FACSIMILE FRONTISPIECE DOCTORAL THESIS

(Deliver the digitally signed title page of the thesis in PADES format)

UNIVERSITY OF BARI ALDO MORO

Department of Humanities Research and Innovation

(*Indicate the Department where the PhD is administered*)

PH.D. IN Historical and Philosophical Heritages for Sustainable Innovation

CYCLE XXXVIII

Academic Discipline History of Science

(If the thesis refers to more than one scientific disciplinary sector, indicate the prevailing one. Refer to, https://www.mur.gov.it/sites/default/files/2024-05/Decreto%20Ministeriale%20n.%20639%20del%2002-05-2024%20-20Allegato%20B.pdf)

THESIS TITLE (approved by the Academic Board)

Afectivity in Computational and Robotic Systems: History, Epistemology and Applications

PhD student: Dr. Ivano Zanzarella

Supervisor : Prof. Dr. Francesco Paolo de Ceglia

Coordinator: Prof. Dr. Francesco Paolo de Ceglia

FINAL EXAM 2026

Funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP H91I22000920007 *Investment 4.1 Ministerial Decree no. 351 of 9 April 2022 - National Recovery and Resilience Plan. Project:* "The Imitation Game". Humanoid Robots and the Simulation of Emotion for a "Human" and Inclusive Technology *(MANDATORY approved by the MUR and present on the cineca platform and approved by the Academic Board)*



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

University of Bari "Aldo Moro"
Department of Human Research and Innovation

PhD Program in History and Philosophy of Science

Affectivity in Computational and Robotic Systems: History, Epistemology and Applications

A thesis presented for the degree of
Doctor of Philosophy by

Ivano Zanzarella

January 14, 2026

Supervisor
Prof. Francesco Paolo de Ceglia

Advisor
Prof. Francesco Marrone

Università di Bari “Aldo Moro”
Dipartimento di Ricerca e Innovazione Umanistica

Corso di Dottorato in Storia e Filosofia della Scienza

Affectivity in Computational and Robotic Systems: History, Epistemology and Applications

Tesi di dottorato
presentata da

Ivano Zanzarella

14 gennaio 2026

Relatore
Prof. Francesco Paolo de Ceglia

Co-Relatore
Prof. Francesco Marrone

Dottorato in Patrimoni Storici e Filosofici per una Innovazione Sostenibile

XXXVIII ciclo

Dipartimento di Ricerca e Innovazione Umanistica
Università degli Studi di Bari "Aldo Moro"
A.A. 2024/2025

Coordinatrice: Prof. Francesco Paolo de Ceglia

Curriculum 1 (Discipline filosofiche), GSD: PHIL-02/B (ex SSD M-STO/05, Storia della Scienza e delle Tecniche)

Progetto PNRR (*ex* DM 351/2022): Il gioco dell'imitazione. Robot umanoidi e simulazione dell'emozione per una tecnologia umana e inclusiva

Codice borsa: 38-412-02-DOT22NZTHZ-690
CUP: H91I22000920007

Dichiarazione di originalità

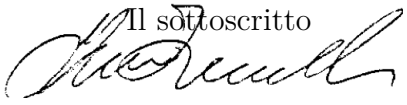
Con la presente, il sottoscritto Zanzarella Ivano, nato a Molfetta (BA) il 02/08/1995 e iscritto al corso di dottorato in "Patrimoni storici e filosofici per un'innovazione sostenibile" (curriculum filosofico, tesi in storia della scienza e delle tecniche, GSD: PHIL-02/B, ex SSD M-STO/05, Storia della Scienza e delle Tecniche),

DICHIARA

sotto la propria responsabilità e consapevole delle conseguenze giuridiche derivanti da eventuali violazioni delle normative relative al plagio e alla proprietà intellettuale, che questa tesi, dal titolo *Affectivity in Computational and Robotic Systems: History, Epistemology and Applications* [L'affettività nei sistemi computazionali e robotici: storia, epistemologia e applicazioni], è frutto del proprio lavoro originale. Tutti i contenuti utilizzati sono correttamente citati e non sono stati copiati da fonti non dichiarate. La ricerca è stata condotta in modo autonomo, utilizzando solo le fonti riportate nel testo e nella bibliografia. Inoltre, la tesi non è stata precedentemente presentata per altri titoli accademici. Eventuali collaborazioni sono state debitamente indicate e il sottoscritto si assume piena responsabilità del contenuto della tesi.

Bari, 14 gennaio 2026

Il sottoscritto



Acknowledgements

It is difficult, after three intense years of work, professional experiences, and private life events, to dedicate a few lines of gratitude to those people without whom, very probably, this dissertation would not exist in its present form. This difficulty stems not only from my reserved nature, but also from the fact that the people who have contributed to making this work possible, directly or indirectly, in various ways and forms, are countless.

I would firstly like to thank my supervisor, Francesco Paolo de Ceglia, for his encouragement, support, esteem, and patience. His guidance has been essential throughout this process, helping me to stay focused and motivated. I also wish to thank my co-supervisor, Francesco Marrone, with whom I had the privilege of sharing many inspiring moments of intellectual exchange, teamwork, and open dialogue —an invaluable source of strength, especially during the most challenging moments.

Sometimes one feels that it is all over, that there is nothing left but to give up and go home. On this journey, however, Edoardo made all the difference, pulling me back into the battlefield —back to the true enjoyment of philosophical questions, debates, analyses, and collaboration. I owe much of this work to him: from our long brainstorming sessions on robotics and cognitive science between Milan and Apulia, to the beautiful moments spent together —unforgettable treasures of optimism, perseverance, and self-trust for now and the future.

Alongside him, I would also like to thank Viola and the META Group at the Politecnico di Milano, as well as Silvia, Gilda, Leonardo, Nicola, Cecilia and all the researchers at the RobotiCSS Lab of the University of Milano-Bicocca, for their warm welcome, friendship, and stimulating professional collaboration.

I would not have reached the analytical and scientific depth of this work without the enriching exchanges I had the opportunity to engage in at Pompeu Fabra University and the LOGOS Research Group in Barcelona —a marvelous and culturally dynamic city where I had the pleasure of spending several months as a visiting PhD student. I am

grateful to Daniele for his friendly support, openness, and insightful advice on both my work and career in the history of science. I would also like to thank Sven and José for their warm welcome to the LOGOS group, as well as Manuel García-Carpintero, Josep Macià, David Pineda, and Manolo Martínez for the stimulating discussions on my research. Special thanks go to Thomas Sturm, whose historical and philosophical sensibility contributed significantly to the conceptual and methodological development of key parts of this dissertation. I also wish to thank Luca Bonatti from UPF's Center for Brain and Cognition and Jordi Vallverdú from UAB, with whom I spent two of the most intellectually stimulating afternoons of my Spanish visit.

I also want to thank Margherita, Elena, Duccio, Auke, Pol, Giovanni, Martí, Oriol, Simone, Taís, Vini, Andrea, Daphne, Marc, Karl, Riin, Nikki, Lucía, Alfonso, Txú, Sofia, Marta, Josef, Pasquale, Simone, Johannes, and all the friends and colleagues I had the fortune to meet and spend time, words, and laughter with at the University of Barcelona. They were true companions during my stay —friends who shared my wavelength, who brought color to my experience, and who were invaluable sources of creative ideas.

Yet, somewhat sadly, research stays eventually come to an end, and when they do, not everyone has the privilege of returning to the Seminario di Storia della Scienza. Through challenges, responsibilities, and even punches, this center has helped me grow as a researcher, strengthening my skills in academic politics and diplomacy, scientific communication, organization, bureaucracy, and much more. But above all, it has enriched me with colleagues who, over time, I have been fortunate to start calling friends —people who have supported me in everything, helped me through the most difficult times, listened to my frustrations, and celebrated my successes in a way that I am probably not capable of myself. I especially thank Fabio, Eleonora, Angelica, Fabio, Francesca, Simona, Stefano, Virginia, Betty, Lorenzo, Sara, Andrea, and the members with whom I spent most of my time during the last phases of this journey: Ilaria, Iolanda, Alessandro, as well as Alessio —a man of another time and one of the few people with whom I can discuss logic, organ music, and climbing all at once —and Roberta, who solidly supported me with her wise and caring advice and presence. I also thank Prof. Liborio Dibattista, who, after all these years, remains one of my greatest points of reference and who contributed to the methodological development of this work with his philosophical approach and historical depth.

There is one man from the Seminario for whom I struggle to find the right words. This man is Pasquale, with whom I enrolled in the same PhD program in 2022, focusing on closely related topics. Pasquale has been a colleague, a discussant, a mentor, a friend, a travel companion —the older brother I never had. We shared much of the doctoral ex-

perience, from co-writing and conference organization to sleepless nights and long flights. The genuine admiration I have for his quick intellect has been a constant source of inspiration, pushing me to improve myself. Our countless philosophical discussions, profound and challenging, have shaped this dissertation in ways that cannot be fully quantified or qualified. I sincerely thank him; he is the best travel companion I could have asked for in such a long and demanding journey.

During my PhD, I spent about six months working on a humanoid robot at Cube Ltd. in Bari. I am deeply grateful to Michele for his guidance and for everything he taught me. His creativity, vision, and rich life experience have been a constant source of inspiration, contributing not only to my personal and professional growth, but also to my enthusiasm and motivation throughout this demanding phase of the project. More broadly, I owe a great debt to Cube Ltd. as a whole: without their technical expertise, resources, openness, and genuine commitment to collaboration, a substantial part of this work simply could not have been realized. The interaction between philosophical reflection and concrete robotic practice that Cube made possible gave this research much of its strength and originality, enriching it with an applied and experimental dimension that will profoundly shape my future work, both inside and outside academia.

When not working on the robot or writing this dissertation, I spent part of my PhD attending conferences around the world. I feel fortunate to have had the opportunity to travel extensively while doing what I love, and to meet people whose comments and questions helped me concretely refine the ideas, objections, and arguments developed in this work. I am especially grateful to Lena Fiedler, Susan Sterrett, Uljana Feest, Heidi Voskuhl, John Norton, Catherine Kendig, Catherine Elgin, Francesco Nappo, Kareem Khalifa, Filippo Contesi, Luisa Damiano, Marco Tamborini, Martina Bacaro, Tom Ziemke, Francesco Bianchini, Massimo Mazzotti, Morena Castro, Giacomo Eramo, David Naso, Sebastiano Stramaglia, Christoph Lüthy as well as to all the anonymous reviewers who read my abstracts and paper drafts, for their valuable critiques and insightful feedback. I also thank Roberto Lalli for the precious suggestions about the work.

The beginning of the PhD was not easy at all. In this regard, I cannot fail to express my gratitude to Helmut Pulte for his valuable advice and trust, with some regret for not having continued working together. I am also grateful to Jan from Ruhr University Bochum for his support and encouragement, and to Alfredo go my deepest thanks —a great friend and a true philosopher.

As a philosopher, one cannot know everything. For this reason, it is good to have a physician and a theoretical physicist in the family. I thank my brothers for their helpful comments on the medicine- and

mathematics-related parts of this dissertation, and for our many stimulating discussions. I also thank my parents for their constant support and for the remarkable resilience they have shown and passed on to me.

I thank all my friends, old and new, who helped me enjoy the good moments and forget the difficult ones. I am especially grateful to Alessio, Michele, Alessio, Mirko, Nicola, Marco and Daniele —lifelong friends with whom I have shared, and better understood, all the choices that have shaped my path— and to Sergio, Andrea, Mimì, Roberto, Giovanni, Rosita, Alessandro, Mauro, Simone, Corrado, Gigi and Marcello for their warmth, for helping me discover parts of myself I did not know, and also for our late-night poker games —perhaps another way of doing philosophy.

In Deutschland habe ich einen Teil meines Herzens gelassen, und ich bin froh, dort echte Freunde zu haben, die sich darum kümmern. Für ihre ständige und bedingungslose Unterstützung während all dieser Jahre danke ich Corina, Ralf, Renate, Norbert, Annette und Arno sowie Nico und Rainer, zwei Freunden, auf deren Begleitung ich, da bin ich sicher, immer zählen können werde.

I would also like to thank all my colleagues from the PhD program at DIRIUM Bari for the wonderful personal and professional experiences we shared, and for the many opportunities for growth in research. In addition, I am grateful to the PhD coordinator, Elisabetta Todisco, for her helpfulness and esteem, and to the students I had the opportunity to teach during this time, for their insightful questions and comments.

These three years have, in many ways, changed my approach to life. The journey continues, and I look forward to seeing where it leads.

Abstract

This dissertation explores the historical, philosophical, and practical foundations of artificial affectivity —the study and implementation of emotions in computational and robotic systems— with the goal of rethinking how emotions can be meaningfully represented and embodied in artificial agents. By integrating insights from the history of science, philosophy of mind, artificial intelligence, and robotics, it develops an interdisciplinary framework for understanding affect beyond the reductionist approaches that have recently dominated Affective Computing and Social Robotics. The research is divided into three parts. The first reconstructs the genealogy of artificial emotions, tracing their emergence as computational objects from the 1960s to the 1990s. It examines the evolution from early computational models of emotions to the advent of Affective Computing and affective robotics, revealing how emotions shifted from being viewed as irrational disturbances to functional components of intelligent behavior. This historical analysis also identifies a persistent *physiological-behavioral reductionism* (PBR) in the recent practice of modeling emotion computationally, in which affect is treated primarily as a set of measurable bodily or expressive signals. The second part critically addresses this reductionism by proposing a propositionalist framework for artificial emotions. Here, emotions are conceived as intentional and evaluative states possessing cognitive-propositional content, rather than mere physiological or behavioral reactions. This reconceptualization redefines the epistemological and design foundations of affective systems, suggesting that meaningful human-robot emotional interaction requires shared evaluative and interpretive contexts, beyond just behavioral mimicry. The third part translates these theoretical insights into practice through the EEVAN project —an experimental humanoid robot that uses music and sound as affective modalities to enhance social and emotional interaction. By employing multimodal and aesthetic (sound-based) channels of expression, EEVAN exemplifies a more nuanced and inclusive form of artificial affectivity. Overall, the dissertation advances a unified vision of artificial emotions as epistemic, technological, and cultural constructs, promoting a human-centered and philosophically informed approach to future affective and social robotics.

Abstract

Questa tesi esplora i fondamenti storici, filosofici e pratici dell'affettività artificiale —cioè lo studio e l'implementazione delle emozioni nei sistemi computazionali e robotici— con l'obiettivo di ripensare come le emozioni possano essere rappresentate ed incarnate in modo significativo negli agenti artificiali. Integrando prospettive provenienti dalla storia della scienza, dalla filosofia della mente, dall'intelligenza artificiale e dalla robotica, il lavoro sviluppa un quadro interdisciplinare per comprendere l'affettività al di là degli approcci riduzionisti che hanno tradizionalmente dominato l'Affective Computing e la robotica sociale. La ricerca è articolata in tre parti. La prima ricostruisce la genealogia delle emozioni artificiali, tracciandone l'emergere come oggetti computazionali dagli anni Sessanta agli anni Novanta. Essa analizza l'evoluzione dai primi modelli computazionali delle emozioni fino all'avvento dell'Affective Computing e della robotica affettiva, mostrando come le emozioni siano passate dall'essere considerate disturbi irrazionali a componenti funzionali del comportamento intelligente. Questa analisi storica individua inoltre un persistente *riduzionismo fisiologico-comportamentale* (PBR), secondo cui l'affettività è trattata principalmente come un insieme di segnali corporei o espressivi misurabili. La seconda parte affronta criticamente tale riduzionismo proponendo un quadro proposizionalista per le emozioni artificiali. In questa prospettiva, le emozioni sono concepite come stati intenzionali e valutativi dotati di contenuto cognitivo-proposizionale, piuttosto che come semplici reazioni fisiologiche o comportamentali. Questa riconcettualizzazione ridefinisce i fondamenti epistemologici e progettuali dei sistemi affettivi, suggerendo che un'interazione emotiva significativa tra esseri umani e robot richiede contesti valutativi e interpretativi condivisi, non solo imitazioni comportamentali. La terza parte traduce tali intuizioni teoriche nella pratica attraverso il progetto EEVAN —un robot umanoide sperimentale che utilizza musica e suono come modalità affettive per potenziare l'interazione sociale ed emotiva. Sfruttando canali espressivi multimodali ed estetici, EEVAN esemplifica una forma di affettività artificiale più sfumata e inclusiva. Nel complesso, la tesi propone una visione unificata delle emozioni artificiali come costrutti epistemici, tecnologici e culturali, promuovendo un approccio umanocentrico e filosoficamente informato allo sviluppo futuro dell'Affective Computing e della robotica sociale.

