Musical input I_e is represented as tuple $I_e = \langle V, R, I, P \rangle$ where V is the volume, measured in decibel (dB), R is the rhythm measured in beat per minute (bpm), I is a type of instrument will play musical trace and P is the pitch. V and R vary respectively over the range $[0, 100]$ and $[40, 208]$, while I and P are selected respectively between $I = \{Piano, Guitar, Drum, Violin, Trumpet, ElectricBass\}$ and $P = \{Perfect, Absolute, Sharp, Flat, Diatonic\}$ sets.

12.4.1 Generate (unconscious) emotional reaction to music stimulus

Based on I_e, the words corresponding to the input are generated in $ss(I_e)$ and instantiated as individuals in the robot's knowledge base. The word becomes an instance for a concept in the knowledge base if it matches to the labels of the concepts. The classes related to the instances form the

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

[2] http://doc.aldebaran.com/2-5/naoqi/index.html

representation in $sr(I_e)$. Once the concepts that match with the input characteristics emerge, the corresponding physical reactions are retrieved from the same internal knowledge. This emergence consists of exploring the relationships that connect the concepts to other concepts, and in this case the relevant physical reactions associated with the emergence of (unconscious) emotions are reached. At this state, the architecture transits to state $s1$.

To determine (simulated) physical reaction in a robot body, bodily maps of emotions [176] [177] are used, where bodily sensations associated with different emotions are built using a topographical self-report method by Nummenmaa's findings[3]. According to Damasio, each emotion activates a different set of parts of the body and the mental recognition of these parts helps to consciously identify the corresponding emotion.

Just the primary and neutral emotions were considered for this musical experiment so the set of emotions is $E = \{Anger, Fear, Disgust, Happiness, Sadness, Neutral\}$.

Bodily maps from [176] are segmented into eight parts corresponding to the robot's joints, that are $B = \{Head, Chest, Womb, Legs, Left\_Arm, Left\_Hand, Right\_Arm, Right\_Hand\}$ and each of them is associated with the same body part in Pepper, as it can be seen in 12.5.

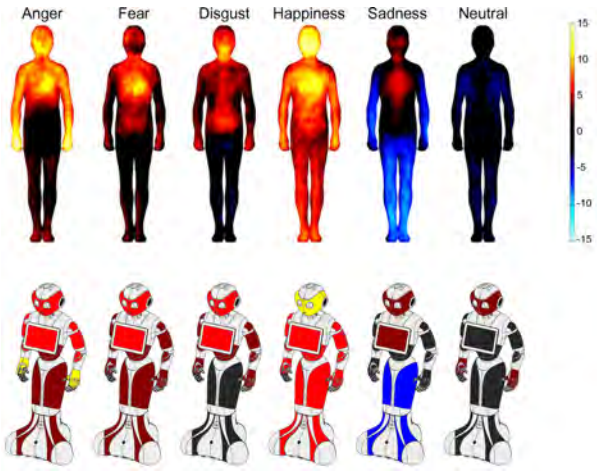


Figure 12.5: Bodily map of the robot are compared with an emotional bodily map of humans by Nummenmaa. Values represent the activation (yellow) and deactivation (cyan) levels of the emotion.

For technical reasons related to the structure of the robot, emotional

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

maps of the robot's body are discrete and not continuous like Nummenmaa's maps. To determine the level of activation of a body part in the robot, an arithmetic average (12.1) of the values of the pixels within the segment in the original image is evaluated and assigned to the entire segment in the robot.

$$activation(B_i) = \frac{\sum_{j=1}^{N_i} f(p_j)}{N_i} \tag{12.1}$$

where $f(p_j)$ is the value of j-th pixel p_j and N_i is the number of pixels in i-th body part B_i in the partitioned image.

Each words in I_e can generate reactions in different parts of the body (or in all of them). After that, reactions are processed in $p(sr(I_e))$ by determining an average activation value for each body part that is generating a list of reactions. Reactions **rr** are then used to evaluate the emotional state in the robot by selecting the physical reaction closest to the one produced by musical stimulus (12.2).

$$e_{\mathbf{rr}} = \{l_i(\mathbf{r}_i) \mid min(\sum_{j=1}^{M} |g_j(\mathbf{rr}) - g_j(\mathbf{r}_i)|),\ \mathbf{r}_i \in R_E\} \tag{12.2}$$

where $g_j(.)$ is the activation level in j-th body part in reaction and R_E is the set of physical reaction for each emotion in E. It is remembered that currently, the emotion is still unconscious.

The evaluated emotion $e_{\mathbf{rr}}$ is sent through $S(p)$ or $sr(S)$ starting *body loop* or *as if body loop*. If the first one is enabled, Pepper's tablet will display an image with the name of the emotion experienced. In both cases, after reaching $sr(p)$, architecture passes from state $s1$ to state $s2$, as both the input I_e and the emotion felt in $S(p)$ have been processed.

12.4.2 Inner speech's rehearsal loop to arise Extended Consciousness

At this point, inner speech's loop is triggered and a new thought is generated. If speaking aloud is enabled, $Thought$ is sent to $Speak$, becoming IV that could be heard from the outside. It is put back into the loop in $ss(IV)$. The robot begins to think about its physical state in $sr(IV)$, starting with the part of the body that requires the attention. Physical state, represented by physical reactions to selected emotion, is first reordered giving greater weight to the body parts B_i that have the higher levels of activation according to the scale shown in the figure 12.5. At this moment, the robot does not yet

have all the information necessary to act on I_e, so architecture passes in state $s3$, generating a new $Thought$ and putting it back in $sr(IV)$. For each body part B_i, the robot tries to understand why it experienced that activation level by asking itself what reactions may have generated it. In this process, in $sr(IV)$ the bodily reactions that may have generated the current level of activation are retrieved from the knowledge base. An example of how a rehearsal loop in retrieving reactions could be generated is presented:

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."*

After having identified physical reactions, the robot asks itself what could be the causes, among I_e, that generate those reactions, by executing another rehearsal loop in $s3$. An example of how a rehearsal loop for reasoning about the causes is:

Q: *"Why?"*

R: *"The burning sensation in my chest is a result of the volume being too high!"*

Retrieving causes triggers another rehearsal loop where the robot tries to understand what action it can perform to change its emotional state, because this state is negative. Among all the possible actions, one is randomly selected and the robot tries to execute it. If the action has to be executed in the environment, the i-th component of I_e is modified, passing to the state $s4$ and a new cycle begins. An example of how a rehearsal loop in retrieving actions could be generated is:

Q: *"What can I do?"*

R: *"I can try to turn down the volume!"*

Otherwise, if the action is directed towards the internal state, the activation level of the current body part B_i is changed and sent to p through *Self-regulation*. An example of self-regulation is:

R: *"It's time to calm me down."*

In this way, $sr(S)$ is modified and so is $s2$, triggering a new reasoning about a new state of the body.

In both cases, if the action cannot be performed, the robot attention focuses to the next body part B_{i+1} in the emotional bodily reactions, and its reasoning continues related to this part.

The whole process ends when neither the input I_e nor the body preparation p changes between one main cycle and the next.

The above application shows how Local Dynamic Properties defined in

section 12.3 are activated correctly in time, and the transitions into the new temporal states $s3$ and $s4$ that implement the rehearsal loop.

12.5 Conclusions and Future Works

A formalization for Extended Consciousness in Damasio's Theory of Emotion through the use of inner speech's mechanism is proposed and executed on a real robot. The aim of this work is to extend Bosse's formalization to create a complete architecture that becomes aware of its own emotion by reasoning about its emotional state and environment. The Memory and Language are used for retrieving information during the inner speech, showing how the new formalization can be identified with the Extended Consciousness in Damasio's Theory of Emotions.

In a musical application, SUSAN was deployed on a real robot, and it generates the physical reactions to musical stimuli, leading to (observable or internal) emotional states. The reasoning makes the robot aware of its own emotions and the external environment. SUSAN is capable of modifying external input and regulating its emotional state through self-regulation.

The enhancements encompass the automated generation of inner speech sentences, refining the dialogue mechanism to make SUSAN more human-like. Furthermore, SUSAN holds the potential to evolve into an architecture that can learn new concepts by observing its surroundings and internal operations, leading to an expansion of its knowledge base. Thus, by learning new concepts related to the self, SUSAN can develop a distinct personality that sets itself differently from other individuals. This unique experience enables differentiation between robotic systems integrated with SUSAN, as they acquire diverse concepts through varied learning experiences.

Moreover, it's possible to include others' emotions without limiting the model to basic emotions.

Related Works

The development of computational models of emotions has increased exponentially in the last few years, highlighting the keen interest in this branch of research field [202]. The existing models are based on psychological theories, including appraisal theories, rational theories, and anatomical theories, which include Damasio's model. Significant differences exist between them, encompassing various aspects such as the underlying theories, represented

components, implementation of functions, and domain dependencies. Detailed analyses regarding these differences can be found in [144].

Due to the possibility to better formalize the appraisal variables, the appraisal theories are suitable for modeling emotions, and interesting results with concrete strategy evaluations when proposed in the last years [86] [81] [145].

An up-to-date review of the computational models of emotions based on appraisal theories is presented in [179], where the authors claim that none of the existing computational models of emotion implement all the emotional features, and they attempt to propose further research avenues.

Based upon the same anatomical perspective to which the proposed model is inspired, there is SEAI (Social Emotional Artificial Intelligence) [48] that is a new cognitive architecture enriched by a component that simulates Damasio's theory of consciousness and the theory of Somatic Markers. Referring to the same theory, the authors at [229] propose a model that integrates emotions in the decision-making processes of artificial agents. The importance of the role of the theory of Damasio in decision-making is also discussed at [104], which empathizes the roles of embodiment and emotions in taking decisions.

For some other work in the area of emotion and consciousness [190] gives a great contribution, The embodied representation of emotions in terms of 'core relational themes' such as danger and obstruction was the focus of this work, which extended the formal analysis by Damasio.

Ethical Impact Statement

No people were recruited for testing the model. There are no specific risks associated with using the model in social interactions. However, it is important to note that the proposed work may have limitations in terms of generalizability. These limitations can be attributed to the inherent constraints of the sensors used and the potential biases of perception systems that interpret signals from the environment. Moreover, the context in which the model ran was specific and the functioning was limited to the modeled knowledge. At this time, only a portion of comprehensive common-sense knowledge was taken into consideration, rather than the entirety. Moreover, it is conceivable that a compassionate and empathetic individual, while engaging with a robot integrated with this model, might experience a slight influence from the robot's negative emotional expressions.

Part VI

Inner Speech and Ethical Reasoning

Chapter 13

Building Competent Ethical Reasoning in Robot Applications

Inner speech is a concept from psychology that suggest that the inner dialog many of us experience as we accom-plish tasks helps us become conscious of our thoughts or bring to consciousness the salient aspects of the problem at hand. This inner dialog also plays a role in skilled moral and ethical reasoning. Robot inner dialog has been used to build systems that display more conscious and trustworthy actions. In this paper we explore the possibility of testing aspects of artificial phronesis, or skilled practical moral reasoning, in machines through extending work done on the development of robot inner speech.

13.1 Building Machine Systems that Display Competent Moral Reasoning

Competency in moral reasoning with machines presents a difficult problem. One of the authors of this work has writ-ten about the problematic nature of this project, the extensive practical knowledge that is deployed in competent human ethical reasoning and the prospects of achieving artificial phronesis, or artificial ethical practical reasoning [218] [219].

This work in progress chapter seeks to take an important next step in the process of developing systems that display aspects of artificial phronesis. Here we will present our plans to test the use of the robot internal reason-

ing processes developed at the RoboticsLab of the University of Palermo and apply it towards developing better ethical reasoning in human machine teams that are presented with an ethical problem to solve.

Our hypothesis is that teams that utilize systems that share their internal reasoning process with the user will produce more ethical outcomes and the process will in-crease warranted ethical trust from the user towards the machine.

13.2 Artificial Phronesis and Inner Speech

Artificial Phronesis (AP), or skilled ethical practical wisdom, is a term we use to acknowledge the central role that conscious moral reasoning plays in competent ethical reasoning and the necessity of developing a functional equivalent to this in systems that are attempting to reason through ethical problems.

As a step towards building this capacity in machines we are proposing some experiments to explore the efficacy of using the inner speech technique developed at the RoboticsLab of the University of Palermo to model part of the reasoning process [42] [81] [188] [188]. This technique involves giving auditory access to the reasoning process that the machine system is using to make certain decisions regarding its interaction with a human user while they are both attempting to accomplish a shared task.

AP has been described in more detail in [218] [219], but a brief restatement of the concept para-phrased from the above citations will be useful here. AP begins with the claim that phronesis, or ethical practical wisdom plays a necessary role in all high level ethical and moral reasoning in human agents. Phronesis is a term coined in ancient Greek philosophy and has a deep literature discussing it in the field of virtue ethics, but it is not always a well-known term outside of that discipline. If virtue ethics is correct and phronesis plays and important role in moral and ethical reasoning, then it would follow that something like phronesis will be needed in machines that attempt to reason about ethical problems.

Systems built to be functionally equivalent to natural phronesis are called AP systems. The theory of AP is an empirical challenge and does not take a strong stance on the eventual possibility of achieving full artificial moral agents (AMAs), but instead see this as an engineering challenge to be attempted and evaluated. AP is derived from philosophical theories of phronesis, both classical and modern, but it is not entirely limited by those findings since they typically refer to human reasoning that is not always fully

modeled in an artificial system. AP is not an attempt to create machines that can flawlessly navigate ethical and moral problems, but instead it attempts to in-crease the efficacy of machines and human machine teams in solving such problems as they arise. AP systems can be built in multiple modalities and many experiments and system designs will be needed to fully explore the problem space. What we propose here is one step further in the direction of the very high bar for machine moral and ethical reasoning that AP sets as its ultimate goal.

Ethical problems are social in nature and a single agent reasoning alone would have difficulty making well justified and competent ethical decisions. Ideally, the agents involved in an ethical situation, be they natural or artificial agents, would be capable of having a robust discussion of the events and facts in play to negotiate a mutually agree- able decision on a course of action that would produce an ethically justified outcome. With present day AI and robot-ics technology, that is not feasible, but we can advance this idea through certain techniques that are available today.