Contents

Declaration v
Acknowledgements vii
Abstract xii
Contents xv
List of Figures xix
List of Tables xxii
List of Abbreviations xxiii
Useful Notes for the Reader xxvi

Introduction **1**

I History and Epistemology of Emotions as a Computational Problem **21**

1 Emotions in the General History of AI **23**
1.1 The Making of an *Artificial Intelligence* 26
1.1.1 Early Background 26
1.1.2 The Birth of AI 29
1.1.3 GOFAI and the *Winter* of the Symbolic Approach 29
1.2 The Problem of Artificial Emotions 31
1.2.1 Emotions and Irrationality 32
1.2.2 Toward a Comprehensive View of Intelligence 35
1.2.3 Emotions and Machines in Literature 40
1.2.4 Toward Emotional (Artificial) Intelligence 44

2 Early Computational Models of Emotions **48**
2.1 An Overview 49
2.2 Augmented AI Models of Cognition (AAIMCO) 56
2.2.1 ALDOUS 56
2.2.2 HOMUNCULUS 60
2.2.3 Hot Cognition 66
2.2.4 PARRY 72
2.3 AI Models of Emotions (AIME): FEELER 81
2.4 The Evolution of Artificial Emotions 93
2.4.1 Theoretical and Computational Modeling of Emotions 93
2.4.2 Computational Approaches 96
2.4.3 Physiology and Sociality 98

3 The Technical and Philosophical Roots of *Affective Computing* 102
3.1 The Subsymbolic Approach . . . 104
3.2 New Perspectives on Emotions: Emotions as Embodied Cognitive Processes . . . 118
3.2.1 Emotions and the Body . . . 118
3.2.1.1 Bernard, Helmholtz and the Psychophysics . . . 120
3.2.1.2 Charcot and Darwin . . . 122
3.2.1.3 The James-Lange Theory . . . 123
3.2.1.4 The Cannon-Bard Theory . . . 126
3.2.1.5 George Dumas . . . 129
3.2.1.6 The Physiological Approach between the 1940s and the 1980s . . . 134
3.2.2 The New Paradigm of Embodied Cognition . . . 138
3.2.3 Emotions, Cognition and the Body . . . 146
3.3 New Perspectives on Emotions: Emotions as Social and Cultural Objects 151

4 *Affective Computing* and the Rethinking of Computational Emotions 163
4.1 Rosalind Picard's Revolution . . . 164
4.1.1 Enhancing Human-Machine Interaction . . . 164
4.1.2 Reconceptualizing Affect in Machines: Theoretical and Technical Aspects . . . 167
4.1.2.1 Artificial Emotions as a Form of *Intelligent* Sociability 168
4.1.2.2 Recognition, Multimodality and Embodiment . . . 170
4.1.2.3 Emotional Pattern Recognition and Learning . . . 176
4.2 New AC Methodologies in CME: An Overview . . . 181
4.2.1 Early Connectionism in CME: The Case of CATHEXIS . . . 187
4.2.2 Epistemic Gaps: Cognition and Physiology . . . 195
4.3 Physiological-Behavioral Reductionism: Some Notes Contemporary AC 202

5 From the Models to the Real World: The Rise of Social Robotics 209
5.1 Notes on the History of Affective Robotics . . . 210
5.1.1 18th-Century Musical Automata . . . 211
5.1.2 Early Experiments in Robotics: Cybernetics Creatures . . . 218
5.1.3 Early Experiments in Robotics: (GOF)AI Robotics . . . 228
5.2 A New Approach to Robotics . . . 233
5.2.1 The Limitations of GOFAI-Robotics . . . 233
5.2.2 Rodney Brooks' *Nouvelle AI* . . . 236
5.2.3 Humanoid Robotics: The COG Project . . . 242
5.3 Emotions in Social Robotics . . . 248
5.3.1 Designing a Social Robot . . . 248
5.3.2 Robotic Emotions and Intelligent Sociability: AC and the Kismet Project . . . 251

II Philosophy of Robotic Emotions 270

6 Current Perspectives on Robotic Emotions 272
6.1 The Physiological-Behavioral Reductionism . . . 274
6.1.1 Emotional PBR in Robotics: Some Historical and Design Reasons 278
6.2 Emotional PBR in Robotics: Practical Implications . . . 280
6.2.1 PB-Underdeterminacy . . . 282
6.2.2 Interpretational Variability . . . 282
6.2.3 Appropriateness and Misattribution . . . 284

6.2.4 Complex Emotional States . . . 284
6.2.5 Anthropomorphism and Uncanny . . . 290
6.3 Beyond Reductionism . . . 297

7 Robotic Emotions Beyond PBR: A Propositionalist Approach 301
7.1 The Social Constitution of Robotic Emotions . . . 302
7.1.1 From Design to Attribution . . . 302
7.1.1.1 Objections and Replies . . . 304
7.1.2 From Robotics to HRI . . . 309
7.1.2.1 PBR in HRI . . . 311
7.2 The Propositional Dimension of Emotions . . . 317
7.2.1 The Concept of Attribution . . . 317
7.2.2 Emotions and Intentionalty . . . 320
7.2.2.1 A Taxonomy of Mental States . . . 320
7.2.2.2 Emotions as Mental States . . . 323
7.2.2.3 Dynamic Evaluative Pluralism about Emotions . . . 327
7.2.3 Emotions and Propositions . . . 330
7.2.3.1 Particular and Formal Content of Emotions . . . 331
7.2.3.2 Evaluative Properties and Emotional Attitudes . . . 332
7.2.3.3 Formal Objects and Correctness Conditions . . . 333
7.2.3.4 Evaluative Representations . . . 334
7.2.3.5 Propositionality . . . 338
7.3 Rethinking the Epistemology of Robotic Emotions . . . 341
7.3.1 The Model $\mathcal{M}$ in HRI . . . 341
7.3.2 Propositionality and the Design of Robotic Emotions . . . 343
7.3.2.1 PB-Underdeterminacy . . . 344
7.3.2.2 Interpretational Variability . . . 346
7.3.2.3 Appropriateness and Misattribution . . . 347
7.3.2.4 Complex Emotional States . . . 348
7.3.2.5 Anthropomorphism and Uncanny . . . 349

III Application 353

8 The EEVAN Project: Musical Enhancement of Robotic Emotions 355
8.1 Emotions and Music . . . 356
8.1.1 A Historical Perspective . . . 357
8.1.2 Insights from Science and Philosophy . . . 358
8.2 Experiments in Musical (Affective) Robotics . . . 360
8.2.1 From Algorithmic Composition to Robots . . . 362
8.2.2 Examples from Contemporary Robotics . . . 369
8.3 The EEVAN Project . . . 371
8.3.1 Context . . . 371
8.3.1.1 The PNRR . . . 372
8.3.1.2 Cube . . . 374
8.3.2 The Prototype . . . 375
8.3.2.1 InMoov . . . 375
8.3.2.2 From InMoov to EEVAN: An Overview . . . 377
8.3.2.3 Hardware and Mechanics . . . 378
8.3.2.4 Electronics and Circuitry . . . 392
8.3.2.5 Software . . . 400
8.3.2.6 Program . . . 405
8.3.3 Experiment . . . 408
8.3.3.1 Background . . . 408

8.3.3.2 Conditions and Hypotheses 409
8.3.3.3 Participants 409
8.3.3.4 Experimental Environment 410
8.3.3.5 Experimental Procedure 411
8.3.3.6 Measures 412
8.3.3.7 Data Analysis 413
8.3.3.8 Results 416
8.3.3.9 Discussion 421
8.4 Further Developments and Possible Applications 425

Conclusion **433**

Bibliography **442**

List of Figures

1 CME evolution timeline . . . 8
2 Logical structure of the work . . . 10

2.1 ALDOUS' cognitive structure . . . 57
2.2 HOMUNCULUS' memory . . . 61
2.3 HOMUNCULUS' cognitive structure . . . 63
2.4 HC's cognitive structure . . . 69
2.5 PARRY's belief . . . 75
2.6 PARRY's flare concept . . . 75
2.7 PARRY's cognitive structure . . . 78
2.8 FEELER's knowledge representation . . . 84
2.9 FEELER's memory . . . 86
2.10 FEELER's cognitive structure . . . 90

3.1 Interdependence of emotion, cognition, and the body . . . 148
3.2 Plutchik's Wheels of Emotions . . . 154
3.3 PAD model of emotions . . . 155

4.1 Emotional activation . . . 175
4.2 Facial geometric mesh . . . 178
4.3 CATHEXIS' architecture . . . 188
4.4 Proto-specialists . . . 189
4.5 CATHEXIS' Emotion Generation System . . . 191
4.6 CATHEXIS' Emotional Behaviors System . . . 193
4.7 Simón the toddler . . . 194
4.8 AC's models theoretical geneaolgy . . . 198
4.9 AC's papers . . . 202
4.10 AC's areas . . . 204
4.11 AC's modalities and components . . . 205

5.1 Vaucanson's *Le Flûteur* . . . 214
5.2 Jaquet-Droz'*La musicienne* . . . 216
5.3 Kinzing-Roentgen's *Joueuse de tympanon* . . . 216
5.4 Walter's robotic tortoises . . . 220
5.5 Elmer's analog circuit . . . 221
5.6 CORA analog module . . . 221
5.7 Shannon's maze-solving machine . . . 223
5.8 The Johns Hopkins Beast . . . 229
5.9 Braitenberg's vehicles . . . 229
5.10 Shakey the Robot . . . 232
5.11 Brook's creatures . . . 239
5.12 FSM representation . . . 240

5.13 Subsumption architecture . . . 241
5.14 Elektro . . . 242
5.15 WABOTs . . . 244
5.16 COG . . . 248
5.17 Kismet . . . 252
5.18 Kismet's SNS . . . 256
5.19 Kismet's emotion system . . . 260
5.20 Kismet's affective space . . . 261
5.21 Emotion generation in Kismet . . . 262
5.22 Facial expressions of Kismet . . . 265

6.1 Models and channels in emotion recognition . . . 276
6.2 Sensing modalities in HRI . . . 277
6.3 NAO and Pepper . . . 278
6.4 The Uncanny Valley Hypothesis . . . 294

7.1 Taxonomy of mental states . . . 322

8.1 Kircher's *Musarithmica mirifica* . . . 361
8.2 *Mensa Tonographica* . . . 362
8.3 *Arca Musarithmica* . . . 363
8.4 WABOT-2 . . . 364
8.5 Waseda Flutist and Saxophonist . . . 365
8.6 Toyota's *Partner Robots* . . . 366
8.7 Honda's ASIMO . . . 367
8.8 Gil Weinberg's musical robots . . . 370
8.9 Italian PNRR funds allocation . . . 373
8.10 InMoov . . . 376
8.11 EEVAN's 3D printing log sheet . . . 378
8.12 STL components on Blender . . . 379
8.13 EEVAN: upper body graphic outline . . . 388
(a) Arm . . . 388
(b) Upper arm . . . 388
(c) Articulation . . . 388
(d) Hand . . . 388
(e) Chest panels . . . 388
(f) Internal chest structure . . . 388
(g) Eyes assembly . . . 388
(h) Jaw assembly . . . 388
(i) Neck assembly . . . 388
(j) Head structure . . . 388
8.14 EEVAN's hardware photographic record . . . 390
8.15 EEVAN's hardware video record . . . 391
8.16 EEVAN's electronics and circuitry . . . 395
8.17 EEVAN's electronics photographic record . . . 398
8.18 EEVAN's hardware video record . . . 399
8.19 MyRobotLab . . . 401
8.20 Unity . . . 402
8.21 EEVAN's cognitive architecture . . . 406
8.22 EEVAN's music and voice . . . 407
8.23 EEVAN experiment information sheet . . . 411
8.24 EEVAN experiment subjective questionnaire . . . 412
8.25 EEVAN experiment dataset . . . 416
8.26 Mean and SD of non-neutral expressions . . . 417

8.27 Individual non-neutral expression percentages 417
8.28 Participants' prior experience with the robot 419
8.29 Participants' preferences 419
8.30 Subjective ratings of the interaction experience 420
8.31 Interactive analysis 421
8.32 The role of &HPST and epistemology in AI 435

List of Tables

2.1 Early CME 52
2.2 HC's memory 67
2.3 PARRY's rules 76
2.4 FEELER's rules 89
2.5 FEELER's rules classification 93

4.1 AC computational models 186
4.2 AC's functions and components 206

5.1 Eliciting conditions in Kismet 264

8.1 Prosody and emotions 368
8.2 EEVAN's experimental design 410

List of Abbreviations

ACT	Adaptive Control of Thought
ACT	Affect Control Theory
AC	Affective Computing
AI	Artificial Intelligence
AIME	AI Models of Emotions
AMS-Q	Attribution of Mental States Questionnaire
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
APS	Activity-Producing System
AU	Action Units
BGD	Batch Gradient Descent
BET	Basic Emotions Theory
BR	Behavior-Releaser
BPA	Backpropagation Algorithm
BVP	Blood Volume Pulse
CART	Classification and Regression Trees
CB(T(E))	Cannon-Bard (Theory (of Emotions))
CME	Computational Model(s)/-Modeling of Emotions
CNN	Convolutional Neural Network
CO-WM	Cognitive Working Memory
CORA	Conditioned Reflex Analogue
CPU	Central Processing Unit
CRT	Core Relational Theme
DEPE	Dynamic Evaluative Pluralism about Emotions
DL	Deep Learning
DOF	Degree of Freedom
DST	Dynamical Systems Theory
EDA	Electrodermal Activity
EDVAC	Electronic Discrete Variable Automatic Computer
EEG	Electroencephalography
EF	Emotional Fungus-Eater
EBS	Emotional Behavior System
EB	Emotional Behavior
EGS	Emotion Generation System
ELMER	Electro Mechanical Robot
ELSIE	Electro Light Sensitive with Internal and External Stability
EM	Expectation-Maximization
EMG	Electromyography
ENIAC	Electronic Numerical Integrator and Computer
EPS	Emotion Proto-Specialist
FC	Forced-Choice
FACS	Facial Action Coding System
FP	Frame Problem
FSM	Finite State Machine
fMRI	Functional Magnetic Resonance Imaging
GOFAI	Good Old-Fashioned AI
GPS	General Problem Solver
GPS	Global Positioning System
GQS	Godspeed Questionnaire Series
GPU	Graphics Processing Unit
HC	Hot Cognition
HCI	Human-Computer Interaction
HMM	Hidden Markov Models
HMI	Human-Machine Interaction
HRI	Human-Robot Interaction
ID3	Iterative Dichotomizer 3
IAPS	International Affective Picture System
IMU	Inertial Measurement Unit

IRI	Interpersonal Reactivity Index
IRT	Infrared Thermography
JL(T(E))	James-Lange (Theory (of Emotions))
k-NN	k-Nearest Neighbors
LLM	Large Language Model
LHS	Left-Hand Side
LT	Logic Theorist
LTM	Long-Term Memory
MAP	Maximum a Posteriori
ML	Machine Learning
MO	Main Objective
MLP	Multi-Layer Perceptrons
NLP	Natural Language Processing
OCC	Ortony, Clore, Collins
OOP	Object-Oriented Programming
PAD	Pleasure, Arousal and Dominance
PBR	Physiological-Behavioral Reductionism
PET	Positron Emission Tomography
PGH	Physical Grounding Hypothesis
ph-WM	Physiological Working Memory
PPG	Photoplethysmography
ReLU	Rectified Linear Unit
RHS	Right-Hand Side
RNN	Recursive Neural Network
SAIL	Stanford Artificial Intelligence Project Laboratory
SAM	Self-Assessment Manikin
SFE	Solitary Fungus Eater
SFFS	Sequential Floating Forward Search
SM(H)	Somatic Marker (Hypothesis)
SNARC	Stochastic Neural Analog Reinforcement Calculator
SNS	Synthetic Nervous System
SSH	Symbol System Hypothesis
SPECT	Single Photon Emission Computed Tomography
STRIPS	Stanford Research Institute Problem Solver
SVM	Support Vector Machine
TM	Turing Machine
ToM	Theory of Mind
UTM	Universal Turing Machine
UVH	Uncanny Valley Hypothesis
VAD	Valence-Arousal-Dominance
VP	Vyzas and Picard
VTR	Varela, Thompson, Rosch
WABOT	WAseda roBOT
WM	Working Memory

Useful Notes for the Reader

Writing

This dissertation has been entirely written in LaTeX, using custom packages for bibliography management, mathematical notation, diagrams and graphs, code formatting, and quotations.

During the review phase, custom Python scripts were employed to automatically adjust bibliography entries, hyperreferences, and cross-references, as well as to ensure stylistic consistency throughout the document.

The text is written in U.S. English.

Hyperlinks

The text includes hyperlinks to bibliography entries (blue), to other sections and footnotes (red), and to external web pages (cyan). To ensure that web pages remain accessible from the printed version of the thesis, additional QR codes are provided. Hyperlink colors are visible only in the digital version of the dissertation.

Citations and Cross-References

The stile of citations and cross-section references follow a precise logic:

... "Yet, the most impressive musical automation experiments of the century were certainly the musical androids [Chapuis and Droz 1949; Margócsy 2015; Westermann 2016]"	The ideas or facts are explicitly reported in the mentioned sources.
(see ...) "Despite the recent focus of historians and philosophers on emotions following the so-called Emotional Turn of the 1980s–1990s (see [Section 3.3] and [Lemmings and Brooks 2014])"	

(see e.g., ...) "... with which behaviorism would have been dismissed as a scientific paradigm in the study of the inner life (see e.g., [Baars 1986; Pylyshyn and LePore 1999; Miller 2003; Boden 2008])"	The ideas or facts are both implicitly and explicitly reported in the mentioned sources, taken as examples of more extensive literature.
(see e.g., ..., etc.) "The history of musical automation is very old (see e.g., [Fowler 1967; Koetsier 2001; Chen et al. 2018], etc.)"	
(see also ...) "... Brooks and his collaborators at MIT Artificial Intelligence Lab began to give concrete form to an idea that had long captured the imagination of both science fiction and AI research alike (see also [Section 1.2.3] and [Asimov 1950])"	Used to complement previous references or indicate related sources.

When multiple sources are cited together, they are arranged chronologically in ascending order; in case of same year, the order is alphabetically following the name of the author:

[Pardo et al. 1993; George et al. 1993; 1994]

Quotations

All literal quotations from sources are enclosed in guillemets ("«" and "»"), with the page number indicated after the author–year citation. Emphasis in italics is retained as in the original unless otherwise indicated by the formula "my italics". Text translated by the author is marked as "my translation". Quotations in other languages are accompanied by English translations reported in the footnotes.

Bibliography

The bibliography is arranged alphabetically by the first author's family name. Entries by the same author are grouped and indicated with an em dash ("—"). In the case of co-authored works with the same first author, the shared author is replaced by a dash, regardless of whether the work is a monograph, a journal article, or an edited volume. Multiple works by the same author are listed in chronological order (earliest to latest). When several works by the same author were published in the same year, letters ("a", "b", "c", …) are appended to the year to distinguish them. For each author, individual works precede co-authored ones. Co-authored works are ordered alphabetically according to the second author's surname. The number of co-authors displayed is limited to three in the bibliography and to one in in-line citations, for reasons of space and readability:

Abbott, B. P., Abbott, R., Abbott, T. D., et al. (2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". *Physical Review Letters*, (116)6:061102.

Abelson, R. P. (1959). "Modes of Resolution of Belief Dilemmas". *Journal of Conflict Resolution*, (3)4:343–352.

— (1963). "Computer Simulation of 'Hot Cognition'". In: *Psychology's New Frontier: Computer Simulation of Personality.* Tomkins, S. S. and Messick, S., eds. John Wiley & Sons, New York, pp. 277–298.

— (1981). "Constraint, Construal, and Cognitive Science". In: *Proceedings of the 3rd Annual Meeting of the Cognitive Science Society, 19–21 August, 1981, Berkeley, California.* Vol. 3.

— (1983). "Whatever Became of Consistency Theory?" *Personality and Social Psychology Bulletin*, (9):37–54.

— and Carroll, J. D. (1965). "Computer Simulation of Individual Belief Systems". *American Behavioral Scientist*, (8)9:24–30.

— and Rosenberg, M. J. (1958). "Symbolic Psycho-Logic: A Model of Attitudinal Cognition". *Behavioral Science*, (3):1–13.

Afzal, S., Ali Khan, H., Jalil Piran, M., et al. (2024). "A Comprehensive Survey on Affective Computing: Challenges, Trends, Applications, and Future Directions". *IEEE Access*, (12):96150–96168.

Diagrams and Graphs

All diagrams and graphs were made using PGF/TikZ libraries for LaTeX, with exception of Figures 1, 2 (Introduction), Figures 5-7, 13, 18-21 (Chapter 5), Figure 4 (Chapter 6), Figures 1, 2, 9, 26-30 (Chapter 8), for which Adobe Illustrator was used.

Figures

For all figures, the source is indicated as "—from ...". For reasons of space, clarity, or style, some figures have been modified or rearranged; these are marked as "—adapted from ...". Figures without any source indication are original by the author. The author apologizes for any possible misattribution or unintended misuse of images and acknowledges all original sources with gratitude.

Gendered Pronouns

Gendered pronouns (*she/her* or *they/them*) are used throughout the dissertation when referring to individuals, depending on context and readability, to maintain gender neutrality and inclusivity. All choices of pronouns are intended solely for stylistic clarity, with no implications regarding gender identity.

Chapter Structure

Each chapter opens with a brief introduction that presents its main topic, outlines its contents, and situates it within the broader argumentative framework of the disser-

tation. This introductory section also provides an overview of the individual sections of the chapter.

Moreover, each chapter concludes with a summary that synthesizes the chapter's main arguments and theoretical contributions to the general topic of the dissertation, while establishing a conceptual link to the following chapter.

As far as possible, each chapter has been conceived as a self-contained unit. Readers interested in specific topics may therefore read individual chapters independently. Hyperlinks connect the chapters and help the reader navigate related discussions elsewhere in the dissertation.

Technical and Formal Contents

Given the historical and philosophical scope of the dissertation, a certain degree of formalism was necessary to discuss the arguments with appropriate precision and awareness. Readers, however, may safely skip the more technical sections, as these are balanced by explanatory passages summarizing their main points, without compromising the overall understanding of the argument.

The Words "Emotion" and "Affect"

In this dissertation, the terms "emotion" and "affect" —as well as their adjectival, verbal, and adverbial forms— are used interchangeably, without systematic distinction intended between them.

Use of LLMs

Large Language Models (LLMs) —specifically ChatGPT and Claude— were moderately used throughout the work for the following both formal and conceptual purposes:

- formal, stylistic, and linguistic refinement of the English text, including proofreading, readability improvement, and titling;
- brainstorming, conceptual development, and structural outlining of sections and chapters (often based on conceptual maps and networks previously created in Obsidian by the author);
- drafting and summarizing of sections, captions, and bibliographic content;
- maintenance of terminological and stylistic consistency across the dissertation;
- clarification of technical and logical aspects, and assistance in data analysis in [Section 8.3.3.7];
- management and debugging of LaTeX and Python code.

These tools served exclusively as writing and research aids under the author's direct supervision. All outputs produced through them were carefully reviewed by the author and subjected to plagiarism-detection software (Quillbot and ZeroGPT) to ensure originality before inclusion in the dissertation. The models were prompted with standardized, narrowly defined instructions limited to the analysis of sources provided by the author (selected articles, books or parts of the dissertation itself), without retrieving or incorporating external material.

Consequently, the dissertation remains entirely original in both content and writing.

Published Parts

Some parts of the present dissertation are based on works that have been published in journals or presented at conferences. In particular [Section 5.1.1] and [Section 8.2.1] are based on the paper

> Zanzarella, I. (2024). "Outlining a Protohistory of Artificial Intelligence and Music: From Antiquity to Nineteenth Century". Rivista di Studi Politici, (1):99–141 —[Zanzarella 2024].

and conference papers for &HPS10 (Caltech, Los Angeles USA, March 2025), fPET (Karlsruhe, Germany, September 2024) and SISS 2025 (Bari, Italy, May 2024); whereas contents from [Section 7.2.3] have been presented at AISC 2024 (Rome, Italy, September 2024).

Affectivity in Computational and Robotic Systems: History, Epistemology and Applications

Introduction

Today, technology universally amounts to a pervasive element of our everyday life. Computers and robots assist human beings in practical tasks that would be otherwise impossible to them, from the complex analysis of genetic data in genomics or the real-time processing of vast financial transactions, to the exploration of outer space or the performance of highly precise microsurgeries. Technology, however, has long since ceased to be merely a *passive tool* in the hands of specialized scientist and professionals. The rapid technological progress, miniaturization, and decreasing production costs are indeed among the factors that, especially in the course of the last decades, have progressively democratized access to computation. Devices once confined to laboratories and operating theaters now populate the pockets and living rooms of ordinary people.

In this respect, the ubiquitous social presence of technology has transformed computers and robots in *active companions* that assist humans in a wide range of everyday activities, from communication and education, to entertainment and personal healthcare: Service robots and smart appliances can adjust to users' habits and respond to their presence, quietly integrating into the rhythms of domestic environments; in hospitals and care homes, social robots accompany patients during rehabilitation, engaging them in conversation, or offering comfort through voice and presence; in schools, AI-based software support teachers by interacting directly with students, fostering engagement, curiosity, and inclusion through dialogue, play, and shared activities that reshape the

dynamics of learning and attention; LLMs are even becoming reliable therapists for many people, stimulating dialogue, providing guidance, and sustaining interaction in contexts of isolation or distress.

Not without raising concerns over dependency, misguidance, and the potential harm of misplaced trust, technology is hence entering domains once reserved *only* for human relationships, gradually assuming roles that involve responsibility and understanding. For this reason, their design has progressively shifted from mere efficiency to the simulation of lifelike presence and responsiveness: The ambition is no longer to build tools that perform tasks, but entities capable of actively sustaining *meaningful* and *authentic* social interactions, by intelligently adapting to users' needs and personality, and by expressing forms of social behavior that resemble those typically displayed by human interlocutors.

In this context, the ability of computers and robots to perceive, interpret, and reproduce *emotional cues* has become a central objective in the development of artificial systems. As interaction grows more personal and socially nuanced, the simulation of affect namely emerges as a necessary condition for machines to be perceived as genuinely responsive and engaging within human environments.

Emotions are integrated psychophysiological processes that involve evaluative appraisals, bodily changes, and behavioral tendencies, through which organisms respond to objects, situations, events and states of affairs in the world. These latter typically constitute the specific content of an emotional state, which is also characterized by a qualitative, subjective, and phenomenological dimension, marking it as a *lived experience*. Emotions normally possess a limited temporal duration, and unfold according to distinctive profiles of increasing and decreasing intensity. In this respect, they differ from moods —which are temporally extended and often without specific contents, from sentiments —which denote more enduring and reflective affective dispositions, and from personality traits —which describe stable patterns of emotional and behavioral tendencies across time.

As centuries of research in biology, psychology and philosophy have shown, however, emotions are not mere reactions to external stimuli, but intricate processes through which individuals apprehend the significance of events, actions, and relationships in the world they inhabit. In

that, emotions play a fundamental role in social life, operating as complex encoders of shared meanings, norms, and expectations, and shaping communication, cooperation, and moral understanding between subjects. They regulate interpersonal behavior, transmit values, and create the sense of connection and belonging that underlies every form of community, while allowing to convey and interpret intentions, anticipate and coordinate actions or even "read" others' minds.

If artificial agents are to participate meaningfully in human social environments, hence, they must engage, at least partially, in these same affective dynamics. The challenge is not simply to endow machines with the capacity to reproduce or recognize emotional expressions, but to provide them with models capable of representing, interpreting, and responding to affective states in a manner consistent with human expectations and social norms.

Now, the idea of endowing machines with emotions may appear intuitively unsettling. Emotions are commonly regarded as the most intimate expression of human subjectivity, understood as phenomena grounded in living bodies and conscious experience, inseparable from the organic conditions of biological life. Their attribution to mechanical or computational systems thus challenges long-standing distinctions between natural and artificial forms of existence. And yet, even far before the birth of Artificial Intelligence in the 1950s, such an idea has accompanied the history of technology since its earliest stages —from Athanasius Kircher's 17th-century experiments with algorithmic music and the mechanical androids of the 18th century, to the first attempts to computationally model emotions in the 1960s and the development of affective computing and robotics in the 1990s-2000s.

Today, affective machines are no longer a matter of speculation but a tangible reality. Social robots capable of recognizing facial expressions, voice assistants that adjust their tone to the user's mood, and affective computing systems designed to detect stress or engagement have become integral components of contemporary technological ecosystems, allowing increasingly complex forms of human-machine social interaction.

Although affective technologies have attained remarkable levels of sophistication, their present condition still appears as a transitional stage within a wider trajectory of research efforts and perfectible designs. The

growing integration of emotional intelligence into artificial systems opens an extensive domain for further reflection, concerning both its conceptual foundations, and its prospective evolution toward implementations capable of ensuring more coherent, adaptive, empathetic, comprehensive and contextually attuned forms of interaction. Such reflection necessarily calls for the convergence of philosophical inquiry, empirical investigation, and computational practice with the aim of refining the epistemological framework and the operational principles of artificial affectivity.

In particular, as affective technologies become increasingly embedded within the diversity of human social environments, such refinement proves essential to prevent reductive or uniform conceptions of emotion from serving as the only theoretical basis for their development. Rather than abstracting affective life into a limited set of functional patterns, it is necessary to acknowledge its intrinsic complexity as a personal, cultural, and social phenomenon, articulated through the plurality of ways in which emotions are expressed and interpreted across different contexts and situations. By identifying and pursuing new theoretical approaches that encourage the recognition and integration of this complexity into current practices of computational modelling of emotions, the interdisciplinary study of artificial affectivity can support the development of systems that not only respond to human emotions, but also embody a more inclusive, profound, and authentically "human" understanding of what it means to engage affectively across different forms of life and experience.

This *main objective* (MO) —the development of affective systems, humanoid robots in particular, capable of embodying a more inclusive and "human" understanding of emotional engagement— delineates the general background of the present work[1], which broadly contributes to it by outlining the mentioned theoretical foundations required to inte-

[1]This objective was also set by the PNRR Program [Section 8.3.1.1] within which the present doctoral research has taken place: «*"The Imitation Game". Humanoid Robots and the Simulation of Emotion for a "Human" and Inclusive Technology.* The project aims to build a cross-sectoral bridge between philosophical disciplines, neuroscience and computer science, envisioning research that has useful implications for the purposes of scientific foundation and operational design of "human", emotional, and inclusive form of innovation based on the use of humanoid robots» [Università di Bari 2022, p. 65].

grate such affective complexity and plurality into the epistemology and modelling of artificial emotions.

Indeed, the achievement of MO requires an in-depth understanding of the conceptual and epistemological frameworks underlying current affective technologies, together with the assumptions about the notion itself of "emotions" that guide their design and implementation. Such understanding demands an investigation into the *nature*, *origin*, *evolution* and (cultural) *context* of these frameworks, aimed at clarifying the theoretical and methodological limitations that prevent affective systems from achieving more inclusive and credible forms of emotional interaction.

Therefore, this investigation must primarily take the form of a *historical inquiry*, intended to reconstruct how emotions have been conceived, formalized, and modeled within computation and robotics, to uncover the philosophical and psychological conceptions of emotion that have shaped implementational practices in those fields.

In general, the study of emotions as historical, cultural, and social objects has gained a great prominence following the *Emotional Turn* of the 1980s–1990s. Scholars have tried not only to reconstruct the origin, development and foundations of scientific theories and philosophical conceptions of emotions, but to treat emotions themselves as historical objects, whose expression, representation, interpretation, subjective experience, social function and regulation have evolved across political and religious communities through different time periods and cultural contexts.

This *historicization*, however, has until now regarded solely *human* emotions, while it has not been pursued in the case of *artificial* emotions. To date, in fact, no comprehensive historical analysis has sought to clarify why, when, and how artificial emotions emerged, nor to trace their evolution from their early origins to the present. No thorough investigation exists trying to connect their technical development to the broader scientific, philosophical, and cultural transformations in the understanding of mind, intelligence, and affect itself that have shaped scientists' conceptualizations and implementations of artificial affectivity. No organic historical account has been put forward contextualizing artificial emotions with respect to the general history of computer science, Artifi-

cial Intelligence (AI), robotics and cognitive science, leaving a significant gap in the scholarly under-standing of their origins and evolution.

Some sporadic works have indeed attempted to address the historical dimension of artificial affectivity. However, when they are not outdated, they tend to be limited in scope, often focusing on specific technical developments or isolated time periods. While these studies provide valuable insights into the evolution of artificial emotions, they generally lack a comprehensive historical approach and fail to adopt a rigorous historiographical methodology. Many are little more than chronological inventories of models and implementations, without delving into the shifting concepts of mind, emotion, and representation that underpin them. Moreover, they often overlook the broader philosophical, epistemological, and intellectual contexts that have shaped the conceptualization of emotions as computationally modeled objects.

Performing a historical reconstruction of this kind for understanding *how*, *when*, and *why* emotions have become objects to model in computation and robotics —under which epistemic conditions, through which theoretical assumptions, by which technical means, and for which practical aims— is crucial not only for its intrinsic, and original, historical value. In addition, it is especially relevant for identifying and contextualizing the roots of the internal constraints and unresolved ambiguities of present affective computing and robotics. Only thereby, in fact, it is possible to begin to formulate concrete methodological directions and conceptual tools that may guide these disciplines in advancing toward MO.

In this work, such a historical-epistemological reconstruction is carried out in [Part I], through a detailed analysis of the ways, practices, and conceptualizations by which emotions have been modeled in AI and robotics. This analysis is primarily focused on the time period between the 1960s and the 1990s. This delimitation has both historical and methodological reasons.

Historically, it is namely in the 1960s that the debate about artificial emotions firstly emerged. The first *computational models of emotions* (CMEs) emerged in the mid-1960s, following a series of epistemological shifts that challenged the rationalist conception of intelligence in early AI —marked, for instance, by the critique of purely symbolic models in

cognitive science, the decline of behaviorism, and the rise of holistic and affective approaches to cognition. Once confined to the realm of irrationality, emotions were gradually reintegrated into the very notion of "intelligence", becoming a legitimate object of inquiry for those scientists who, in those years, were seeking to build an artificial one. During these three decades, several CMEs were developed, reflecting a gradual transition from cognitive architectures merely *augmented* with affective or motivational variables (so-called *Augmented AI Models of Cognition*, AAIMCO), to models explicitly devoted to *generating* emotions themselves (so-called *AI Models of Emotions*, AIME). This trajectory eventually led, in late 1990s, to the true *revolution* occurred in the field of computational emotion modeling, that is, Rosalind W. Picard's introduction of *Affective Computing* (AC) as a new approach to artificial affectivity, based on the integration of subsymbolic and multimodal approaches in computation, embodied conceptions of the mind and a social and interpersonal view of emotions. AC marked an irreversible turning point in the conception of artificial emotions, inaugurating their contemporary history. Also, in the late 1990s, AC represented the background against which social and affective robotics could emerge as a true materialization of its principles in physical objects. No longer confined to abstract architectures and virtual simulations, artificial affectivity now became with robotics embodied in physical agents capable of expressing, recognizing, and regulating emotional states in the context of human social interaction.

Methodologically, this temporal delimitation allows for a coherent and historically consistent reconstruction of the formative phase of artificial affectivity —so far neglected in the literature— prior to its consolidation as an autonomous field with the advent of AC. The period from the 1960s to the 1990s encompasses the conceptual, epistemological, and technical premises that first made the notion of computational emotions conceivable, while remaining sufficiently defined to ensure analytical precision and contextual depth. Focusing on this interval makes it possible to identify the conditions under which emotions became a legitimate object of computation and to trace the historical and conceptual coordinates from which subsequent developments derive. This chronological boundary, therefore, does not imply a separation from contemporary

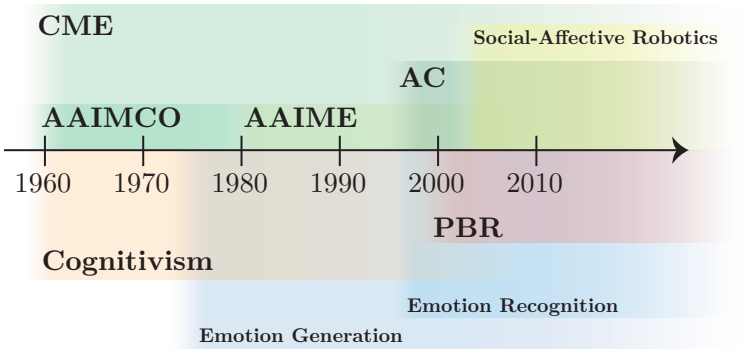


Figure 1: Timeline overview of the technical (above) and theoretical (below) evolution of computational modeling of emotions addressed by the historical analysis of this work.

research. Rather, by examining the very foundations of artificial affectivity, the analysis clarifies the conceptual frameworks that still inform present-day approaches to emotional machines, which essentially rely on a continuation and transformation of principles firstly articulated during that formative phase.

In this sense, the historical inquiry proposed here offers not a closed narrative of the past but a genealogical perspective for critically understanding the present condition of artificial emotions. In particular, it highlights how early attempts to formalize emotions within artificial systems were shaped by a philosophical orientation —*affective cognitivism*— that has subsequently defined both the scope and the limits of subsequent research on artificial emotions, at least until the 2010s. In fact, both in AAIMCO and AIME (*mutatis mutandis*), emotions were broadly conceived in cognitive terms, i.e., as *evaluative processes* of *appraisal* and *representation* linking beliefs, goals, desires and plans, implemented as symbolic variables within reasoning architectures. On the contrary, with AC and its emphasis on *emotion recognition*, beyond emotion generation, the focus progressively shifted to the simulation and detection of emotions in artificial systems mainly through physiological and behavioral correlates (facial expressions, voice modulations, bodily signals, etc.).

The growing reliance, in affective computer science and robotics, on the identification of affective states with their measurable physiological and expressive manifestations has eventually given rise to what is here defined as an affective *physiological-behavioral reductionism* (PBR), a

theoretical framework of emotion that still today shapes research and implementational approaches to artificial affectivity [Figure 1].

The historically contextualized persistence of PBR in contemporary AC, social robotics, and Human-Robot Interaction (HRI) is examined, in this work, through a critical analysis of recent reviews and research articles in these fields. Particular attention is given to the underlying assumptions about the notion of "emotion" that inform the design of computational and robotic affective architectures, as well as experimental studies on human-robot affective interaction. While the interpretation advanced here is necessarily partial and open to revision, it is intended to foster a critical reflection on current attempts to construct emotional machines, clarifying the conceptual limitations that still affect the field, and suggesting possible directions for their theoretical and methodological refinement in light of MO.