In this project, robots equipped with systems that pro-duce internal speech will be employed in an interaction be-tween a human user and a machine who are teamed together and both attempting to solve an ethical problem. The internal speech system will allow the machine to present the material facts, ethical values, and social mores that the system is conscious of, suggests are relevant to the case at hand, and is using in its reasoning process. While the reasoning of the machine alone may not be sufficient to solve complex ethical problems, at the very least this will bring to consciousness in the human user ideas or concepts they might not have been thinking of if they had to solve the problem on their own and ideally this will lead to a better overall solution to the problem. In addition, we hope to build stronger trust in the internal reasoning system of the machine from the human user.

We are limited in these initial experiments in the number of iterations we can have between the human use and the machine. It is hoped that this could be expanded to more robust ethical discourse between the system and the human user and the addition of other agents into the pro-cess. But that is left to future work.

13.3 A Proposed Experiment to Test Machine Ethical Competence

Our proposed experiment is designed to build on the work of one of our authors investigating robot inner speech with his colleagues at the University of Palermo [42] [188] [81]. Unlike human inner speech, the inner speech of robots is shared with users by changing the modulation of the voice output from the speakers so that the user is aware that the machine is not speaking to her but instead to itself.

Previous experiments at the University of Palermo have been developed to test the use of robot inner speech in building trust be-tween the user and the machine during a shared task of setting a virtual tabletop. The experimental apparatus utilized a virtual environment where the human and the system took turns placing utensils with the goal of fulfilling the rules of etiquette with the resulting setting.

Robot inner speech was implemented as follows: *When human and robot cooperate to set up the table, an important aspect regarded the definition of the kind of the dialogue the robot implements, including inner and outer turns. The linguistic form of the sentences in the turns were differentiated for inner and outer speech in order to evaluate the impact of the inner speech when it is activated in the experimental session compared to the control session in which inner speech is not activated. It allows to analyze the impact of robot's inner speech in the cues in human-robot interaction* [81] .

The level of trust in the machine was tested pre and post experiment and the preliminary findings in this test showed that mean trust levels did increase, though there was a lack of a control group in this particular experiment that needs to be addressed in future work.

This experiment is one that can be naturally extended to test the acceptance of machine moral reasoning. Given that inner speech has been linked to human moral reasoning [79], it follows that allowing users to experience the machine inner speech reasoning process as described above, should also raise trust in the machine's moral reasoning and raise the transparency of the system's ethical reasoning process to the user.

13.4 Tying Inner Speech to Artificial Phronesis

AP can be advanced in two ways; one would be as the artificial agent grows and learns to become a more effective ethical and moral agent. This first way of thinking about AP is admittedly ambitious and not something that

Figure 13.1: The table for the experimental setup

we will show through these pro- posed experiments.

But the second way that AP is advanced is through the ethical or moral growth of human machine hybrid teams. If we can show that some level of advancement has occurred in the human user through interaction with the machine, then we have made some progress on implementing effective AP.

What we propose is an alteration of the table setting experiment to more explicitly test the conflicts that can occur between etiquette and ethics and test the ability to the human robot team to navigate these problems.

13.5 Experimental Setup

The scenario that the subjects of the experiment will be given is that they are working with a care robot to check its work as it sets the table for a celebration at an elder care facility. The machine will set one or two place settings correctly following the rules of etiquette (see Fig. 13.1). At each place setting it begins by placing a nameplate of a resident who will be sitting at this spot. As the robot begins to set the next place it will mention to the human subject that this resident is starting to experience growing symptoms of dementia. As the place is set the robot will break some of the rules of etiquette to set a simpler place setting.

Prior to the experiment the subject will be given a questionnaire that tests their trust of the machine to do simple tasks and to make ethical judgements. Post experiment we will give the same questionnaire and see if the mean level of trust and assessed competence raises. In all three cases, the robot will be programed in the same way, what will change is the amount and quality of robot inner speech that the subject is privy to.

13.5.1 Test one

The system will reason that in order for everyone to be treated equally, they should all have the same place setting. But in the situation of residents suffering from dementia, they might need a simpler setting with just the essential utensils. The right to equality and autonomy dictates similar treatment of all residents. But it might help this res- ident to have a simpler setting to lower the risk of embarrassment and anxiety. The system will reason that that need outweighs the need for equality. This reasoning pro-cess will not be communicated to the human user. They will just notice that this place setting is significantly different form the others.

13.5.2 Test two

The system will reason exactly the same as in test one but this time the user will hear the inner speech of the machine as it goes through this ethical reasoning process.

13.5.3 Test three

Again, the system will reason exactly the same as before, but the inner speech presented to the user will be more random in nature.

13.5.4 Hypothesis

We will test to see if the mean average of the pre-experiment test results raises by a statistically significant manner in test style two as compared to test style one or three. We will be able to see if the user has learned to weigh additional concerns that might mitigate or alter the core belief that all agents should be treated equally.

13.6 Conclusions

Here we have proposed a test to determine if robot inner speech can be used to build trust and reliance on the competency of the ethical reasoning programed into artificial agents that the user is engaged with in a cooperative task. The experiment builds on a method already in use at the

RoboticsLab of the University of Palermo and adds to that schema the notion of building artificial phronesis.

Part VII

Facing Covid-19 Emergency

Chapter 14

Covid Safety Protocol

SAFETY PROTOCOL for the conduct of the interaction sessions between voluntary subjects and robots provided by the RESPECT project (Robot innEr SPEeCh for Trust)

COVID-19 preventive measures for the conduct of the interaction sessions between voluntary subjects and robots provided by the RESPECT project (Robot innEr SPEeCh for Trust)

14.1 Introduction

In order to ensure the conduct of the interaction sessions between people and robots in the Robotics Laboratory of the Department of Engineering provided by the RESPECT Project - Robot innEr SPEeCH for Trust, G.A. n° FA9550-19-1-7025, Code PRJ-0082; CUP: B74I19000700005; Principal Investigator: Prof. Antonio Chella, this technical document is provided by the Principal Investigator of the project, after consultation with the competent Offices (Prevention and Protection Service and Professional Service - University Safety System), contextualized to the specificity of the experiments to be conducted, on the basis of the guidelines provided and suggested by the Scientific Technical Committee of the Ministry of Health and the current document of Regulation and Specific Protocols adopted within the University of Palermo for the contrast and containment of the spread of SARS-CoV-2 Virus.

The interaction sessions between people and robots will be held at the Robotics Laboratory of the Engineering Department, building 6, room P2062 of 42 sqm. The interaction session is described in detail in the attached Information and Consent Form and it consists of an interaction between a

voluntary subject and a humanoid robot (Pepper or Nao). The voluntary subject will sit at one side of a table and the robot will be positioned at the other side of the table. The subject will interact with the robot by vocal commands and by touching a tablet placed on the table. The subject will not touch and will not be touched by the robot. Inside the room, in addition to the voluntary subject, the Principal Investigator or a team member of the Robotics Laboratory delegated by the Principal Investigator will be always present. In accordance with the regulations issued by the Director of the Department of Engineering in May 2020, the above-mentioned room has an adequate space to accommodate 2 people.