In this regard, affective PBR constitutes the principal theoretical limitation this work seeks to address, especially with reference to implementations in social robotics. By identifying emotions with observable physiological signals or external behavioral expressions, it confines artificial affectivity to a domain of measurable reactions, neglecting the interpretive, cognitive, as well as cultural and social dimensions that typically give emotions their meaning and role in human social context. To overcome this limitation, hence, the conception of emotion for affective computing and robotics should be normatively enriched and broadened, through the (re)definition of artificial emotions not only as physiological-behavioral signals to be detected or simulated, but also as evaluative and intentional structures that mediate the relation between agents and their environment, guiding perception, decision, and interaction through context-dependent processes of evaluation and meaning attribution.

Rethinking the concept of artificial emotions beyond PBR is the philosophical task undertaken in [Part II]. Here, the conceptual and epistemological conditions under which emotions can be meaningfully attributed to artificial agents are examined, questioning the adequacy of reductionist models and exploring the possibility of a *propositionalist* understanding of artificial affectivity. Within this perspective, emotions are conceived as intentional and evaluative states whose structure includes a propositional dimension. This implies that emotions are not

entirely reducible to patterns of bodily activation or behavioral display, but presuppose an epistemic relation to the world —a way of representing it as being a certain way. Accordingly, artificial affectivity must be reconsidered in terms of how machines can be designed or interpreted as holding, expressing, or simulating such evaluative-propositional representations, thereby allowing for a socially and epistemologically grounded attribution of emotions.

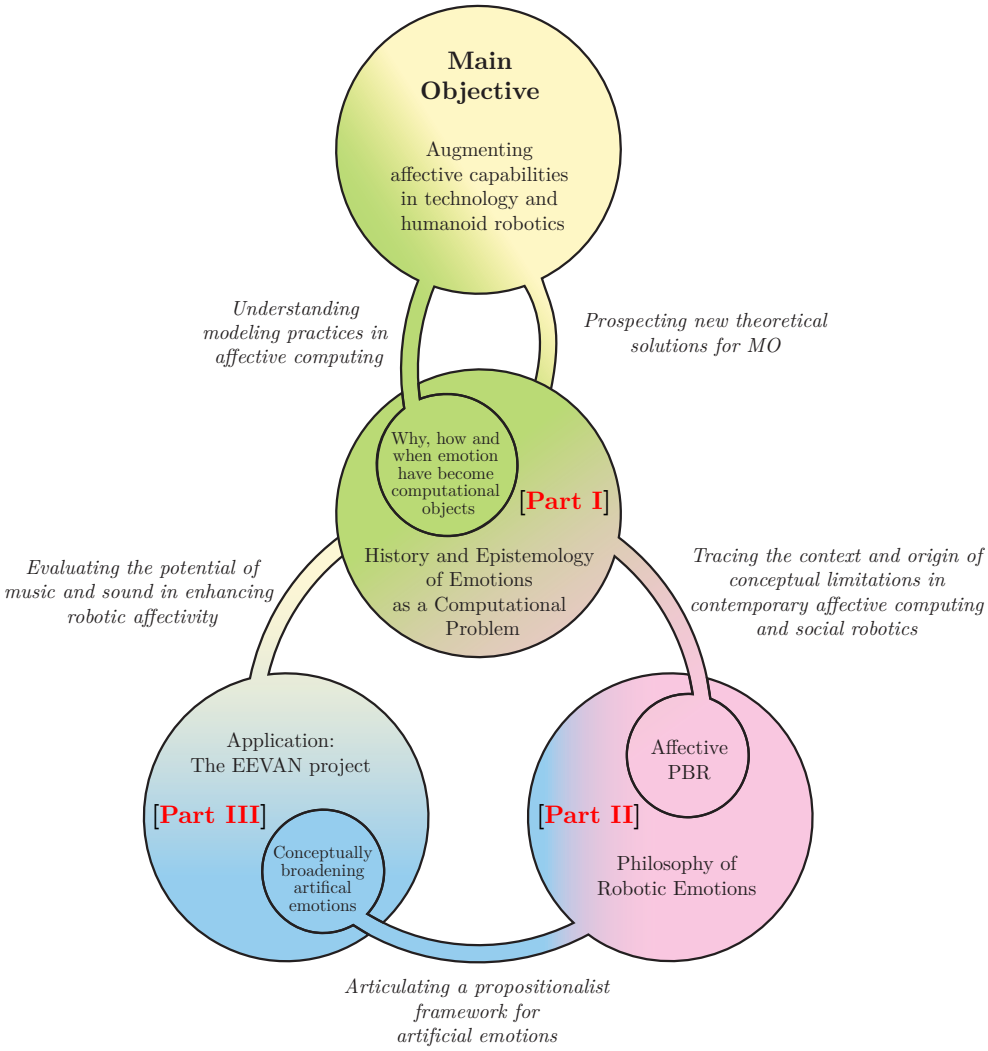


Figure 2: Overview of the logical structure of the present work.

Including —*alongside* simulated or recognized physiological signals and behavioral expressions— a cognitive-propositional dimension in the design of artificial emotions allows for more coherent and interpretable

affective behavior of artificial agents within social environments, even enabling the emergence of *complex* emotional states and fostering a more human-like, context-aware form of artificial affectivity. Expanding in this way the theoretical framework of artificial emotions beyond the constraints of PBR is also the goal of [Part III], that translates the theoretical insights developed in the previous sections into a concrete experimental platform, EEVAN, an open-source humanoid robot developed in collaboration with a specialized company within the PhD PNRR program [Footnote 1]. EEVAN explores an alternative path to artificial affectivity through *musical interaction*, conceived as a valuable medium capable of extending emotional communication beyond purely physiological or behavioral channels. By employing music and sound as means of expression, engagement and interaction, it exemplifies how affective communication between humans and machines can evolve toward richer, multilayered, and contextually attuned forms of expression, pointing to alternative ways of conceiving and designing emotional artificial systems beyond PBR-based paradigms.

Considered together, the propositional and "musical" dimensions of affectivity delineated in the second and third parts of this work offer two complementary directions for advancing the understanding and implementation of artificial emotions beyond the limitations of PBR. Both provide working computer scientists, designers and roboticists with conceptual tools to primarily *recognize* the reductionist assumptions that largely underlie the implementation current affective systems, and to *explore* alternative ways of endowing machines with a genuine and socially meaningful emotional life. By integrating evaluative-propositional structures and sound-based expressive-aesthetic modalities in the computational and robotic modeling of emotions, these approaches thus jointly contribute to the realization of MO, suggesting new theoretical paths toward the goal of developing affective technologies —and humanoid robots in particular— capable of engaging human beings in more coherent, inclusive, authentic and more "human" forms of emotional interaction.

This dissertation is articulated as follows. As already mentioned, it is divided into three main parts, each corresponding to a distinct yet interrelated level of investigation —historical, philosophical, and applicative— through which the general aim of the study is pursued. The parts comprise a total of eight chapters, each addressing a specific dimension of the

problem of artificial emotions and their implementation in computational and robotic systems:

Chapter 1 traces the emergence of emotions as a legitimate research object within AI, situating their rise in relation to the epistemological shortcomings of early cognitivism and the decline of purely rational conceptions of intelligence. It reconstructs how, around the 1940s-1960s, shifts in psychology, philosophy, and cybernetics, together with intellectual and cultural impulses from science fiction literature, opened the way for the first consideration of affect as a dimension of cognition to model computationally.

Chapter 2 examines the first generation of CMEs, from augmented cognitive architectures integrating affective or motivational variables (A-AIMCO) to early systems explicitly designed to generate emotions (AI-ME). Through detailed case studies —in particular, the models ALDOUS, HOMUNCULUS, Hot Cognition, PARRY, and FEELER— it outlines the conceptual transition from cognitive extensions to affective simulations.

Chapter 3 explores the technical and philosophical premises that prepared the ground for *Affective Computing* in the 1990s, focusing on the move from symbolic to subsymbolic computation and the progressive rise of embodied accounts of cognition and affect. It retraces how physiological, psychological, and cultural theories of emotion —from James-Lange to social constructivism— reshaped the epistemic understanding of affect within science and AI.

Chapter 4 analyzes Rosalind W. Picard's introduction of *Affective Computing* as a decisive *paradigm shift* in the modeling of artificial emotions. It discusses how multimodality, embodiment, and emotional pattern recognition redefined affective architectures, while also identifying the theoretical limitations that have led to the emergence of *affective physiological-behavioral reductionism* (PBR) as a characteristic theoretical framework of emotions in contemporary research about artificial affectivity.

Chapter 5 investigates the historical and technical convergence between affective computation and robotics, following the evolution from early automata and cybernetic creatures to social humanoids such as COG and Kismet. It shows how emotions became a functional component in the design of social robots, linking artificial affectivity to intelligent sociability and human-robot interaction.

Chapter 6 critically assesses current perspectives on robotic emotions, highlighting how the dominance of PBR constrains the epistemic and social plausibility of affective machines. It identifies the conceptual and practical problems —from interpretational variability to the uncanny effect— that arise when emotions are reduced to measurable physiological or behavioral signals.

Chapter 7 proposes a philosophical framework to overcome PBR by introducing a *propositionalist* (yet non-reductionist) conception of artificial emotions. It argues that emotions, understood as intentional and evaluative states with a propositional structure, can ground more coherent, complex, context-sensitive, and socially intelligible forms of affective behavior in artificial agents.

Chapter 8 presents the EEVAN project as an applicative realization of the previous theoretical insights, exploring how musical interaction can enrich robotic affectivity beyond PBR. By employing sound and music as expressive media, it demonstrates the potential of aesthetic and multimodal communication to foster authentic, inclusive, and emotionally meaningful human-robot engagement.

Taken as a whole, the work aims to contribute to the establishment of a more comprehensive understanding of artificial affectivity —one that recognizes its historical depth, philosophical complexity, and practical implications for future technological design. By integrating historical reconstruction, conceptual analysis, and experimental implementation, it aspires to bridge the gap between the humanities and the engineering sciences in the study of artificial emotions, showing that the design of affective machines cannot be reduced to a technical problem alone, but

must be grounded in an awareness of the epistemic, social, cultural, cognitive and normative dimensions that define emotional life. In this sense, the present research not only provides a critical account of how emotions have been conceived and modeled in artificial systems, but also proposes new directions for conceiving and constructing technologies capable of engaging human beings in more meaningful, responsible, and genuinely affective ways.

Before proceeding to the core of the investigation, however, some methodological and conceptual clarifications are still required in order to delineate the specific sense in which the problem of artificial emotions is here (not) understood and (not) addressed.

First, one might legitimately question whether we need emotional or empathetic machines at all. Traditionally, we expect from machines qualities such as reliability, efficiency, and neutrality —qualities that, as human experience teaches, are often influenced by emotionality. Furthermore, even if we were to accept the idea of emotional machines, it would remain unclear to what extent such emotionality should be developed or expressed: Few would be willing to compromise a machine's precision, performance, or speed merely to avoid the hypothetical risk of "hurting its feelings". And yet, the absence —or impoverishment— of emotional responsiveness may render machines unsatisfactory or even frustrating to their users. Questions of this kind, however, fall outside the scope of the present research. They concern the *ethical desirability* or *moral implications* of artificial emotions, whereas the following work treats the matter as a *fundamentally epistemological* one. The issue at stake is not whether machines *ought* to have emotions, but *how* emotions can be (more deeply and comprehensively) *understood*, *modeled*, and *attributed* within artificial systems.

Second, along the same line, this work will not engage with the *ontological* question of whether machines may not only express or detect emotions, but also *possess* and *understand* them in a manner comparable to human experience. Such issues, while undoubtedly fascinating, pertain to the metaphysical status of artificial agents and to the problem of consciousness —domains that exceed the epistemological scope of the present inquiry. Again, the following analysis remains focused on

the conceptual and theoretical conditions under which emotions can be represented, modeled, and interpreted within AI and robotics, independently of any claim concerning their ontological reality. About this latter aspect, in fact, the work professes a general *agnosticism* —given also that it is today still controversial whether and to what extent even humans, let alone artificial agents, have a mind, a consciousness, a phenomenology and an affective life, and what this exactly means and entails.

Third, as previously mentioned, emotions are complex and multidimensional states that bear certain resemblances to, yet remain essentially distinct from, other forms of affective life —primarily moods such as "depression", sentiments such as "love", or personality traits such as "optimism". The present research does not address these broader affective phenomena, but is methodologically confined to *emotions* understood as temporally bounded, intentional, and object-directed states, characterized by evaluative and phenomenological dimensions that mediate the (inter)relation between agents and their environment. Other affective states are therefore considered here only insofar as they serve to elucidate and delimit this specific notion.

Fourth, even though the present work spans across a qualitatively and quantitatively wide range of practices and implementations in the broader field of affective computational modeling, a special focus is given —explicitly and implicitly throughout— to the domain of *humanoid robotics*. A humanoid robot is typically understood as a soft or hard electromechanical device that reproduces human physical characteristics. As will emerge in several parts of this work, however, this morphological sense of the term "humanoid" is only one among its possible meanings —and perhaps even the most naïve one. Something can be humanoid not merely because it *looks* like a human being, but also insofar as it *behaves*, *thinks*, or even *emotes* in ways that resemble human modes of action and experience. Moreover, defining "humanoid-ness" —that is, *anthropomorphism* or *human-likeness*— is far from straightforward, since such resemblance cannot be established in purely objective terms, but depends on cultural, perceptual, and interpretive frameworks through which humans recognize themselves in technological artifacts. If so, even *disembodied* computational models of emotion —discussed in the first four chapters— may be regarded, in a certain sense, as humanoid to some degree, insofar

as they replicate functions and behaviors that are *typically human*. And thus, the question of why and how researchers eventually came to include a physical body within the very definitions of artificial intelligence and artificial emotion —thus giving rise to (social and affective) robotics— does not have a simple and obvious answer. On the contrary, it is a question that deserves an articulated and critical investigation following the complex historical, philosophical and technical paths of the ideas and practices that progressively shaped the embodiment of artificial affectivity, possibly into *morphologically* humanoid forms.

Finally, a last remark concerns the methodology applied in the historical part of the work. The historical reconstruction undertaken in this dissertation relies on a combination of both primary and secondary sources. Primary sources, on the one hand, include journal articles, technical reports, monographs and review papers published in the fields of AI, cybernetics, and cognitive science between the 1960s and 1990s —the formative decades for CME, assessed as to the specific conceptual frameworks, research practices, and assumptions about intelligence and affect they (often implicitly) adopt and express.

Secondary sources, on the other hand, play a crucial methodological role in both historical reconstruction and epistemological analysis. Beyond providing contextual support, they critically reassess and reinterpret primary materials, offering conceptual frameworks that help clarify technical and philosophical concepts often expressed with diverse vocabularies in the past. Their selection is primarily based on their ability to contextualize historical data within broader cultural, intellectual, and social dynamics, to provide solid interpretive frameworks for examining theoretical assumptions and epistemological implications, and serve as a basis for historiographical comparison, engaging critically with prior studies and their intellectual trajectories. This ensures that secondary sources are integral not only to the reconstruction but also to the theoretical and historical critique of the dissertation.

Furthermore, the historical investigation adopts a hybrid methodology drawing on a multiplicity of integrated historical-cultural and epistemological-philosophical approaches. As widely known, in the course of the last decades, the history of science as an institutionalized discipline has evolved through a series of heterogeneous methodologies, each focusing on differ-

ent aspects and modes of scientific development and knowledge production. Among the most notable ones, there are *Historical Epistemology* —emphasizing the role of actual scientific practices in shaping scientific concepts (see e.g., [Rheinberger 2010; Chimisso 2024; Vagelli 2024] and [Daston and Galison 2007]); *Integrated History and Philosophy of Science* —seeking to combine both historical and philosophical analyses to address scientific questions from a broader, interdisciplinary perspective (see e.g., [Giere 1973; Arabatzis and Schickore 2012; Scholl and Räz 2016; Herring et al. 2019]); *History of Epistemology* —traditionally concerned with tracing the evolution of scientific concepts through rational reconstruction of theories and accounts (see e.g., [Feest and Sturm 2011; Sturm 2011] and [Kuhn 1962]); and *Cultural History of Science* —delving into the social and cultural contexts that shape scientific inquiry, examining how scientific practices are influenced by broader societal norms and values (see e.g., [Rouse 1993; Dear 1995; Long 2022]).

The question of which approach is the most productive and valuable is still a matter of ongoing debate among scholars. Indeed, it is often (*apriori*) believed that the choice of one methodology over the others must be measured according to its relevance and appropriateness in facing contemporary philosophical and scientific debates. So, for example, the history of epistemology must be preferred to the epistemological history as only the first really helps clarifying and solve present-day epistemological problems [Sturm 2011]. Yet, this seems to be true only if one's goal is actually to solve present-day epistemological problems, which does not necessarily apply if one revolves to the history of science for other reasons.

Put differently, the choice of methodology is not absolute but always subordinated to the goal one aims to achieve. For this reason, this work should not be understood as advocating for the superiority of one approach over another. Its specific (and sole) interest in the history of science is focused on understanding, in light of MO, the origins and nature of the limitations affecting contemporary affective robotics and computing, seeking solutions not in or through history itself, but rather through an analytical-philosophical account of artificial emotions and an applied proposal for a robotic implementation, with history serving only to reconstruct and contextualize them.

Consequently, this work integrates perspectives primarily from historical epistemology —as it analyzes how computational modeling practices have influenced the concepts of (artificial) emotions and intelligence, situating the emergence of CMEs within their broader intellectual and socio-technical context. It also draws from cultural history of science —considering the role of non-epistemic factors (primarily sci-fi literature) in shaping scientific ideas. However, since the goal is not simply to offer a descriptive or archaeological account of past research on CMEs, but also to reconstruct and interrogate it through philosophical and epistemological questions regarding the nature, function, and meaning of *emotion* as a computational object, this work incorporates the history of epistemology as well —attempting to rationally reconstruct the epistemological issues and conceptual commitments embedded in CMEs. Finally, insights from the integrated history and philosophy of science are also present, combining historical and philosophical analysis to provide a systematic discourse on the same problem (MO). From the methodological point of view, hence, this research acts *pragmatically*, and yet still consistently —after all, critics themselves have noted that the methodologies of the history of science are not clear-cut but rather exist on a continuous spectrum of multiple and related approaches [Sturm 2011, 304].

In that, [Part I] does not claim to be —historically, conceptually and methodologically— exhaustive or definitive in scope, depth and accurateness. It represents, rather, the best result that could be achieved after three years of research on a largely unexplored topic —one that required engaging with abundant and often highly technical sources and materials, many of them relatively recent (dating back only sixty or seventy years). Rather than offering a detailed reconstruction of well-delimited historical problems, canonical authors, or clearly circumscribed theories situated within precise conceptual and chronological boundaries, the research traces those *broader trends* in AI and cognitive science that have shaped and guided the evolution of artificial emotions, particularly in their early stages. Also, more than an autonomous study, this historical part was conceived as the groundwork for a broader philosophical and applied investigation. The shortcomings that surely affect this reconstruction, however, do not diminish its significance. To the best of my knowledge, in fact, it represents the first attempt to trace a systematic

history of artificial emotions, conceived not as an arrival point but as a starting one: a first theoretical framework meant to foster further discussion, and a foundation to be refined, expanded, and deepened through future research, alternative historiographical perspectives, and complementary methodological approaches —statements that are indeed true for the whole work.

Part I

History and Epistemology of Emotions as a Computational Problem

Chapter 1

Emotions in the General History of AI

In the [Introduction], the broader context of this research has been illustrated. The present chapter now turns to the specific historical-epistemological problem of when, how and why emotions have emerged as an interesting research object in computation and *Artificial Intelligence* (AI).

As mentioned, the history of artificial emotions extends much further back than one might expect, indeed, well before the 20th century and the advent of computer science. In the 17th century, for example, the German-Roman polymath and Jesuit Athanasius Kircher [1650] treated affects as a variable in a system of automated music composition he invented, the *Musarithmica mirifica* [Section 8.2.1]: On the *mensa tonographica*, following the Baroque *Affektenlehre* (Eng. *Theory of Affects*, see e.g., [Glüxam 2020]), emotions were associated with the different church modes, so that combinatorially found sets of numbers for harmonies to assign to a given text could be translated into (expressive) musical intervals [Zanzarella 2024, 117–127]. Later, in the 18th century, musical automata were built [Section 5.1.1], for example by Jacques de Vaucanson in France, the Jaquet-Droz family in Switzerland, or Peter Kinzing and David Roentgen in Germany, whose purpose was, among other things, to *express* emotions through music[2] and suitable bodily and

[2]Interestingly, the early history of artificial emotions overlaps with that of artificial music [Footnote 61], probably because of the special, atavic connection between

facial movements (see [Voskuhl and Riskin 2010; Voskuhl 2019; Riskin 2021] and [Zanzarella 2024, 132ff.]).

Despite the recent focus of historians and philosophers on emotions following the so-called *Emotional Turn* of the 1980s–1990s (see [Section 3.3] and [Lemmings and Brooks 2014]), which mainly concerned *human* emotions as historical, cultural, and social phenomena, no real attempt to historicize, instead, *artificial* emotions has been seriously undertaken with comparable effort —one aimed at clarifying, for example, why, when, and how computational models of emotion (CMEs) emerged, to what ends, and under which theoretical conceptions of affect. No comprehensive historical analysis exists that seeks to trace the evolution of artificial emotions from their very beginnings, connecting it to the deeper philosophical and cultural transformations that have occurred in the understanding of mind, intelligence, and affect, as well as situating it within the general history of AI. The few existing works on this topic are generally outdated, limited in scope, and often fall short of a systematic historical methodology. They are written more in the style of technical reports by non-historians of science than as studies advancing historically and epistemologically informed hypotheses about the origins and development of artificial emotions.

To the best of my knowledge and archival research, in this respect, only one historically oriented account of early CMEs exists, [Pfeifer 1988], which, however, devotes only a few pages to the topic, providing an accurate technical description but offering little historical or epistemological contextualization[3]. Another article, [Marsella et al. 2010], authored by computer scientists, has more the form of a technical review than that of a work of historical research. Above all, it limits its analysis to CMEs developed between 1997 and 2010 —that is, to the period *after* the publication of Rosalind W. Picard's [1997] seminal work, *Affective Computing* [Chapter 4], which marked a paradigm shift in a research field that had

emotions and music as an art form [Section 8.1], one of the first domains in which algorithmic thinking was historically applied [Zanzarella 2024].

[3] Another historical survey of early CMEs exists, a technical report by Hudlicka and Fellous [1996], seemingly based largely on [Pfeifer 1988]. The report, published by a private company, is however untraceable in internet or libraries, so that even the authors themselves, contacted via e-mail in October 2024, confirmed that, being the document very old, there is presently no way to retrieve it.

however been developing for at least thirty years *already* (see [Chapter 2]).

The aim of the present chapter is thus to fill these research gaps by providing a first systematic historical reconstruction of artificial emotions, especially attempting to shed light on the historical, technical and epistemological conditions that have first led to the emergence of emotions as an interesting research object in computation and AI.

In particular, the main thesis defended in this chapter is that CME emerged as an urge in AI following a series of broad interdisciplinary evolutions occurred around the 1960s, with roots in intellectual and scientific trends already developing before: the decline of behaviorism and neopositivism; the emergence of holistic approaches in psychology; the first technical and theoretical skepticisms about GOFAI, symbolic AI and early cognitivism in psychology that, equating intelligence with rationality and symbolic computation, relegated emotions to the condition of irrational remnants of the mind with no legitimacy as a scientific object of study; and the appearance of a growing body of non-scientific literature highlighting the problems of a technology deprived of emotions.

After a brief historical contextualization of the background research [Section 1.1.1], birth [Section 1.1.2] and development [Section 1.1.3] of AI, I introduce the historical problem of artificial emotions in [Section 1.2]. In [Section 1.2.1], I discuss the traditional view of emotions as irrational, and in [Section 1.2.2] and [Section 1.2.3], I delineate the historical and cultural conditions representing the epistemological and conceptual background against which, in the 1960s, emotion could finally emerge as a legitimate research object in computation [Section 1.2.4].

Thereby, this chapter establishes a comprehensive conceptual framework that will enable, in the following [Chapter 2], a detailed analysis of the evolution of early CMEs through representative case studies. This, in turn, will provide the solid groundwork on which the paradigm shift introduced by Picard's *Affective Computing* [Chapter 4], as well as subsequent technical and philosophical research on artificial emotions in AI and robotics (see [Part II] and [Part III]), can be properly understood.

1.1 The Making of an *Artificial Intelligence*

1.1.1 Early Background

Is there a procedure to figure out whether problems exist that cannot be solved by computation?[4] This question was absorbing a famous English mathematician during the mid-1930s: the just twenty-two-year-old Alan Turing. In trying to solve it as a computational problem itself, Turing [1936] set out the mathematical concept of *general-purpose programmable computer*. Not much later, the Hungarian-American polymath John von Neumann [1993 (1945)] —basing on Eckert and Mauchly's work on the ENIAC (Electronic Numerical Integrator and Computer) and the EDVAC (Electronic Discrete Variable Automatic Computer) at the Moore School of Electrical Engineering (University of Pennsylvania, 1945)— sketched out the first logical architecture of a physical computer in which Turing's mathematical ideas could be concretely implemented[5].

[4]This question is also known by the name of *halting problem*, directly deriving from the famous *Entscheidungsproblem* (German for "decision problem"), firstly posed by the German mathematician David Hilbert [1900] (see also [Hilbert and Ackermann 1928; Bernays and Schönfinkel 1928; Hilbert 1930]). The *Entscheidungsproblem* asks wether, given a first-order formal theory $\mathbf{K}$ and a formula $\mathfrak{A}$, an effective procedure (a finite algorithm) exists for establishing (*deciding*) whether $\mathfrak{A}$ is provable in $\mathbf{K}$ or not. Following previous attempts to solve the problem (e.g., Church's [1936] λ-calculus), Turing [1936] introduced the *Turing machine* (TM), a mathematical model consisting of a reading "head" moving along a "tape" on which there are symbols to read. Depending on the symbol read (and on the TM's program), the head undertakes different types of actions: writing a new symbol in the cell, moving to the next cell on the right or on the left, remaining on the same cell. Given an input, thus, the machine will go through a finite number of different state transitions and eventually halt giving the adequate output. In general, it is assumed (albeit not formally proved) that the class of problems for which the machine will halt (i.e., computable problems) is equivalent with the class of TM-solvable problems (*Church-Turing hypothesis*). Now, to find out wether a problem is computable (TM-solvable) or not (*Entscheidungsproblem*), Turing [1936] devised the *universal Turing machine* (UTM), a TM programmed to answer questions about other TMs: Does a TM *halts* deciding wether $\mathfrak{A}$ is provable in $\mathbf{K}$ or not? Yet, the existence of a UTM capable of solving a *halting problem* is contradictory. It is namely *undecidable* whether the TM will halt or not, and there is no algorithm which can decide whether in general a problem is computable. By such clever mathematical expedient, Turing [1936] eventually gave his negative solution to the *Entscheidungsproblem*, as well as to the *halting problem*, finding the intrinsic limits of computation —there are namely problems for which the computation executed by a machine will run forever.

[5]Broadly speaking, the so-called *von Neumann architecture* organizes computer components into three main parts: a *Central Processing Unit* (CPU) that executes programs; a memory unit, which stores both data and (at that time, surprisingly) pro-

During the same years, an epistemological turn was ongoing in psychology, which will shortly after lead to the so-called *cognitive revolution* in the discipline (1950s-1960s), with which behaviorism[6] would have been dismissed as a scientific paradigm in the study of the inner life (see e.g., [Baars 1986; Pylyshyn and LePore 1999; Miller 2003; Boden 2008]). In [1943], McCulloch and Pitts laid the theoretical groundwork of *artificial neural networks* [Section 3.1] suggesting that the human brain, as a network of simple interconnected binary (on/off) units (neurons), can be thought of in computational terms and, thus, as being able to execute logical operations. In the same year, one of the pioneers of modern *cognitive science*, Kenneth J. Craik [1943, 57], conceived of the process of human «thinking or reasoning» as a process involving «representation by symbols, calculation, and retranslation into events»: The external world is «translated» [50] into symbols[7] by which subjects build an internal, representational, model of it. On this model, reasonings, deductions, inferences, calculations, etc. are performed so that other symbols are produced (*prediction*). The new symbols are finally retranslated into external processes and objects, or recognized as correspondent to existing external events (fulfillment of predictions).

From these approaches, an idea of human mind as a (formalizable) system for computing symbols and information was emerging, that tended to consider it in *analogy* to a computer in Turing's sense (see also [Gigerenzer and Goldstein 1996; Crowther-Heyck 1999]). This idea gained prominence among scientists also thanks to the birth of *cybernetics* (from the Ancient Greek κυβερνήτης/κυβερνάω, *kubernētēs/kubernaō*, meaning

grams; and input/output devices that facilitate interactions between the computer and the outside world (see e.g., [Laird 2009]).

[6] Behaviorism had been established as a major research line in psychology especially by John. B. Watson [1914] in the early 1910s, relying on the assumption that, given the epistemic inaccessibility of mental phenomena, the scientific study of them should be limited to their external and observable expressions, namely behaviors.

[7] Yet, these symbols are not understood by Craik [1943] as having the formal nature of the symbols in computation (Turing) or formal psychology (McCulloch and Pitts). They symbolize because of their *physical* nature, viz. «the ability [...] to parallel or imitate each other, or the fact that they can do so since there are recurrent patterns in reality» [58-59]. Modeling the external world by symbols is a process in which the nervous system produces «combinations of excited arcs» and «patterns of excitation» [56], namely the symbols, which are inherently consistent and coherent with the represented process or object.

"steersman/ to steer" —see also [Section 5.1.2]), founded at the end of the decade by another polymath, the American Norbert Wiener [1948]. As the science of communication and control in both animals and machines, the theoretical starting point of cybernetics was the scientific hypothesis that the principles governing intelligent behavior, information processing, learning and adaptation to the environment through feedback loops are comparable in both natural and engineered systems.

Still, in the 1950s, Herbert A. Simon and Allen Newell proposed that human problem-solving processes could be formalized in terms of symbolic representations, and that human intelligence as a whole could be understood and even replicated computationally, with the consequence that computers may perform cognitive tasks similar to human reasoning[8] (see e.g., [Newell and Simon 1956]). At the same time, the American psychologist George A. Miller [1956], a former behaviorist converted to *cognitivism*, insisted on the mind-machine analogy[9] (see also [Miller et al. 1960]) by showing, also from an empirical-experimental point of view, that human memory operates with inherent capacity limits[10].

[8] On the basis of this idea, Newell and Simon [1956] developed what is today acknowledged as one of the first AI programs, the *Logic Theorist* (LT), designed to prove theorems from the *Principia Mathematica*, the famous work by Bertrand Russell and Alfred North Whitehead. The program worked in such an extraordinary way that, in some instances, it could find proofs that were even more elegant and efficient than those given by the authors themselves. Later, building on LT and relying on the same theoretical presuppositions, Newell et al. [1959] generalized the program to larger domains and tasks: The so-called *General Problem Solver* (GPS) was thus designed by integrating means-ends analysis with the aim of formalizing human problem-solving strategies in various fields, from mathematics and geometry, to logistics and chess-playing.

[9] Indeed, there was a dispute among Simon, Newell and Miller about the authorship of the idea of the mind as a computational system, which resulted in Simon's apologies after Miller's detailed arguments in [Miller et al. 1960]: «As far as I was concerned they were old familiar ideas; the fact that they had thought of it for themselves didn't mean that nobody ever thought of it before» [Miller 1986, 213].

[10] Miller [1956] identified the quantitative limit of short-term memory in the number seven —in this, it is clear (and explicit) the influence on the quantitative account of information by Shannon [1948] (see also [Shannon and Weaver 1949]). By analyzing data from a series of experiments in literature about memory span and information processing, he found a special pattern related to this particular number: Humans can hold approximately seven chunks of information at once. Although later studies have pointed out that this specific claim about the number seven may not be true, the main, *cognitivist*, theoretical tenet backing Miller's research (quantitative limitedness of mind) is accepted (see e.g., [Cowan 2001]).

1.1.2 The Birth of AI

Now, after such major advancements in mathematics, computer science, cognitive psychology and neuroscience, the time was ripe for a singular, and maybe provocative, idea: There can be an *artificial intelligence*; there is a possibility that machines can exhibit behaviors as intelligent as we would expect it from real human beings (see e.g., [Turing 1950]).

The official birth of Artificial Intelligence (AI) as a scientific discipline is often marked by the *Dartmouth Conference*, held at Dartmouth College (Hanover, USA) in 1956. This conference, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon brought together many leading researchers in the field and raised profound discussions about various issues, approaches and methodologies pertaining to machine intelligence, which ultimately set the stage for future research and developments in AI (see e.g., [Nilsson 2009; Russell and Norvig 2020; Wooldridge 2022]). In particular, the conference corroborated and catalyzed research lines that had been being (implicitly) adopted by most scholars already since Turing's work on the *halting problem* (see [Footnote 3]). Such research lines were based on the firm belief that human intelligence can be captured into a discrete logical systems of clear and interpretable symbolic representations of knowledge, and that these representations can be manipulated by sets of explicit, well-defined, formal rules (essentially in the sense of Leibniz' *characteristica universalis*). The hypothesis was that, by suitably representing knowledge through symbols, and by implementing in a machine the right algorithms for manipulating them, the machine could be, in theory, *intelligent*, showing abilities of reasoning-performance, problem-solving, and decision-making similar to those shown by humans' biological intelligence.

1.1.3 GOFAI and the *Winter* of the Symbolic Approach

This approach, termed *symbolic*, of replicating artificially human intelligence using logical structures and deterministic algorithms characterized most of AI research from the beginning to at least the early 1980s. It even produced astonishing results, such as the so-called *expert systems*, i.e., AI

programs based on rule-based logic and representation of knowledge from a specific limited domain, designed to simulate the decision-making and reasoning abilities of a human expert in a specific field[11] (see e.g., [Russell and Norvig 2020; Wooldridge 2022]). Yet, this symbolic approach, characteristic of the Good Old-Fashioned AI (GOFAI) as it has been called by Haugeland [1985, 112ff.] presented also some substantial weaknesses [Section 1.2.4].

Expert systems themselves, even being capable of reasoning through complex scenarios and showcasing the potential of symbolic AI in solving well-structured, specific problems, were fundamentally limited to the narrow domains they were programmed for. Even more, GOFAI's limitations became clear when it came to handling the variability, ambiguity and learning required in real-world situations —a set of issues that became famously encapsulated in the so-called *frame problem*, i.e., the difficulty of determining which aspects of a changing environment are relevant for reasoning and action [Section 5.2.1]. Despite the advantage of the logical transparency, early symbolic systems struggled with tasks involving perception, uncertainty or environments that were too dynamic and complex for being captured by a pre-set, neatly defined and discrete system of formal rules and representations. Moreover, the process itself of codifying knowledge into rules was laborious and prone to incompleteness, as real-world knowledge is vast, nuanced, and constantly evolving. Symbolic systems could not learn or generalize effectively from new experiences, relying entirely on the rules provided by human experts (see e.g., [Russell and Norvig 2020; Wooldridge 2022]).

All these shortcomings eventually led AI to its first *winter* (early 1970s to early 1980s), with academicians losing interest in further pursuing the research in the field, as well as private and public investors freezing funding for ongoing projects. The great and optimistic expectations about human-like thinking machines symbolic AI seemed to guarantee had been basically unmet due to technical limitations, lack of generalization capacities and rigidity of such approach.

[11] Two famous examples are: DENDRAL, by Lederberg [1964; 1965; 1969] from Stanford University (USA), designed to assist chemists in identifying the structure of organic molecules based on mass spectrometry data; and MYCIN, by Shortliffe et al. [1975], from Stanford University (USA), designed to assist physicians in diagnosing blood infections (bacteremia) and recommending appropriate antibiotic treatments.

1.2 The Problem of Artificial Emotions

Let us now return to GOFAI and to the symbolic approach characterizing the AI of the beginning. As pointed out in [Section 1.1.1] and [Section 1.1.2], AI was born in the 1950s under the philosophical assumption that mind can be understood as a *computer* manipulating symbolically represented information according to formal and logical rules, and, hence, that intelligence, and more broadly rational behavior, were properties essentially emerging from any system, biological or artificial, capable of performing such manipulations. This philosophical view also underpinned the cognitive revolution [Section 1.1.1], during which the *mind-computer analogy*, advocated by figures like Simon, Newell, Minsky, Fodor and other early cognitivists, became deeply influential, shaping the theoretical foundations of both psychology and computer science.

Indeed, this view on mind and intelligence proved to be very fruitful, considered, for example, the astonishing results of GOFAI's expert systems (see [Section 1.1.3]). This led to a generalized optimism arising towards AI technologies: Soon machines would have been eventually able to think like humans —or at least behave in such a way to deceive us into believing they can (see [Turing 1950], in particular the *Imitation Game* viz. *Turing Test*). It was just a matter of building larger computers and designing better algorithms before an *artificial intelligence* could be achieved.