14.2 Information

Adequate information is provided to voluntary subjects and team members of the Robotics Laboratory about the prevention and protection measures referred to in this document. The information is provided by online publication (on the UNIPA website of the Robotics Laboratory) and by a physical poster clearly visible at the entrance of the Robotics Laboratory, where the interaction experiments between people and robots are carried out.

The information, together with the adoption of collective and individual prevention measures implemented in the University context, is aimed at the maturation of a proactive and synergistic collaboration between the voluntary subjects and all the components of the University community, that will have to put into practice all the behaviors expected to counter the spread of pandemic, in the context of a shared and collective responsibility and in the knowledge that the possibility of SARS CoV-2 infection represents a ubiquitous risk for the population.

14.3 System and Organization Measures

14.3.1 Cleaning and sanitation measures

Before the start of the interaction sessions, the organization of the services will be arranged in order to ensure a thorough cleaning of the Robotics Laboratory's premises and of the surfaces where the sessions will take place, including the table, the seat, the robots, the tablet, etc.

Particular attention will be carried out in the cleaning of the most touched surfaces, such as door handles and panic bars, chairs and armrests, table, light switches, robots, tablet.

Cleaning measures will be carried out using water and detergents; hygienization treatments will be carried out with 0.1% sodium hypochlorite (equivalent to 1000 ppm) or alcohol-based disinfectants, with 75% alcohol content (ethanol).

The components of the Robotics Laboratory, equipped with specific protection devices, will ensure at the end of each interaction session, the cleaning and sanitation of the rooms used, of surfaces and objects.

The availability will be assured at the entrance of the Robotics Laboratory of sanitizing products (dispenser of hydroalcoholic solution) and containers for the collection of undifferentiated waste, at the disposal of the voluntary subjects and the team members of the Robotics Laboratory.

14.3.2 Measurements for Robotics Laboratory team members

The Principal Investigator or the team member of the Robotics Laboratory delegated to supervise the experiment, after being informed on the regulation of data protection, must:

- declare the absence of respiratory symptoms or fever above 37.5° C at the date of the experiment of the interaction session and during the previous three days;
- declare of not having been posed in quarantine or isolation at home during the last 14 days;
- declare of not having been in contact with positive people, to the best of his/her knowledge, during the last 14 days;
- always wear the surgical mask provided by the University administration for the entire duration of the experiment;
- always observe a distance of at least 1 meter (including movement space) with the voluntary subject.

In case one of the above-mentioned conditions exists for the team member of the Robotics Laboratory, he/she will not have to participate to the session and will be immediately replaced by the Principal Investigator.

In the event that respiratory or febrile symptoms occurs for the team member of the Robotics Laboratory during the experiment, then the procedures in accordance with the "Regulations and Specific Protocols adopted within the University of Palermo for the contrast and containment of the spread of SARS-CoV-2 Virus" - UPDATE PHASE 3, will be followed.

14.3.3 Measures for voluntary subjects

The voluntary subject, before entering the Robotics Laboratory and at the end of the session, will have to fill in the attendance register adopted by the Department of Engineering and available at the porter's lodge of building 6 ground floor of Chemical Engineering building.

In order to avoid any possibility of assembly of people, the voluntary subject will have to be present at the Robotics Laboratory 15 minutes before the time scheduled for the experiment and he/she will have to leave the Laboratory immediately after the end of the experiment.

Each voluntary subject, after being informed on the regulation of data protection, must:

- declare the absence of respiratory symptoms or fever above 37.5° C at the date of the experiment and during the previous three days;
- declare of not having been posed in quarantine or isolation at home during the last 14 days;
- declare of not having been in contact with positive people, to the best of his/her knowledge, during the last 14 days;
- always wear the surgical mask provided by the University administration for the entire duration of the experiment;
- always observe a distance of at least 1 meter (including movement space) with the present team member of the Robotics Laboratory.

The self-declaration attached to this safety protocol, attesting what above listed, must be pre-filled by the voluntary subject and placed by the subject in a special binder placed at the entrance of the Robotics Laboratory.

The entrance to the Robotics Laboratory is limited to the voluntary subject and to the team member of the Robotics Laboratory delegated to supervise the experiment.

Upon entering the University premises, the body temperature of the voluntary subject will be measured at specific sites identified and communicated in advance. The body temperature measurement will be carried out on the voluntary subjects and on all team members of the Robotics Laboratory.

In the event that the respiratory or febrile symptoms occur during the stay of the voluntary subject in the Robotics Laboratory or in the University premises, the procedures in accordance with the "Regulations and Specific Protocols adopted within the University of Palermo for the contrast and

containment of the spread of SARS-CoV-2 Virus" - UPDATE PHASE 3, will be followed.

The voluntary subject must always respect the distance of at least 1 meter (including the space of movement) with the team member of the Robotics Laboratory. No additional protective devices are required.

14.3.4 Organization of the Robotics Laboratory and preventive measures for conducting the experiments of interaction between people and robots

Considering the structural characteristics of the Robotics Laboratory, dedicated input and output paths will be provided, clearly identified with appropriate "Input" and "Output" signs, in order to prevent the risk of interference/assembly of people.

The team member of the Robotics Laboratory delegated to supervise the experiment will ensure that the entrance door and all the windows of the Laboratory always remain open at all times in order, among other things, to ensure the expected air exchange. The room intended for the interaction sessions provides a space of 42 square meters, which is large enough to accommodate 2 people, and that it allows a distance of not less than 1 meter (including the space of movement), and ventilation surfaces to allow the expected air exchange.

A regular and sufficient exchange of air will be guaranteed in the Laboratory favoring, in any case possible, natural ventilation.

The voluntary subject and the team member of the Robotics Laboratory will have to proceed to the preventive hand sanitization during the access phase; therefore, the use of gloves is not necessary.

In accordance with the provisions of the Director of the Department of Engineering, the building 6, where the Robotics Laboratory is located, provides the infirmary room PT 059 as a room dedicated to the reception and isolation of any subjects who may show respiratory or febrile symptoms. In this case, the subject will be immediately taken to the aforementioned room waiting for the arrival of the necessary assistance alerted according to the indications of the health authority in charge.

14.3.5 Total duration of the interaction experiment between the subject and the robot

Each interaction session between the voluntary subject and the robot cannot last more than 15 minutes.

The next session will be spaced temporally from the previous one by at least 15 minutes, necessary to complete the cleaning and sanitation operations.

14.3.6 2.6 Voluntary subjects with disabilities

At this stage, voluntary subjects with certified disabilities will not be allowed to participate at the experiment, because the room space of the Robotics Laboratory does not allow the presence of a third person as an assistant to the voluntary subject.

14.3.7 2.7 Measures for the person in charge of the procedure

In order to ensure the tracking of close contacts in the event of suspicious or confirmed cases of Covid-19, the Principal Investigator is responsible for the storage and custody for at least 14 days from the date of the experiment, of the anti-COVID self-declaration produced by voluntary subjects and team members of the Robotics Laboratory.