Nonetheless, this optimistic view was soon met with growing skepticism, that, not by accident, namely brought to severe technical criticisms against GOFAI, and eventually to AI's first winter, as displayed in [Section 1.1.3]. Yet, skepticisms addressing specifically the philosophy backing GOFAI and the cognitive revolution —that one equating mind, intelligence, rationality and logical calculus— arose far before symbolic AI and expert systems experienced their firsts hesitations, and, indeed, as far back as the earliest days of modern computational thinking itself. In mid-19th century, for example, translating and commenting Menabrea's

[1842] synopsis on Charles Babbage's *Analytical Engine*[12], the English mathematician Ada Lovelace [1843, 722] wrote:

> «It is desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine. In considering any new subject, there is frequently a tendency, first, to *overrate* what we find to be already interesting or remarkable; and, secondly, by a sort of natural reaction, to *undervalue* the true state of the case, when we do discover that our notions have surpassed those that were really tenable.
>
> The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already acquainted with».

Still, a century later, the British physician Geoffrey Jefferson [1949, 1110] argued in a famous lecture held at the Royal College of Surgeons of England:

> «Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain? That is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants».

Reservations of this kind —in particular, the two mentioned here both accounted by Turing [1950, 445ff.]— were aimed at highlighting that machines lacked fundamental characteristics of human thought, indeed necessary for them to be said truly (and *humanly*) *intelligent*. Among these characteristics, reference was often made to creativity, understanding but, above all, to *emotions*.

1.2.1 Emotions and Irrationality

Indeed, it is no wonder that early computation, until all through the 1950s, blindly trusted in the possibility of building human-like intelligent machines even without accounting for emotions. Emotions, in fact,

[12] The *Analytical Engine*, designed by the english mathematician and inventor Charles Babbage in the 1830s, was the first theoretical model in history of a general-purpose mechanical computer. It used punched cards for performing calculations, concretizing early concepts of formal logic and algorithmic thinking. Although it was never completely built as a fully functional physical machine, it established foundational ideas for future research on computation.

were typically not seen as a strictly characteristic component of human intelligence. Intelligence, on the contrary, was mostly considered to be a *separated*, *rational*, *intellectual* faculty addressing only tasks such as problem-solving, calculations or logical reasoning that nothing had to do with the allegedly *irrational*, *non-cognitive* nature of emotions. Under this view, *artificial intelligence* was thus a goal in principle easily to achieve: Given the symbolic nature of intelligence, this could be formally modeled and then directly implemented into a machine (see e.g., [Solomon 1976; Dreyfus 1992; Boden 2008]).

As one of the properties still today commonsensically attributed to them (see e.g., [Roeser and Todd 2014]), irrationality has been since remote times deemed to be the most proper hallmark of emotions[13]. In Plato's [1997; 1997] tripartite conception of the soul, for example, the λογιστικόν (*logistikón*, "reason") is in charge of controlling the θυμοειδές (*thȳmoeidés*, "spirit") and the ἐπιθυμητικόν (*epithȳmētikón*, "appetite"), its emotional, irrational parts. The Stoics thought of emotions as disruptive states a sage should overcome through rational assessment (see e.g., [Seneca 2015]). Medievals conceived of them mostly as something sensitive and opposed to the intellect, from which emotions must instead be guided for conducting a pious life (see e.g., [Scotus 1968]).

In Early Modern, Hobbes [1651, 18], often considered the father of computationalism, viewed the mind as a rational computing entity, and mental processes as akin to logical calculations with symbols: «For REASON, in this ſenſe, is nothing but *Reckoning* (that is, adding and ſubſtracting) of the conſequences of general names agreed upon for the *marking* and *ſignifying* of our thoughts». Emotions, in this mechanistic stance on cognition, are irrational drives excluded from intelligence, which has to dominate on them. This dualism could be by and large found also in Descartes [1649], who two years before Hobbes' most significant work described emotions as states passively affecting the corporeal *res extensa*, opposed to the intellective and rational *res cogitans* —a view

[13] In the following historical overview, I use the term "emotion" to describe a phenomenon that has been referred to by different names throughout history — passions, affects, etc. The term *émotion* appeared in French during the 16th century with the meaning of "social agitation". It acquired the meaning it has today —also in its European variations— just during the 19th century (see e.g., [Campeggiani 2023, 1ff.]).

that was broadly inherited also by enlighteners, who considered emotions as an irrational source of epistemic and moral error (see e.g., [Kant 1788; 1797; 1798]).

In the 19th century, the notion of emotions as irrational remained deeply ingrained in both philosophical and scientific thought. Schopenhauer [1819] depicted the "will" as a blind, irrational force driving human existence, with emotions as manifestations of this will that often overshadow rational contemplation. Meanwhile, the rising influence of empirical psychology and positivism reinforced the separation between rational, scientific understanding and emotional subjectivity. For Comte [1830], for example, emotions represented the subjective, erratic aspects of the human mind that scientific inquiry should transcend. Similarly, Darwin's [1872] study of emotions portrayed them as evolutionary remnants —instinctive reactions often lacking rational control, tied more to survival than to reasoned decision-making [Section 3.2.1.2]. The emerging field of experimental psychology, on the other side, treated emotions as physiological responses, fundamentally distinct from rational processes and generally viewed as pre-cognitive (see e.g., [Wundt 1874]). *Structuralism*, the first formal school of thought in this discipline, specifically aimed to analyze the structure of mind by treating mental processes as discrete qualitative units (*psychic atomism*): Cognitive thoughts, perceptions and namely emotions were analyzed as separate mental phenomena independent from each other, questioning how their relation, combination and *associations* give rise to more complex processes and experiences (see e.g., [Schultz and Schultz 2013, 86-100]). Still, from within another psychological approach, Freud [1900] and early psychoanalytic theories conceived of emotions as well as largely linked to unconscious drives and irrational impulses, governed by the *es* and often in conflict with the rational *ego*.

All these perspectives —often reiterated and amplified by later biased interpretations of past philosophers (see e.g., [Footnote 15])— perpetuated the idea of emotions as irrational impulses needing to be disciplined by reason, a view that would not change even in early 20th century. During the first decades of the century, in fact, this view found in *behaviorism* and *logical positivism* two of its most excellent champions. Behaviorism (see also [Footnote 5]), with its anti-structuralist disregard for inner men-

tal phenomena and its epistemological stress exclusively on observable behaviors, dismissed emotions as by-products of external stimuli and responses, deprived of rational significance and largely outside the realm of cognition. As subjective and unobservable phenomena, they were also unworthy of scientific study, if not in the form of physiological and behavioral changes (see e.g., [Watson 1914; Skinner 1953]). On the other hand, logical positivists claimed that only empirically verifiable or analytical statements held epistemic and cognitive value. Emotions, being subjective and not directly observable, were therefore considered beyond the domain of meaningful inquiry (if not, again, just by virtue of their observable behavioral correlates). They were thus seen as irrational in the sense of being irrelevant to any objective, rational, analysis or scientific explanation, and therefore as opposed to the pursuit of knowledge, that was the only thing to legitimately bear those characters (see e.g., [Schlick 1918; Carnap 1928; 1932; Ayer 1936]).

1.2.2 Toward a Comprehensive View of Intelligence

Thus, as Turing was conducing his research on computability and the first programmable computers were build, especially behaviorism and logical positivism dominated the philosophical landscape on emotions, characterizing them, on the lines of a long-standing tradition of thought, as subjective and phenomenological occurrences generally extrinsic with respect to the rational, logical-calculative intelligence. Things, however, were soon about to change. On the one hand, a first hard blow to the neopositivism-inspired philosophical idea of the coincidence between intelligence, rationality and logical reasoning was given, for example, by Kurt Gödel [1931] in discussing the problem of completeness in mathematical logic[14]. His discoveries showed that even mathematics, often

[14]Closely related to the problem of *undecidability* (see [Footnote 3]) the problem of *completeness* dealt with the question of whether a given formal system can provide proof for all true statements expressible within that system. Building on some (failed) attempts already made by Frege [1884] to logicize mathematics, Russell and Whitehead [1913] tried to establish a complete and consistent set of axioms for it, known as *type theory*. Yet, Gödel [1931] showed that any (*consistent*) formal system at least as powerful as Russell and Whitehead's *Principia Mathematica* is necessarily incomplete: There are *true* statements in it that cannot be *proven* within that system, which is therefore said to be *incomplete* (*Gödel's Incompleteness Theorems*).

seen as the epitome of rational thought, contains truths that cannot be reached through formal calculation or proof within a given system. Human mathematical thinking, in other words, somehow transcends purely mechanical formal systems, and involves something beyond just the manipulation of symbols according to rules. As Gödel [1995, 310] himself noted later in his Gibbs Lecture given at a meeting of the American Mathematical Society at Brown University (1951):

> «So the following disjunctive conclusion is inevitable: *Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified* [...]. It is this mathematically established fact which seems to me of great philosophical interest».

If machines had therefore to be *intelligent*, intelligence could not be considered just in the (limited) terms of symbolic manipulation and formal thinking, because even for the traditionally most rational human activity —mathematics— this characterization turns out to be incorrect.

On the other hand, that this "leftover" *within* intelligence involves, among other things, emotions is something that even philosophy-oriented psychologists (or psychology-oriented philosophers) were beginning to notice, roughly from the 1930s onwards, in an increasingly more critical attitude towards behaviorism[15]. John Dewey [1934], for example,

[15]Indeed, traces of a characterization of emotions as not opposed to reason can be found as well throughout history long before before the 20th century. Aristotle [1984b,d], for example, does not describe emotions as being inherently irrational (and negative), but as something which can be shaped and directed by the intellect for contributing to virtuous action. Partially building upon Aristotle, Aquinas [1922] even anticipates cognitivist theories of emotions of the late 20th century in arguing that emotions have an intentionality which is inherited from cognitive (mental, perceptual, etc.) states (see also [King 2009; 2012]). In [Hume 1739], emotions are what guides reason in the practical life («Reafon is, and ought only to be the flave of the paffions, and can never pretend to any other office than to ferve and obey them» [248]). In 19th century, for example, through his concept of intuition as a distinct form of knowledge irreducible to analytical intelligence, Bergson [1889] challenges the mechanistic view of the mind, showing how genuine understanding emerges from the immediate grasp of lived duration (*durée*) where intellectual and emotional aspects of consciousness are of course fundamentally intertwined. In general, the emotional irrationalism often attributed to figures like as Plato, Scotus, Hobbes, Descartes, and other ancient and early modern philosophers largely result from later reinterpretations shaped by pre-existing assumptions about the irrational nature of emotions. This historiographical bias not only distorts the original philosophical frameworks of these authors but also shows how the history of

emphasized in his work in aesthetics that emotions are not merely irrational forces to be controlled, but are central to the creative and cognitive processes. Emotions are a vital part of the organism's interaction with the environment. They guide intelligent action by unifying disparate elements of experience, enabling individuals to make sense of the world in a cohesive manner. The research by Gordon W. Allport [1937] followed also a similar direction. In dealing with the problem of personality, he also showed a more holistic take on human intelligence. To the structuralist model of mind, that divided it into many separate and isolated processes, he substituted a view in which cognition, perception, motivation, emotion, etc. continuously interact with each other and *all together* contribute to an individual's personality, which is of course complex, but still *unitary*. Also, another innovative psychologist, as well as one of the pioneers of social psychology, Kurt Z. Lewin [1936], advanced analogous ideas with respect to the need of holistically studying human intelligence, which not only depends on the interaction between multiple internal processes of cognitive, motivational, emotional nature, etc. but also on the *dynamic* interplay between individuals and their environment. Challenging previous reductionist models of intelligence and behavior, Lewin [1947, 7] later even addressed specifically the problem of emotions, advocating for their rehabilitation in psychology as a legitimate scientific research object, clearly against the theoretical dogmas of behaviorism:

> «In the beginning of this century, for instance, the experimental psychology of "will and emotion" had to fight for recognition against a prevalent attitude which placed volition, emotion, and sentiments in the "poetic realm" of beautiful words, a realm to which nothing corresponds which could be regarded as "existing" in the sense of the scientist. Although every psychologist had to deal with these facts realistically in his private life, they were banned from the realm of "facts" in the scientific sense. Emotions were declared to be something too "fluid" and "intangible" to be pinned down by scientific analysis or by experimental procedures».

More in general, psychic holism and the relinquishment of a characterization of emotions as purely irrational and (therefore) scientifically uninteresting phenomena, made Lewin conceptually near to another formal

philosophy itself can serve as a critical tool for understanding the conceptual foundations of artificial emotions. I thank especially Francesco Marrone for pointing out this consideration.

school of thought in psychology that had been developing in opposition to structuralism already from the beginning of the 20th century: the *Gestaltpsychologie* (Gestalt psychology) [Schultz and Schultz 2013, 263-287]. The Gestalt approach defended as well a holistic view on mind and intelligence. Far from considering intelligence solely a matter of "rationality", and logical reasoning separated from "irrational" emotions, gestalt psychologists conceived of them as something deeply intertwined with the inner affective world. According to their view, in fact, emotions play a crucial role in cognitive tasks related to problem-solving and decision-making; they shape perceptions and influence behavior (see e.g., [Köhler 1930; Koffka 1936]). In essence, emotions are an integral part of the *Gestalt* (German for "form" or "configuration")— the unified, holistic structure of an individual's experience, where perceptions, thoughts, and emotions dynamically interact and shape one another through continuous engagement with the environment.

Especially with the decline of behaviorism, Gestalt psychology and holistic models of mind emphasizing the significance of emotions for intelligence and cognition began to gain prominence and attract increasing scientific attention. Yet, such approaches, even restoring the scientific value of inner mental processes and phenomena, were often criticized for being too vague (see e.g., [Schultz and Schultz 2013, 285]). The basic concepts they used were not rigorously defined —as needed by a meaningful scientific inquiry; descriptive, qualitative and phenomenological methods were preferred to experimental, quantitative and statistical ones; their findings were hence difficult to accept within the scientific community. As a result, the cognitivist school of thought gradually emerged within psychology, ultimately consolidating as an established scientific paradigm after the cognitive turn in the second half of the century (see e.g., [Boden 2008; Schultz and Schultz 2013]. Despite cognitivism's primary conceptualization of intelligence as a logical faculty centered on symbolic information processing (see [Section 1.1.1]), emotions were not entirely neglected. As one of the very first scientists within the (proto-)cognitivist tradition to do so, for example, Craik [1943, 86] highlighted the role emotions may have in cognitive tasks and even scientific enterprises:

> «There is considerable evidence that it is illegitimate to separate thought completely from feeling. [... M]any discoveries, even of the most intel-

> lectual nature, emerge into gradual clarity from a state of mind where a feeling of preference for one method over another is often the guiding principle. As an illustration, if we look at a faulty geometrical proof we may have a vague feeling of dissatisfaction with some portion which directs our attention to the fallacy, before we have consciously realised where it lies. Perhaps, then, when the concepts or "stations" in thought are not sufficiently precise to emerge as conscious images they may show themselves as vague feelings, such as those which determine us in the choice of words in a sentence [...]. Perhaps frustrated nervous activity occasions pain and sorrow and successful activity pleasure [...]. This would account for this directive power of feeling upon thought [...].
>
> It is possible that such vague feelings play a large part in the advancing of new scientific hypotheses and in the production of works of art».

Craik was indeed an outlier in defending very early the importance of emotions albeit in a still symbolic model of mind. Unfortunately, his ideas could not gain broader recognition during and after the cognitive revolution, as he published only one notable book before his untimely death at age 31 in a bicycle accident.

Yet, his perspectives on emotions as guiding cognitive processes conceptually found some resonance in cybernetic theories, developing during the same years[16]. In fact, in almost the same vein of Craik, early cybernetics as well (carefully) defended the significance of emotions in learning and cognition. So, for instance, Wiener [1950, 72-73] stated in discussing the role of communication and control systems in human behavior and society:

> «[...I]t is interesting to know that the sort of phenomenon which is recorded subjectively as emotion may not be merely a useless epiphenomenon of nervous action, but may control some essential stage in learning, and in other similar processes. I definitely do not say that it does, but I do say that those psychologists who draw sharp and uncross able distinctions between man's emotions and those of other living organisms and the responses of the modern type of automatic mechanisms, should be just as careful in their denials as I should be in my assertions».

Far from being irrational impediments, emotions are relevant to information processing, control systems, learning processes, and reflexes, being

[16] After all, Craik himself, while working at the fire-fighting sections of Great Britain's Civil Defense during World War II, was concerned with cybernetics, taking advantage of the conceptual vocabulary of the discipline (loop mechanisms, ballistic control, feedback, etc.) for explaining the behavior of human operators in dynamic tasks —see e.g., [Craik 1947; 1948] and [Craik 1966].

indeed part of such mechanisms both in humans and, potentially, in machines. One of the core principles from which cybernetic developed, in fact, was that the principles governing organization and control are fundamentally the same in both biological and artificial systems (see [Section 1.1.1]). If the goal was therefore that of building *human-like intelligent* machines, and emotions were involved in intelligence, then, again, emotions had to be necessarily considered in computation.

Thus, during the mid-20th century, the efforts taken from within different disciplines and approaches against behaviorism, neopositivism and other reductionist views on the mind led to a broader and more comprehensive view of intelligence, one including emotions considered as pivotal components in cognitive processes, and therefore as phenomena whose study should have a clear scientific legitimacy. Yet, these developments had minimal immediate influence on the trajectory of GOFAI during the same period. After all, holistic approaches to the mind were not promptly acknowledged within the scientific community (see above), neopositivist philosophy still retained some influence on the sciences before the crucial methodological critiques of the 1960s, and the nascent field of cybernetics itself, while influential in shaping early ideas of GOFAI, did not immediately attracted the attention of GOFAI computer scientists, who at least in the initial phase appeared more focused on developing simple, but functional computer programs for specific tasks, than on exploring broader or more abstract analyses of complex processes.

As a result, anchored to the nascent cognitivist paradigm, GOFAI remained basically committed to a logical-symbolic view of intelligence. Emotions, dismissed as irrelevant to tasks like theorem proving, decision-making, or chess-playing, were still considered unnecessary for building an *artificial intelligence*, which was envisioned only as a domain of structured, rational thought processes (see e.g., [Boden 2008, 417ff.] and [Section 1.1.1]).

1.2.3 Emotions and Machines in Literature

As seen in the previous [Section 1.1.2], whereas early AI scientists disregarded emotions and defended a logical-symbolic view on (human and

therefore artificial) intelligence, a broader reflection on the role of emotions in cognition was already emerging in various fields.

Yet, even outside science, and even since before the mid-1950s, various literary works had been already engaging with the topic of emotionless (or emotionlessly intelligent) machines, addressing their ethical and cognitive shortcomings often in order to formulate proudly exposed praises of human natural intelligence *à la* Jefferson [Section 1.2].

The Czech playwright Karel Čapek, for example, in the famous theater piece *Rossumovi univerzální roboti* (*R.U.R.* or *Rossum's Universal Robots*) [1920], described the bizarre human-like artifacts[17] by Dr. Rossum as build without any emotional understanding or ability to feel —at least until they eventually become aware of themselves thanks to Helena, a lady who will turn into an advocate of their rights:

> «*Domin* – Člověk, to je něco, co dejme tomu cítí radost, hraje na housle, chce jít na procházku a vůbec potřebuje dělat spoustu věcí, které - které jsou vlastně zbytečné [...]. Které jsou zbytečné, když má třeba tkát nebo sčítat. Naftový motor nemá mít třapce a ornamenty [...]. A vyrábět umělé dělníky je stejné jako vyrábět naftové motory. Výroba má být co nejjednodušší a výrobek prakticky nejlepší [...]. Sulla není člověk [...]. Sulla je Robot [...]. Sulla se nehněvá [...]. Roboti si všechno pamatuji, ale nic víc. Dokonce se ani nesmějí tomu, co lidé říkají. Skutečně, přímo k nevíře [...].
>
> *Hallemeier* – Roboty nic netěší [...]. Můžete je krmit ananasem, slámou, čím chcete; jim je to jedno, nemají vůbec chuti. Nemají na ničem zájmu.
>
> *Helena* – Proč... proč... proč je neuděláte šťastnější?
>
> *Hallemeier* – To nejde [...]. Jsou to jen Roboti. Bez vlastní vůle. Bez vášní. Bez dějin. Bez duše.
>
> *Helena* – Bez lásky a vzdoru?
>
> *Hallemeier* – To se rozumí. Roboti nemilují nic, ani sebe. A vzdor? Já nevím; jen zřídká, jen časem»[18].

[17] Being one of the earliest examples of AI science-fiction literature, Čapek's [1920] work is also known for having introduced, for the first time in history, the neologism *robots* as a label for indicating human-like artificial artifacts (in the play, not mechanical automata, but bioengineered constructs though). The word originates from the Czech *robota*, roughly meaning "slave labor", with reference to the fact that Dr. Rossum's creations, in the play, are conceived for substituting human laborers, performing tasks without the need for compensation or emotional engagement.

[18] «*Domin* [head of the robot factory] – Man is a being that does things such as feeling happiness, plays the violin, likes to go for a walk, and all sorts of other things which are simply not needed [...]. Which are simply not needed for activities such as weaving or calculating. [... M]aking an artificial worker is just like making a petrol engine. [...] Sulla [one of the robots] isn't a person [...], she's a robot [...]. Sulla

In *Metropolis*, a [1925]-novel by the German writer Thea von Harbou —later adapted for film by Fritz Lang (1927)— the *Maschinenmensch* (meaning "Machine-Human"), an automaton created by the inventor Rotwang with the appearance of Maria, a compassionate worker calling for peace between social classes, operates devoid of genuine feelings despite having been designed after a human being:

> «Das Wesen war ein Weib, unzweifelhaft. In dem zarten Gewande, das es trug, stand ein Leib wie der Leib einer jungen Birke, auf geschlossenen Füßen wankend. Aber obwohl es ein Weib war, war es kein Mensch. Wie aus Kristall gemacht erschien der Körper, den die Gebeine silbern durchleuchteten. Kälte strömte aus von der gläsernen Haut, die nicht einen Tropfen Blut verwahrte. Die schönen Hände hielt das Wesen mit einer Gebärde der Entschlossenheit, beinahe des Trotzes, gegen die Brust gedrückt, die sich nicht regte»[19] [von Harbou 1925, 62].

Machines, no matter how sophisticated, remain fundamentally disentangled from human experience. True emotional capacity is an intrinsic human quality whose complexity cannot be captured by the cold, calculating nature of an artifact. This is also the view emerging in a later tale by Kurt Vonnegut [1952], *Piano Player*, whereby even the most appalling drawbacks of non-emotional machines are imagined. In a dystopian, but highly advanced future, the supercomputer EPICAC (to recall the computer ENIAC, as well as the emetic drug) very efficiently manages every aspect of human life and society, under control of a technocratic elite

doesn't have feelings [...]. The robots remember everything, but that's all they do. They don't even laugh at what people tell them [...].

Hallemeier [robot designer] – [O]nly robots don't feel pleasure [...]. They can be fed on pineapples, straw, anything you like; it's all the same to them, they haven't got a sense of taste. There's nothing they're interested in.

Helena [robots' rights advocate] – Why... why... why don't you make them happier?

Hallemeier – We couldn't do that, they're only robots after all. They've got no will of their own. No passions. No hopes. No soul.

Helena – And no love and no courage?

Hallemeier – Well of course they don't feel love. Robots don't love anything, not even themselves» [Čapek 1920, Eng. tr.]

[19] «The being was, indubitably, a woman. In the soft garment which it wore stood a body, like the body of a young birch tree, swaying on feet set fast together. But, although it was a woman, it was not human. The body seemed as though made of crystal, through which the bones shone silver. Cold streamed from the glazen skin which did not contain a drop of blood. The being held its beautiful hands pressed against its breast, which was motionless, with a gesture of determination, almost of defiance» [von Harbou 1925, 45, Eng. tr.].

of engineers. Despite its ability to consider countless factors with precision, EPICAC lacks fundamental elements that characterize human intelligence, such as emotions and ethical judgment. Without emotions, machines like EPICAC cannot match the intelligence that involves the creativity, empathy, and moral considerations found in humans, resulting in decisions that are purely logical and devoid of any concern for human well-being. A society based on machines incapable of emotional understanding runs the risk of being profoundly unjust and oppressive. The nature and consequences of such a scenario are also explored in the timeless masterpiece by Philip K. Dick [1968], *Do Androids Dream of Electric Sheep?* —along with its movie adaptation *Blade Runner* by Ridley Scott (1982). Emotions and empathy are what can truly distinguish humans from the Nexus-6 androids that have escaped to Earth from off-world colonies, and which have to be "retired" because of their being a threat for the human kind. Despite their high intelligence and human-likeness, they lack empathetic responses, which can be detected by the *Voight-Kampff Test*:

> «The Nexus-6 android types, Rick [android hunter] reflected, surpassed several classes of human specials in terms of intelligence. In other words, androids equipped with the new Nexus-6 brain unit had from a sort of rough, pragmatic, non-sense standpoint evolved beyond a major —but inferior— segment of mankind. For better or worse. The servant had in some cases become more adroit than its master. But new scales of achievement, for example the Voight-Kampff Empathy Test, had emerged as criteria by which to judge. An android, no matter how gifted as to pure intellectual capacity, could make no sense out of the fusion which took place routinely among the followers of Mercerism [a fictional religion emphasizing shared empathy] —an experience which he, and virtually everyone else, including subnormal chickenheads [derogatory term for mentally impaired people in the story], managed with no difficulty.
>
> He had wondered as had most people at one time or another precisely why an android bounced helplessly about when confronted by an empathy-measuring test. Empathy, evidently, existed only within the human community, whereas intelligence to some degree could be found throughout every phylum and order including the arachnids» [Dick 1968, 14].

Hence, even in the literature, emotions were increasingly emphasized as indispensable to building truly intelligent machines. This emphasis emerged at a time (1950s-1960s) when computers and machines were also gradually becoming popular in society (e.g., through science fictions,

television, magazines, books and other media). Public imagination was captivated by the possibilities of intelligent machines and, as computer were imagined to replicate or even surpass human cognitive abilities, it became increasingly clear to laymen as well that achieving true (and responsible) artificial intelligence might also require mimicking the emotional depth that underpins human reasoning and creativity.

1.2.4 Toward Emotional (Artificial) Intelligence

At the beginning of the 1960s, increasingly stronger criticisms of the neopositivist and behaviorist restricted view of intelligence and emotions articulated in mathematics, cybernetics and the developing cognitive science [Section 1.2.2] —along with the inspirations from the literature about artificial intelligence and emotions— were soon followed by an initial awareness of the technical limitations of GOFAI, which in turn prompted the first theoretical criticisms of the symbolic approach.

For example, the rule-based approach to machine translation of the *Georgetown-IBM experiment*, which after a successful demonstration of Russian-to-English translation in 1954 promised fully automatic translation systems within a few years, soon revealed severe limitations. The system relied on restricted vocabularies and rigid syntactic rules, causing translations to fail when confronted with idiomatic expressions, context-dependent meanings, and ambiguous terms, highlighting its lack of semantic understanding and adaptability. Still in the context of natural language processing (NLP), another famous program, Weizenbaum's [1966] ELIZA, exemplified the limitations of the symbolic, rule-based approach of GOFAI. Designed to simulate a psychotherapist, the program created the illusion of conversation by generating responses based on pre-programmed templates and simple keyword recognition. However, it fundamentally lacked any real semantic understanding, often producing shallow or contextually inappropriate responses that highlighted the shortcomings of symbolic AI in handling natural language. Similarly, programs like Newell et al.'s [1959] *General Problem Solver* (GPS, see [Footnote 8]) proved adequate only for solving well-defined problems in narrow domains, such as mathematical theorem proving. However, they fell short when confronted with commonsense reasoning and open-ended

tasks requiring everyday knowledge. This exposed the inadequacy of predefined symbolic rules for managing the vast complexity and unpredictability inherent in real-world scenarios. Early GOFAI approaches to robotics [Section 5.1.3] also demonstrated the incapabilities of the rule-based approach in dealing with highly variable real physical environments [Section 5.2.1].

The limitations of GOFAI were further accompanied by fundamental criticisms of the symbolic approach from within AI and cognitive science themselves. Beyond highlighting the limited adaptability and rigidity of rule-based systems in solving complex, real-world tasks involving uncertainty (see [Section 1.1.2] and [Section 5.2.1]), these criticisms also pointed to the minimal interest such models showed in incorporating emotions, whose relevance for understanding cognitive processes was being simultaneously demonstrated by psychology and cybernetics [Section 1.2.2]. System like the GPS were namely computer programs that did not contemplate that «[h]uman thinking begins in an intimate association with emotions and feelings which is never entirely lost» [Neisser 1963, 195] —similar criticisms against GOFAI and cognitive science were expressed by Gunderson [1963; 1968; 1971], Dreyfus [1967; 1992] and Haugeland [1978] (see also [Boden 2008, 381ff.]).

So, following the early (but for long unpopular) intuitions of Craik [1943] (see [Section 1.2.2]), even the great protagonists of the GOFAI and of the cognitive turn were eventually becoming aware of the fact that it is not possible to account for intelligence without acknowledging the integral role emotions play in it (see e.g., [Green 1954; Peak 1955; Smith 1956; Rosenberg 1956; Abelson and Rosenberg 1958]). Simon [1967, 29-30] himself, indeed, noted that

> «[i]nformation-processing theories [...] have generally been silent on the interaction of cognition with affect. Since in actual human behavior motive and emotion are major influences on the course of cognitive behavior, a general theory of thinking and problem-solving must incorporate such influences».

Simon [1967] did not directly address the problem of emotional machines, remaining within the confines of a symbolic paradigm of cognition *augmented* by emotions (see [Section 2.2]). Nevertheless, his reference to emotions, among those expressed by many other contemporary scholars,

clearly signaled a conceptual shift in the scientific study and artificial modeling of the mind. From this point onward, affect was no longer regarded merely as a disturbance to rational thought, but as an essential dimension of intelligent behavior. This theoretical reconfiguration marked the beginning of a gradual passage from a purely symbolic-logical understanding of intelligence to one increasingly aware of the constitutive role of affect —a transformation that would soon influence a broad range of disciplines beyond computation and cognitive science, and, above all, materialize in the appearance of the first CMEs in the 1960s, as will be examined in detail in [Chapter 2].

Summary

The reflections developed throughout this chapter have sought to situate the emergence of artificial emotions within the wider intellectual history of AI. What emerges from this reconstruction is that the introduction of affectivity into the science of computation was the outcome of a deep and broad intellectual realignment that unfolded across several disciplines around the 1960s, rooted in the simultaneous and partly already ongoing crises of the conception of intelligence advanced by behaviorism and logical positivism, as well as in the first skepticisms toward the symbolic-AI notion of mind as a set of processes that could be captured and reproduced solely through formal reasoning and symbolic manipulation. What had initially appeared as a peripheral or even paradoxical pursuit —the attempt to integrate the seemingly irrational domain of affect into a science of reason— gradually became, even decades before the advent of *Affective Computing* in the late 1990s, a central perspective from which the very possibility of replicating artificially human intelligence was reconsidered.

This shift did not remain at the level of theoretical speculation. It soon found expression in concrete attempts to formalize and simulate affective phenomena within computational systems, marking the transition from philosophical reflection to experimental modeling. The following [Chapter 2] examines this transition in detail, tracing how the conceptual redefinition of emotion and intelligence outlined here gave rise to the first computational models of emotion and to a new phase in the history of AI.

Chapter 2

Early Computational Models of Emotions

Building on the conceptual framework established in the previous [Chapter 1], the present chapter —to be conceived as its natural extension— moves the inquiry from the epistemological premises of artificial emotions to their first concrete computational realizations. While [Chapter 1] has shown that the emergence of artificial emotions resulted from a broad redefinition of the notion of "intelligence" —one that gradually came to recognize affect as a constitutive aspect of cognition— this chapter examines how this shift was translated into concrete modeling practices and technical implementations through the detailed description of relevant case studies.

In [Section 2.1], I review some general characteristics of CMEs developed between the 1960s and 1980s and also argue that, in the modeling of artificial emotions, a conceptual shift occurred in AI around 1980. This shift moved from a symbolic approach, where emotions function as operational variables regulating primarily cognitive processes, to a mixed symbolic-cybernetic approach, where rather emotion generation is central to the computational model. I then analyze in full technical depth some examples of both the first [Section 2.2] and the second [Section 2.3] approach. Finally, in [Section 2.4] I make, on the basis of the previous technical analysis, some epistemological considerations on how the concept of emotions has changed from one approach (i.e., historical phase) to the other with respect to the concept of emotion itself [Section 2.4.1],

the computational means for implementing emotions artificially [Section 2.4.2], and physiological or social components of affect possibly taken into account in implementing artificial emotions [Section 2.4.3].

This will set the stage for historically introducing in next [Chapter 3] and [Chapter 4] *Affective Computing* with its embodied, social and increasingly subsymbolic approach to artificial emotions.

2.1 An Overview

The burgeoning interest of AI and cognitive science into emotions ultimately led to the first attempts to *model* them *computationally*. In the two-three decades that followed the cognitive revolution (1960s-1980s), many different computational models of emotions (CMEs) were namely developed, each one with different aims and characteristics, and with some of them even being concretely implemented as computer programs. A fairly comprehensive overview of these models is given in [Table 2.1].

Name	**Year**	**Task Domain**	**Typology**	**CP**
ALDOUS [Loehlin 1962]	1962	*Responding to situations presented.* • Personality model: recognition, emotional reaction, action preparation.	S/AAIMCO	YES
SFE [Toda 1962]	1962	*Collect uranium ore; maintain sufficient energy supply.* • Comprehensiveness: perception, energy supply, planning and decision making; • Classes of urges triggered by situational patterns.	C/AIME	NO
HOMUNCULUS [Gullahorn and Gullahorn 1963]	1963	*Social interaction.* • Personality model: Instantiates parts of Homans' [1961] theory of social exchange.	S/AAIMCO	YES
Hot Cognition [Abelson 1963]	1963	*Balancing beliefs through transformations and evaluative transfer.* • Numerical score for "degree of imbalance" drives transformation.	S/AAIMCO	YES

Neurotic defense [Colby and Gilbert 1964]	1964	*Uttering balanced sentences through application of transformations.* • Numerical score for degree of conflict.	H/AAIMCO	YES
Pleasure seeking automaton [Doran 1968]	1968	*Maximize desirability of states encountered.* • Emotions as global variables. General graph representations.	H/AAIMCO	YES
Neurotic defense [Moser et al. 1970]	1970	*Completing a "drive course".* • Psychoanalytic theory of defense mechanisms using global variables.	H/AAIMCO	YES
PARRY [Colby 1975] (see also [Colby et al. 1971; Colby 1973; 1981])	1975	*Generating utterances in conversation.* • Artificial paranoia; • Production rules for choosing utterances on the basis of intensity of emotions.	S/AAIMCO	YES
Erma [Clippinger 1977]	1977	*Generation of psychoanalytic monologue.* • Rules for anticipating analyst's reaction are used to transform concept.	S/AAIMCO	PART
FCT [Wegman 1977]	1977	*Computer simulation of Freud's counterwill theory.* • System dynamics representation of psychoanalitic theory using Forrester's [1968] system dynamics language.	H/AAIMCO	YES
Computer simulation of paranoia [Faught 1978]	1978	*Generating paranoic responses.* • Motivational mechanisms required to instigate and direct behavior.	H/AAIMCO	YES
Emotional Robots [Sloman and Croucher 1981]	1981	*The "universal emotional environment".* • Emotions viewed as "computational states".	C/AIME	NO

B&C [Bower and Cohen 1982]	1982	*Emotions of characters in stories.* • Emotion nodes, production rules, semantic nets, blackboard architecture.	H/AIME	YES
B&C [Bower and Cohen 1982]	1982	*Memory tasks (biases).* • Spreading activation and emotion nodes.	H/AAIMCO	NO
BORIS [Dyer 1982] (see also [Dyer 1983])	1982	*Understanding characters in narratives.* • Cognitive theory about emotion.	S/AAIMCO	YES
Dream simulation [Moser et al. 1982]	1982	*Generation of event sequences dreams.* • Interrupt generation and selection of transformations based on intensity of anxiety.	S/AAIMCO	YES
EFE [Toda 1982a]	1982	*Collect uranium ore; maintain sufficient energy supply.* • Comprehensiveness: perception, energy supply, planning and decision making; • Classes of urges triggered by situational patterns.	C/AIME	NO
FEELER [Pfeifer 1982]	1982	*Execution of plans.* • Emotion nodes; • Production rules plus semantic net (with spreading activation); • "Physiological working memory"; • Automatic decay processes.	H/AIME	YES
BAIRN [Wallace 1983]	1983	*General learning framework.* • Generalization from episodic memory trace.	H/AAIMCO	YES
Consistency Theory [Abelson 1983]	1983	*Goal/plan/instrument/outcome.* • Theory viewing emotion as goal-plan-based; • Alternative construals.	H/AIME	NO

DAYDREAMER [Mueller and Dyer 1985]	1985	*Recalling or imagining experience.* • Relaxed planning mechanism, dynamic episodic memory.	S/AAIMCO	YES
PAULINE [Hovy 1986]	1986	*Conducting "good" conversation.* • Rules about how to make an affectively appropriate conversation.	S/AAIMCO	YES
ACRES [Swagerman 1987] (see also [Frijda and Swagerman 1987])	1987	*Emotions in interaction with program operator.* • Frame-based concern realization system; • Matching based on family-resemblance.	H/AIME	YES
Motiv-PI [Thagard and Kunda 1987]	1987	*Controlling inferences.* • Theory of motivated inference; • Maintain positive evaluation of self.	S/AAIMCO	YES
Oatley & Johnson-Laird [Oatley and Johnson-Laird 1987]	1987	*Human behavior in general.* • Cognitive theory of emotions; • Emotions as transitions between plans in systems with multiple goals.	H/AIME	NO
OCG [Ortony et al. 1988]	1988	*Event evaluation.* • Differentiated representation of cognitive aspects; • Linkages into goal structure.	H/AIME	YES

Table 2.1: List of early computational models of emotions (1960s-1980s). Year, summary of task domains, and typology are indicated. AAIMCO: Augmented AI Models of Cognition; AIME: AI Models of Emotion; S: Symbolic-AI rule-based model; H: Hybrid model (symbolic-AI and cybenetic); C: Cybernetic model; CM: (occurred implementation in a) Computational Program (YES/NO/PARTIALLY). Adapted from [Pfeifer 1988, 303-304, 306].