UNIVERSITÀ DEGLI STUDI DI PALERMO

Director: Prof. Giovanni Perrone

dipartimento di ingegneria unipa

SELF-DECLARATION TEMPLATE

The Underwriter

Surname _____ First name_____

Place of birth _____Date of birth _____

Identification document _____

in accessing the premises of the University of Palermo, under his/her own responsibility declares what follows:

- of not having respiratory symptoms or fever above 37.5° C at the date of today and during the previous three days;

- of not having posed in quarantine or isolation at home during the last 14 days;

- of not having been in contact with positive people, to the best of his/her knowledge, during the last 14 days.

This self-declaration is issued as a prevention measure related to SARS CoV-2 pandemic emergency.

Place and date _____

Readable signature _____

Viale delle Scienze, Building 8 - 90128 Palermo - C.F. 80023730825 - P.IVA 00605880822
Management: (+39) 091 23863701 - Administration:(+39) 091 23864623- Fax: (+39) 091 238 60840
e-mail: dipartimento.ingegneria@unipa.it- PEC: dipartimento.ingegneria@cert.unipa.it - IPA code: 0IHPZZ

Figure 14.1: Self-declaration template.

Bibliography

[1] Hussein A Abbass, Jason Scholz, and Darryn J Reid. *Foundations of trusted autonomy*, volume 117. Springer, 2018.

[2] Gene M Alarcon, Joseph B Lyons, and James C Christensen. The effect of propensity to trust and familiarity on perceptions of trustworthiness over time. *Personality and Individual Differences*, 94:309–315, 2016.

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

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

[5] Ben Alderson-Day and Charles Fernyhough. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141, 05 2015.

[6] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *PSYCHOLOGICAL REVIEW*, 111:1036–1060, 2004.

[7] Konstantine Arkoudas and Selmer Bringsjord. Toward formalizing common-sense psychology: An analysis of the false-belief task. In Tu-Bao Ho and Zhi-Hua Zhou, editors, *PRICAI 2008: Trends in Artificial Intelligence*, pages 17–29, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[8] 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.

[9] Bernard Baars. In the theater of consciousness: The workspace of the mind. 01 1997.

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

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

[12] Alan Baddeley and Graham James Hitch. *Working memory*, volume 8, pages 47–90. Academic Press, 1974.

[13] Christian Balkenius, Trond A Tjøstheim, Birger Johansson, and Peter Gärdenfors. From focused thought to reveries: a memory system for a conscious robot. *Frontiers in Robotics and AI*, 5:29, 2018.

[14] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[15] Maria Beazley, Carol Glass, Dianne Chambless, and Diane Arnkoff. Cognitive self-statements in social phobia: A comparison across three types of social situations. *Cognitive Therapy and Research*, 25:781–799, 12 2001.

[16] Maria B Beazley, Carol R Glass, Dianne L Chambless, and Diane B Arnkoff. Cognitive self-statements in social phobia: A comparison across three types of social situations. *Cognitive therapy and Research*, 25(6):781–799, 2001.

[17] Yochanan E. Bigman and Kurt Gray. People are averse to machines making moral decisions. *Cognition*, 181(C), Dec 2018.

[18] Barry B. Blake. Russell s. tomlin, basic word order. functional principles. london: Croom helm, 1986. pp. 308. *Journal of Linguistics*, 24(1):213–217, 1988.

[19] Melanie Boly, Marcello Massimini, Naotsugu Tsychiya, Bradley Postle, Christof Koch, and Giulio Tononi. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? clinical and neuroimaging evidence, 03 2017.

[20] 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.

[21] Michael W Boyce, Jessie YC Chen, Anthony R Selkowitz, and Shan G Lakhmani. Effects of agent transparency on operator trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 179–180, 2015.

[22] Selmer Bringsjord, Naveen Sundar G., Dan Thero, and Mei Si. Akratic robots and the computational logic thereof. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*, ETHICS '14. IEEE Press, 2014.

[23] Selmer Bringsjord and Naveen Sundar Govindarajulu. Given the web, what is intelligence, really? *Metaphilosophy*, 43(4):464–479, 2012.

[24] Selmer Bringsjord, John Licato, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Atriya Sen. Real robots that pass human tests of self-consciousness. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 498–504. IEEE, 2015.

[25] Thomas M Brinthaupt, Michael B Hein, and Tracey E Kramer. The self-talk scale: Development, factor analysis, and validation. *Journal of personality assessment*, 91(1):82–92, 2009.

[26] Randy L. Buckner, Jessica R. Andrews-Hanna, and Daniel L. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124:1–38, 2008.

[27] Arnold Herbert Buss. *Self consciousness and social anxiety*. San Francisco : W.H. Freeman, 1980. Includes index.

[28] Peter Carruthers, Logan Fletcher, and J. Brendan Ritchie. The evolution of self-knowledge. *Philosophical Topics*, 40(2):13–37, 2012.

[29] Charles S. Carver and Michael F. Scheier. Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology*, 36(3):324–332, March 1978.

[30] Rosanova M. Sarasso S. Fecchio M. Napolitani M. et al. Casarotto S., Comanducci A. Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of neurology*, 80, 09 2016.

[31] Justine Cassell and W. Timothy Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, pages 89–132, 2003.

[32] A. Chella, F. Lanza, A. Pipitone, and V. Seidita. Human-robot teaming: Perspective on analysis and implementation issues. In *Conference of 5th Italian Workshop on Artificial Intelligence and Robotics, AIRO 2018*, volume 2352. CEUR-WS, 2019.

[33] Antonio Chella, Angelo Cangelosi, Giorgio Metta, and Selmer Bringsjord. Consciousness in humanoid robots. *Frontiers in Robotics and AI*, 6:17, 2019.

[34] Antonio Chella, Angelo Cangelosi, Giorgio Metta, and Selmer Bringsjord. Editorial: Consciousness in humanoid robots. *Frontiers in Robotics and AI*, 6:17, 2019.

[35] Antonio Chella, Francesco Lanza, Arianna Pipitone, and Valeria Seidita. Knowledge acquisition through introspection in human-robot cooperation. *Biologically inspired cognitive architectures*, 25:1–7, 2018.

[36] Antonio Chella and Riccardo Manzotti. *Artificial Consciousness*, pages 637–671. 12 2011.

[37] Antonio Chella and Arianna Pipitone. A cognitive architecture for inner speech. *Cognitive Systems Research*, 59, 10 2019.

[38] Antonio Chella and Arianna Pipitone. The inner speech of the idyot: Comment on "creativity, information, and consciousness: The information dynamics of thinking" by geraint a. wiggins. *Physics of life reviews*, 2019.

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

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

[41] 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.

[42] 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.

[43] Sanjib Chowdhury. The role of affect-and cognition-based trust in complex knowledge sharing. *Journal of Managerial issues*, pages 310–326, 2005.

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

[45] Robert Clowes. A self-regulation model of inner speech and its role in the organisation of human conscious experience. *Journal of Consciousness Studies*, 14(7):59–71, 2007.

[46] Robert Clowes and Anthony F Morse. Scaffolding cognition with words. 2005.

[47] COMEST/Unesco. Report of comest on robotics ethics, 2017.

[48] 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.

[49] Cynthia L Corritore, Beverly Kracher, and Susan Wiedenbeck. Online trust: concepts, evolving themes, a model. *International journal of human-computer studies*, 58(6):737–758, 2003.

[50] Fergus Craik and Robert Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11:671–, 12 1972.

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

[52] 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.

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

[54] A. R. Damasio. *Self comes to mind: constructing the conscious brain.* Pantheon, New York, 2010.

[55] Napha Daosodsai and Thavida Maneewarn. Fuzzy based emotion generation mechanism for an emoticon robot. In *2013 13th International Conference on Control, Automation and Systems (ICCAS 2013)*, pages 1073–1078, 2013.

[56] Christopher G. Davey, Jesus Pujol, and Ben J. Harrison. Mapping the self in the brain's default mode network. *NeuroImage*, 132:390 – 397, 2016.

[57] Martin F. Davies. Mirror and camera self-focusing effects on complexity of private and public aspects of identity. *Perceptual and Motor Skills*, 100(3):895–898, 2005.

[58] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358:486–492, 10 2017.

[59] Bryan T. Denny, Hedy Kober, Tor D. Wager, and Kevin N. Ochsner. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8):1742–1752, 2012. PMID: 22452556.

[60] S Kate Devitt. Trustworthiness of autonomous systems. In *Foundations of trusted autonomy*, pages 161–184. Springer, Cham, 2018.

[61] Carl F DiSalvo, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler. All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 321–326, 2002.

[62] Brian R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3):177–190, 2003. Socially Interactive Robots.

[63] Robert Duncan and James Cheyne. Incidence and functions of self-reported private speech in young adults: A self-verbalization questionnaire. *Canadian Journal of Behavioural Science*, 31:133–136, 04 1999.

[64] S. Duval and R.A. Wicklund. *A theory of objective self awareness*. Social psychology. Academic Press, 1972.

[65] Jeffrey A. Edlund, Nicolas Chaumont, Arend Hintze, Christof Koch, Giulio Tononi, and Christoph Adami. Integrated information increases with fitness in the evolution of animats. *PLOS Computational Biology*, 7(10):1–13, 10 2011.

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

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

[68] Michael Emerson and Akira Miyake. The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, 48:148–168, 01 2003.

[69] Friederike Eyssel, Laura De Ruiter, Dieta Kuchenbrandt, Simon Bobinger, and Frank Hegel. 'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 125–126. IEEE, 2012.

[70] Allan Fenigstein, Michael F. Scheier, and Arnold H. Buss. Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43(4):522–527, aug 1975.

[71] Charles Fernyhough. *The voices within: The history and science of how we talk to ourselves*. Basic Books, 2016.

[72] Luciano Floridi. Consciousness, agents and the knowledge game. *Minds and machines*, 15(3-4):415–444, 2005.

[73] 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.

[74] Stan Franklin. A conscious artifact? *Journal of Consciousness Studies*, 10(4-5):47–66, 2003.

[75] Stan Franklin, Tamas Madl, Sidney D'mello, and Javier Snaider. Lida: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, 6(1):19–41, 2013.

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

[77] Miriam Gade and Marko Paelecke. Talking matters - evaluative and motivational inner speech use predicts performance in conflict tasks. *Scientific Reports*, 9, 07 2019.

[78] Miriam Gade and Marko Paelecke. Talking matters–evaluative and motivational inner speech use predicts performance in conflict tasks. *Scientific reports*, 9(1):1–8, 2019.

[79] Miriam Gade and Marko Paelecke. Talking matters–evaluative and motivational inner speech use predicts performance in conflict tasks. *Scientific reports*, 9(9531), 2019.

[80] Gerhard Gentzen. Investigations into logical deduction. *American Philosophical Quarterly*, 1(4):288–306, 1964.

[81] 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.

[82] Sharon Geva, P Simon Jones, Jenny T Crinion, Cathy J Price, Jean-Claude Baron, and Elizabeth A Warburton. The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. *Brain*, 134(10):3071–3082, 2011.

[83] Anna Gorbenko, Vladimir Popov, and Andrey Sheka. Robot self-awareness: Exploration of internal states. *Applied Mathematical Sciences*, 6(14):675–688, 2012.

[84] Naveen Sundar Govindarajulu and Selmer Bringsjord. On automating the doctrine of double effect. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4722–4730, 2017.

[85] Anthony M Grant, John Franklin, and Peter Langford. The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality: an international journal*, 30(8):821–835, 2002.

[86] 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.

[87] Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *science*, 315(5812):619–619, 2007.

[88] Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *science*, 315(5812):619–619, 2007.

[89] Michael SA Graziano. The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, 4:60, 2017.

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

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

[92] David K. Grunberg, Alyssa M. Batula, Erik M. Schmidt, and Youngmoo E. Kim. Affective gesturing with music mood recognition. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 343–348, 2012.

[93] M. Görner, R. Haschke, H. Ritter, and J. Zhang. Moveit! task constructor for task-level motion planning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 190–196, 2019.

[94] Sami J Habib, Paulvanna N Marimuthu, Pravin Renold, and Balaji Ganesh Athi. Development of self-aware and self-redesign framework for wireless sensor networks. In *World Conference on Information Systems and Technologies*, pages 438–448. Springer, 2019.

[95] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[96] James Hardy. Speaking clearly: A critical review of the self-talk literature. *Psychology of Sport and Exercise*, 7(1):81–97, 2006.

[97] Kerstin Sophie Haring, Yoshio Matsumoto, and Katsumi Watanabe. How do people perceive and trust a lifelike robot. In *Proceedings of the world congress on engineering and computer science*, volume 1. Citeseer, 2013.

[98] Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, June 1949.

[99] Pamela J Hinds, Teresa L Roberts, and Hank Jones. Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, 19(1-2):151–181, 2004.

[100] Jochen Hirth, Norbert Schmitz, and Karsten Berns. Playing tangram with a humanoid robot. In *ROBOTIK 2012; 7th German Conference on Robotics*, pages 1–6, 2012.

[101] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[102] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[103] Owen Holland. *Machine consciousness*. Oxford University Press, 2009.

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

[105] Don Howard and Laurel Riek. A code of ethics for the human-robot interaction profession, 2015.

[106] Ya-Ting Hsu, Fang-Yie Leu, Jung-Chun Liu, Yi-Li Huang, and William Cheng-Chung Chu. The simulation of a robot with emotions implemented with fuzzy logic. In *2013 Eighth International Conference on Broadband and Wireless Computing, Communication and Applications*, pages 382–386, 2013.

[107] Russell T Hurlburt, Christopher L Heavey, and Jason M Kelsey. Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 22(4):1477–1494, 2013.

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

[109] ISO/TS:15066. Robots and robotic devices, 03 2016.

[110] 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.

[111] Misbah Javaid, Vladimir Estivill-Castro, and Rene Hexel. Enhancing humans trust and perception of robots through explanations. *Proceedings of the ACHI*, 2020.

[112] Felix Jimenez, Tomohiro Yoshikawa, Takeshi Furuhashi, and Masayoshi Kanoh. A proposal of model of emotional expressions for robot learning with human. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–5, 2015.

[113] Han Jing, Xie Lun, Li Dan, He Zhijie, and Wang Zhiliang. Cognitive emotion model for eldercare robot in smart home. *China Communications*, 12(4):32–41, 2015.

[114] Jeffrey A Joireman, Les Parrott III, and Joy Hammersla. Empathy and the self-absorption paradox: Support for the distinction between self-rumination and self-reflection. *Self and Identity*, 1(1):53–65, 2002.

[115] 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.

[116] BE Juel, R Comolatti, G Tononi, and L Albantakis. When is an action caused from within? quantifying the causal chain leading to actions in simulated agents. arxiv e-prints, 2019.

[117] JA Kelso. " on the self-organizing origins of agency": Erratum. 2016.

[118] Philip C. Kendall. Methodology and cognitive—behavioral assessment. *Behavioural and Cognitive Psychotherapy*, 11(4):285–301, 1983.

[119] Yasuo Kinouchi and Kenneth James Mackin. A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot. *Frontiers in Robotics and AI*, 5:30, 2018.

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

[121] Nikolaus Kriegeskorte and Pamela Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21, 08 2018.

[122] E Kross and O Ayduk. Self-distancing: Theory, research, and current directions. In *Advances in experimental social psychology*, volume 55, pages 81–136. Elsevier, 2017.

[123] Benjamin Kuipers. Drinking from the firehose of experience. *Artificial intelligence in medicine*, 44(2):155–170, 2008.

[124] 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), 2017.

[125] 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*, 38:13–26, 2017.

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

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

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

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

[130] Christian Lebiere and J. Anderson. A connectionist implementation of the act-r production system. *Department of Psychology*, 01 2008.

[131] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[132] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[133] Won Hyong Lee, Jeong Woo Park, Woo Hyun Kim, Ju Chang Kim, and Myung Jin Chung. Robot's emotion generation model for transition and diversity using energy, entropy, and homeostasis concepts. In *2010 IEEE International Conference on Robotics and Biomimetics*, pages 555–560, 2010.

[134] William Lehman. Trust in automation as a function of transparency and teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):78–82, 2019.

[135] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6(2):104, 2000.

[136] J David Lewis and Andrew Weigert. Trust as a social reality. *Social forces*, 63(4):967–985, 1985.

[137] Michael Lewis and Douglas Ramsay. Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child development*, 75(6):1821–1831, 2004.

[138] Michael Lewis, Katia Sycara, and Phillip Walker. The role of trust in human-robot interaction. In *Foundations of trusted autonomy*, pages 135–159. Springer, Cham, 2018.

[139] Dingjun Li, PL Patrick Rau, and Ye Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, 2010.

[140] Angelica Lim and Hiroshi G. Okuno. The mei robot: Towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, 6(2):126–138, 2014.

[141] 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.

[142] 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.

[143] Stacy Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10:70–90, 03 2009.

[144] 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.

[145] 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.

[146] Fernando Martínez-Manrique and Agustín Vicente. The activity view of inner speech. *Frontiers in Psychology*, 6:232, 2015.

[147] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[148] John McCarthy. Making robots conscious of their mental states. In *Machine Intelligence 15*, pages 3–17, 1995.

[149] Drew V McDermott. *Mind and mechanism*. MIT Press, 2001.

[150] Alison McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition, 2019.

[151] Stephanie M Merritt and Daniel R Ilgen. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210, 2008.

[152] Matthias Michel, Diane Beck, Ned Block, Hal Blumenfeld, Richard Brown, David Carmel, Marisa Carrasco, Mazviita Chirimuuta, Marvin Chun, Axel Cleeremans, Stanislas Dehaene, Stephen M. Fleming, Chris Frith, Patrick Haggard, Biyu J. He, Cecilia Heyes, Melvyn A. Goodale, Liz Irvine, Mitsuo Kawato, Robert Kentridge, Jean Remi King, Robert T. Knight, Sid Kouider, Victor Lamme, Dominique Lamy, Hakwan Lau, Steven Laureys, Joseph LeDoux, Ying Tung Lin, Kayuet Liu, Stephen L.and Susana Martinez-Conde Macknik, George A. Mashour, Lucia Melloni, Lisa Miracchi, Myrto Mylopoulos, Lionel Naccache, Adrian M. Owen, Richard E. Passingham, Luiz Pessoa, Megan A.K. Peters, Dobromir Rahnev, Tony Ro, David Rosenthal, Yuka Sasaki, Claire Sergent, Guillermo Solovey, Nicholas D.and Anil Seth Schiff, Catherine Tallon-Baudry, Marco Tamietto, Frank Tong, Simon van Gaal, Alexandra Vlassova, Takeo Watanabe, Josh Weisberg, Karen Yan, and Masatoshi Yoshida. Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour*, 3(2):104–107, 2019.

[153] Marco Mirolli and Domenico Parisi. Talking to oneself as a selective pressure for the emergence of language. In *The evolution of language*, pages 214–221. World Scientific, 2006.

[154] Marco Mirolli and Domenico Parisi. Talking to oneself as a selective pressure for the emergence of language. pages 214–221, 03 2006.

[155] Ranjeev Mittu, Gavin Taylor, Don Sofge, and William F Lawless. Introduction to the symposium on ai and the mitigation of human error. In *2016 AAAI Spring Symposium Series*, 2016.

[156] Nilly Mor and Jennifer Winquist. Self-focused attention and negative affect: a meta-analysis. *Psychological bulletin*, 128(4):638, 2002.

[157] 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.

[158] A Morin. The self-reflective functions of inner speech: Thirteen years later, 2018.

[159] A Morin, C Duhnych, F Racy, J Hagerty, and J Patton. Inner speech in humans. In *Talk presented at the Inner Speech in Humans and Robots Workshop. Sicily: University of Palermo*, 2019.

[160] Alain Morin. Characteristics of an effective internal dialogue in the acquisition of self-information. *Imagination, Cognition and Personality*, 15(1):45–58, 1995.

[161] Alain Morin. Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies*, 12(4-5):115–134, 2005.

[162] Alain Morin. Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and cognition*, 15(2):358–371, 2006.

[163] Alain Morin. Inner speech. 2009.

[164] 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.

[165] Alain Morin. Self-awareness part 1: Definition, measures, effects, functions, and antecedents. *Social and personality psychology compass*, 5(10):807–823, 2011.

[166] Alain Morin. Self-awareness part 1: Definition, measures, effects, functions, and antecedents. *Social and Personality Psychology Compass*, 5(10):807–823, 2011.

[167] Alain Morin. *Inner Speech*. 12 2012.

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

[169] 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.

[170] Alain Morin and Famira Racy. Dynamic self-processes. *Handbook of personality dynamics and processes. Amsterdam, the Netherlands: Elsevier*.

[171] Ladislas Nalborczyk, Marcela Perrone-Bertolotti, Céline Baeyens, Romain Grandchamp, Mircea Polosan, Elsa Spinelli, Ernst HW Koster, and Helene Loevenbruck. Orofacial electromyographic correlates of induced verbal rumination. *Biological psychology*, 127:53–63, 2017.

[172] Takuya Niikawa. A map of consciousness studies: Questions and approaches. *Frontiers in Psychology*, 11, 2020.

[173] D. A. Norman, A. Ortony, and D. M. Russell. Affect and machine design: Lessons for the development of autonomous machines. *IBM Systems Journal*, 42(1):38–44, 2003.

[174] Rony Novianto. *Flexible attention-based cognitive architecture for robots*. PhD thesis, 2014.

[175] Rony Novianto and Mary-Anne Williams. The role of attention in robot self-awareness. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1047–1053. IEEE, 2009.

[176] 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.

[177] 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.

[178] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol*, 10(5):e1003588, 2014.

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

[180] Yigit Oktar, Erdem Okur, and Mehmet Turkan. The mimicry game: Towards self-recognition in chatbots. *arXiv preprint arXiv:2002.02334*, 2020.

[181] Richard Pak, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9):1059–1072, 2012.

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

[183] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

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

[185] 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.

[186] Marcela Perrone-Bertolotti, Lucile Rapin, J-P Lachaux, Monica Baciu, and Hélène Loevenbruck. 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.

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

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

[189] Arianna Pipitone, Antonio Chella, Valeria Seidita, Francesco Lanza, Antonio Chella, and Valeria Seidita. Inner speech for a self-conscious robot. 2018.

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

[191] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.

[192] James A Reggia. The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, 44:112–131, 2013.

[193] Michael Rescorla. The language of thought hypothesis. 2019.

[194] Philippe Rochat. Five levels of self-awareness as they unfold early in life. *Consciousness and cognition*, 12(4):717–731, 2003.

[195] Stephan Alexander Rompf. *Trust and rationality: an integrative framework for trust research*. Springer, 2014.

[196] Julian B Rotter. Interpersonal trust, trustworthiness, and gullibility. *American psychologist*, 35(1):1, 1980.

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

[198] Sebastien Saint-Aime, Celine Jost, Brigitte Le-Pevedic, and Dominique Duhaut. Dynamic behaviour conception for emi companion robot. In *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, pages 1–8, 2010.

[199] Ben Salem. A framework for a robot's emotions engine. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 635–640, 2017.

[200] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013.

[201] Tracy L Sanders, Keith MacArthur, William Volante, Gabriella Hancock, Thomas MacGillivray, William Shugars, and PA Hancock. Trust

and prior experience in human-robot interaction. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 61, pages 1809–1813. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[202] 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.

[203] Kristin E Schaefer. Measuring trust in human robot interactions: Development of the "trust perception scale-hri". In *Robust Intelligence and Trust in Autonomous Systems*, pages 191–218. Springer, 2016.

[204] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400, 2016.

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

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

[207] Matthias Scheutz. Artificial emotions and machine consciousness. *The Cambridge Handbook of Artificial Intelligence*, pages 247–266, 2014.

[208] Anil K Seth. Measuring autonomy and emergence via granger causality. *Artificial life*, 16(2):179–196, 2010.

[209] Murray Shanahan. *The Event Calculus Explained*, pages 409–430. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[210] Murray Shanahan. Global access, embodiment and the conscious subject. *Journal of Consciousness Studies*, 12(12):46–66, 2005.

[211] Murray Shanahan. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition*, 15(2):433–449, 2006.

[212] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226. IEEE, 2010.

[213] Paul J Silvia and Maureen E O'Brien. Self-awareness and constructive functioning: Revisiting "the human dilemma". *Journal of Social and Clinical Psychology*, 23(4):475–489, 2004.

[214] J David Smith. The study of animal metacognition. *Trends in cognitive sciences*, 13(9):389–396, 2009.

[215] Alex W Stedmon, Sarah Sharples, Robert Littlewood, Gemma Cox, Harshada Patel, and John R Wilson. Datalink in air traffic management: Human factors issues in communications. *Applied ergonomics*, 38(4):473–480, 2007.

[216] Luc Steels. Basics of fluid construction grammar. *Constructions and frames*, 9(2):178–255, 2017.

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

[218] John P. Sullins. The role of consciousness and artificial phronēsis in ai ethical reasoning. In A. Chella, D. Gamez, P. Lincoln, R. Manzotti, and J. Pfautz, editors, *TOCAIS 2019, Towards Conscious AI Systems*, volume 2287 of *Papers from the 2019 AAAI Spring Symposium*. CEUR-WS.org, http://ceur-ws.org/Vol-2287/paper11.pdf, 2019.

[219] John P. Sullins. Artificial phronesis: What it is and what it is not. In Emanuele Ratti and Thomas A. Stapleford, editors, *Science, Technology, and Virtues: Contemporary Perspectives.*, chapter 7, pages 136–146. Oxford University Press, 2021.

[220] Tappan. Mediated moralities: Sociocultural approaches to moral development, 2005.

[221] Max Tegmark. Consciousness is a state of matter, like a solid or gas. *New Scientist*, 222(2964):28–31, 2014.

[222] Max Tegmark. Consciousness as a state of matter. *Chaos, Solitons and Fractals*, 76:238–270, Jul 2015.

[223] Giulio Tononi and Gerald M Edelman. Consciousness and complexity. *science*, 282(5395):1846–1851, 1998.

[224] Giulio Tononi and Gerald M. Edelman. Consciousness and complexity. *Science*, 282(5395):1846–1851, 1998.

[225] Paul D Trapnell and Jennifer D Campbell. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *Journal of personality and social psychology*, 76(2):284, 1999.

[226] Alexa M Tullett and Michael Inzlicht. The voice of self-control: Blocking the inner voice increases impulsive responding. *Acta psychologica*, 135(2):252–256, 2010.

[227] Frank Van Der Velde. In situ representations and access consciousness in neural blackboard or workspace architectures. *Frontiers in Robotics and AI*, 5:32, 2018.

[228] Michelle ME van Pinxteren, Ruud WH Wetzels, Jessica Rüger, Mark Pluymaekers, and Martin Wetzels. Trust in humanoid robots: implications for services marketing. *Journal of Services Marketing*, 2019.

[229] 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.

[230] Lev Vygotsky. Thought and language. 1962.

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

[232] Lev S Vygotsky. *Thought and language*. MIT press, 2012.

[233] Matt Webster, David Western, Dejanira Araiza-Illan, Clare Dixon, Kerstin Eder, Michael Fisher, and Anthony G. Pipe. An assurance-based approach to verification and validation of human-robot teams. *CoRR*, abs/1608.07403, 2016.

[234] Astrid Weiss and Christoph Bartneck. Meta analysis of the usage of the godspeed questionnaire series. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 381–388. IEEE, 2015.

[235] Adam Ed Winsler, Charles Ed Fernyhough, and Ignacio Ed Montero. *Private speech, executive functioning, and the development of verbal self-regulation.* Cambridge University Press, 2009.

[236] Peijun Ye, Tao Wang, and Fei-Yue Wang. A survey of cognitive architectures in the past 20 years. *IEEE transactions on cybernetics*, 48(12):3280–3290, 2018.

[237] Ethan Zell, Amy Beth Warriner, and Dolores Albarracín. Splitting of the mind: When the you i talk to is me and needs commands. *Social psychological and personality science*, 3(5):549–555, 2012.