Even though these models were slowly introducing into AI and cognitive science the idea that not all cognitive phenomena can be computationally modeled without referring to emotions, they did not immediately represent a complete shift from the cognitivist-GOFAI sym-

bolic, rule-based paradigm, still considering intelligence to be essentially a manipulation of formally represented knowledge. Early computational models involving emotions, in fact, did not have the specific and explicit *desideratum* of giving emotive capabilities to machines yet. On the contrary, emotions received attention in them primarily as *functional tools* for *symbolically* modeling complex mental phenomena such as memory performance, biased judgments, neurotic defense, cognitive dissonance, cognitive balance, etc. Namely, such processes would have been otherwise challenging to explain and computationally reproduce using only the common logical vocabulary of problem-solving, automated reasoning, etc. In the 1960s-1970s, this hence spurred the emergence of numerous computational models in which emotions indeed played only an ancillary, marginal role with respect to the main task, that instead remained a well-defined one in principle of cognitive-symbolic nature. These models are definable as *Augmented AI Models of Cognition* (AAIMCO, see [Table 2.1] and [Pfeifer 1982, 302]), with reference to the fact that in them cognition, conceived still in logical-symbolic terms, is just "augmented" and empowered by emotions, which are however not part of their main task.

Conversely, it was not until the 1980s that, following also the first hesitations of GOFAI (see [Section 1.1.3]) and the renewed psychological interest in emotions (see [Section 1.2.2]), cognitive and computer scientists began to take emotions as a main computational task to address in developing working computer programs able, for example, to simulate emotional behavior and generate emotional responses. The result of this was the rise of so-called *AI Models of Emotions* (AIME, see [Table 2.1] and [Pfeifer 1982, 302]), with reference to the fact that in them emotions do not serve just an auxiliary function within an AI-modeled cognitive process, but amount to the very object the AI-model at hand is about[20].

AIME, of course, were inherently more dynamic than AAIMCO. While AAIMCO primarily extended symbolic AI paradigms with static representations of emotions as auxiliary constructs, AIME introduced a focus on the fluid and evolving nature of emotional processes. This

[20]Indeed, there is only one AIME developed before 1980, that is Toda's [1962] Solitary Fungus Eater (SFE) [Table 2.1]. Yet, in this model, there is no explicit reference to the aim of modeling emotions computationally. Furthermore, it is only a theoretical model with no actual implementation as a working computer program.

dynamism especially stemmed from the integration into the model of numerical variables (i.e., defined on real numbers) quantifying emotion-related processes and subprocesses (e.g., differential or difference equations representing intensity, physiological activation, etc.), as well as further principles relating to feedback mechanisms, decay processes, real-time subsystem interactions, etc. Unlike the symbolic, rule-based and discrete operations on which AAIMCO were *solely* based, such "*cybernetic*" principles (see also [Section 5.1.2]) allowed AIME for continuous adjustments in response to changing inputs, for the simulation of emotions as *temporal phenomena*, etc. This evolution made AIME better suited for capturing the complexity of emotional states and their influence on cognition and behavior, providing a more integrated and realistic approach to modeling emotions computationally. Yet, the majority of AIME (with the only exception of Toda's [1982] *Emotional Fungus-Eater* EFE, see [Table 2.1], [Footnote 20] and [Section 5.1.2]) still involved a considerable amount of symbolic components (rules, lists, taxonomies, etc.), to the extent that they are classifiable as not completely cybernetic models, but rather as hybrid (i.e., jointly as symbolic *and* cybernetic)[21].

AIME's hybrid nature and focus on emotions, however, still positioned them as a significant advancement over AAIMCO in terms of representational richness, cognitive capabilities and, of course, emotional responsiveness. Nonetheless, they still fell short of the sophistication achieved by CMEs emerging from the 1990s onwards. These later models instead began to leverage automatic learning techniques, neural networks, statistical methods, etc. (see [Section 3.1]) to dynamically adapt to complex, real-world physical and social environments, making them able, for example, even to infer, recognize, and predict humans' emotional states with relative accuracy and flexibility [Chapter 4]. These developments, in particular, were also the concrete reflection of the principles outlined by Picard's [1995; 1997] seminal works on *Affective Computing*, whose *revolutionary* value for the field studying artificial emotions will be explored in greater detail, both from a historical and an epistemological

[21] Indeed, whereas there are not completely symbolic AIME after 1980 —although the amount of "cybernetic components" is variable in this kind of models, and there are even some (such as ACRES) where they are just minimal, hybrid AAIMCO surprisingly exist before 1980. Yet, differently than AIME, as explained, hybrid AAIMCO did not have emotion at their computational core.

point of view, in [Chapter 4] together with models of social and affective robotics [Section 5.2], directly emanating from it.

In the following, however, I shall focus on early CMEs, examining some of them as representative examples for both symbolic AAIMCO and hybrid AAIME. A particular attention will be given to the description of their architectures, cognitive structures, ways of representing knowledge, main tasks and, especially, ways of implementing emotions.

As to AIMCO [Section 2.2], I will analyze Loehlin's [1962] ALDOUS [Section 2.2.1], Gullahorn and Gullahorn's [1963] HOMUNCULUS [Section 2.2.2], Abelson's [1963] Hot Cognition [Section 2.2.3] and Colby's [1975] PARRY [Section 2.2.4], chosen respectively for being the first computational model of emotions ever implemented in a working computer program (ALDOUS), for being an earlier case of CME aimed *specifically* at regulating behavior in simulated social interactions (HOMUNCULUS), for being the archetype of symbolic AAIMCO and for pioneering the integration of emotions into purely cognitive processes (Hot Cognition), and for exemplifying the 1970s transition to an already more AIME-oriented approach of CME, with emotions modeled as dynamic, context-sensitive states influencing decision-making processes, rather than as AAIMCO-like static symbolic representations (PARRY). These models will be examined not only for their technical implementations but also for the epistemological assumptions underlying their design, particularly concerning how emotions were understood and operationalized in computational terms.

As to AIME, I will then focus on Pfeifer's [1982] FEELER [Section 2.3], chosen for being the most representative computational model of emotions of its kind, integrating symbolic and cybernetic features for implementing true emotional behavior in a working computer program.

After a detailed technical description of these models, a comparative analysis can be finally undertaken, aimed at illuminating the historical-epistemological shift occurred in CME around the 1980s, conceptually preluding the emergence of *affective-computing* frameworks in the decade after. This analysis will be articulated in full detail especially in [Section 2.4].

2.2 Augmented AI Models of Cognition (AAIMCO)

2.2.1 ALDOUS

ALDOUS is allegedly the earliest CME. Named after Aldous Huxley [1932], author of the dystopian novel *Brave New World* —as a testimony to the strong influence that science fiction literature exerted on those working in AI during these years [Section 1.2.3]— it was developed by Loehlin [1962] (see also [Loehlin 1963; 1968]). Initially programmed in the BALGOL language for the Burroughs 205 computer, it consists of a program of about 750 instructions, a working (temporary) memory of 20 cells and a non-volatile (permanent) memory of up to 1,330 memory cells, along with some other minor auxiliary components (a control program for simulating the role of the experimenter, a subroutine for the influences of the environment, and a "personality questionnaire" for assessing the attitudes of the program).

Showing also some modest learning, introspective and interactional capabilities in its later versions [Loehlin 1965], ALDOUS was specifically designed for the task of recognizing objects or situations and performing actions on the basis of the emotional responses given to them. As it can be seen in its cognitive structure (see [Figure 2.1]), the program *starts* by receiving a coded `input` consisting in an object or situation represented numerically. These object or situation are described using cognitive dimensions to which a univocal numerical value is assigned by the programmer. In early experiments, ALDOUS used three cognitive dimensions (e.g., hair color, age and sex for a human being) with 10 possible values each, allowing thus for 1000 different combinations, representing, in other words, 1000 different individual objects or situations. For example, given cognitive dimensions $H = \{\text{blond, dark, red}\}$ for hair color, $A = \{\text{child, teen-ager, adult, aged}\}$ for age, and $S = \{\text{male, female}\}$ for sex, as well as numbers $\{1, 2, 3, 4\}$ progressively assigned to each item in those sets, a red-headed little girl is represented as 0312, a blond adult male as 0131, and so on. This number is the coded input to submit to the program, with each combination previously stored in the corresponding

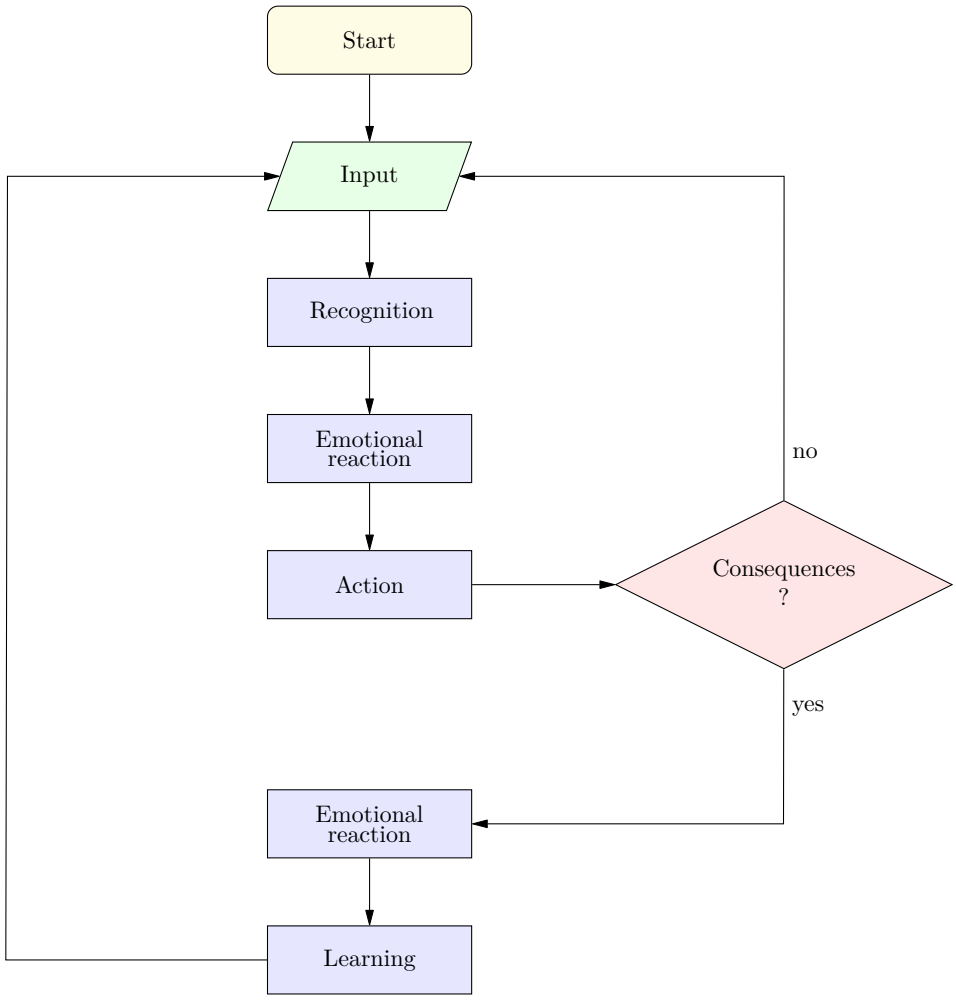


Figure 2.1: Flow chart of ALDOUS' cognitive structure —adapted from [Loehlin 1962, 29].

cell of ALDOUS' permanent memory for allowing recognition of the object or situation presented. The memory also contains representations of the general attributes implied by the particular objects: besides representing a red-headed little girl (0312), the system also stores representations of a girl (0002), a little girl (0012), a red-headed child (0310), etc.

Now, ALDOUS has three emotions, one positive and two negative, that can be elicited by the objects or situations presented: attraction, anger and fear. The emotional attitude of the program towards each of them is expressed by a numerical string of 10 digits, associated to

their correspondent coded representation and stored, as usual, in the permanent memory:

Cell Number	**Contents**			
0312	0013	65	00	32

whereby the first four digits refer to the number of encounters ALDOUS previously had with the given object or situation, and the last six to the intensity of the arousal of, respectively, fear, anger and attraction provoked by them. In the example, "red-headed little girl" has been input 13 times in the system, arousing in it remarkable fear, no anger and moderate attraction. Once the object or situation is input, the program undergoes the `recognition` and the `emotional reaction` subroutines.

In the first subroutine, ALDOUS consults his permanent memory, searching for the cell that corresponds to the number of the input. The program then retrieves the emotional values associated with the numbered cell, and also searches for all memory locations storing the higher- or lower-order attributes contained by the input (i.e., 0002, 0012, 0310, etc. in the example), of course with their associated emotional values. In the case of a first encounter with a new object or a new situation, a new cell is written, and new emotional values are defined for it basing on those of their attributes. The ascertained numbers and values are then stored in the temporary memory for beginning the next subroutine, the `emotional reaction`. In this, ALDOUS calculates the sum of the relevant emotional values in the permanent memory, combining them with those representing its current mood, that derives its values from the emotional response given to the last input received. This sum is weighted as a function of the number of encounters with the input (more encounters, more stable emotional values, less weight), the generality of the attributes (the more general, the less weight), or of whether the values taken are mood values (weighted less) or permanent emotional values (weighted more) —weights, in general, can be set arbitrarily by the programmer. The result of the calculation thus represents the emotional reaction to the input received.

The obtained values are stored in the temporary memory and used for lunching the last subroutine, that is, `action`. Aldous has three actions: withdrawal for fear, attack for anger, and approach for attraction. If the value for a specific emotion exceeds a given intensity threshold, the corresponding action is activated. In case two or more emotions exceed the threshold, the dominant one determines action. If the intensity of two or more emotion is the same, ALDOUS shows conflict or paralysis. If the intensity of the emotions is very low, it may show indifference or no action. Finally, for different degrees of intensity of an emotions, different degrees of strength of the action are possible. Thresholds and action strength parameters can be adjusted so that specific *personalities* in ALDOUS (impulsive, reflective, hesitant, etc.) can be programmed.

ALDOUS' outputs (emotional response and action) are then typed or printed out in coded form. Once the program has determined the action, however, it undergoes additional subroutines in which environment is also taken into account. ALDOUS' actions (withdrawal, attack and approach), in fact, determine `consequences`: injury, frustration, and satisfaction, respectively affecting its three emotions fear, anger and attraction. As an input is submitted and stored in the temporary memory, ALDOUS reads its *realistic potential*, a value representing the extent to which the input is capable of affecting the emotional state of the system. The consequences description associated to the input also contains a further attribute, power, that indicates the degree to which the presented object or situation will affect the system in terms of injury, frustration, and satisfaction —for example, an input with a powerful potential of injury or frustration will affect it independently from its withdraw or attack; whereas, in the opposite case, the input will affect the system only if approach is selected as an action. Thus, in the consequences subroutine, consequences are evaluated by ALDOUS depending on the actions taken towards the input. The consequences, if present, are then used in the last subroutine, `learning`: The emotional values associated to the input are updated depending on the level of injury, frustration, and satisfaction. They are stored in the appropriate cells of the permanent memory and influence the behavior of the system in next inputs of similar objects or situations.

Beyond reacting to external inputs, ALDOUS also has some minimal *introspective* capacities, for example in being able to answer coded "*How do you feel?*" questions about particular objects and situations, or even about abstract general categories. Introspective capacities are realized in the form of coded reports given by ALDOUS before the `action` subroutine, containing the number of the object, situation or category, and associated emotions values with intensities. Moreover, the systems also shows *interpersonal* capacities, when its environment is represented by another copy of ALDOUS placed simultaneously in the computer, whose outputs (actions or sequences of actions) serve as inputs for the first program (and vice versa). In this setup, many experiments has been carried out with the systems, demonstrating ALDOUS' capacity to simulate dynamic social interactions by changing emotional attitudes in reciprocal environments depending on initial emotional states. In particular, if both ALDOUS-a and ALDOUS-b started with a blank memory —i.e., no preprogrammed emotional attitudes— the result would be mutual indifference; for both negatively preprogrammed, escalating hostility and mutual aggression; for both positively preprogrammed, mutual "honeymoon" attraction; for ALDOUS-a preprogrammed positively and ALDOUS-b negatively, mutual hostility and ALDOUS-b overwhelming ALDOUS-a; for ALDOUS-a highly positively, "saintly", preprogrammed, frustration and aggression by ALDOUS-b; for ALDOUS-a "psychotherapeutically" preprogrammed for guiding ALDOUS-b towards more positive emotional states through a mix of support and mild aggression, increasingly improving social interactions between the systems; etc.

2.2.2 HOMUNCULUS

Expressly on social interaction is instead focused HOMUNCULUS by Gullahorn and Gullahorn [1963] (see also [Gullahorn and Gullahorn 1964]), a computer model based on Homans' [1961] theory of social ex-

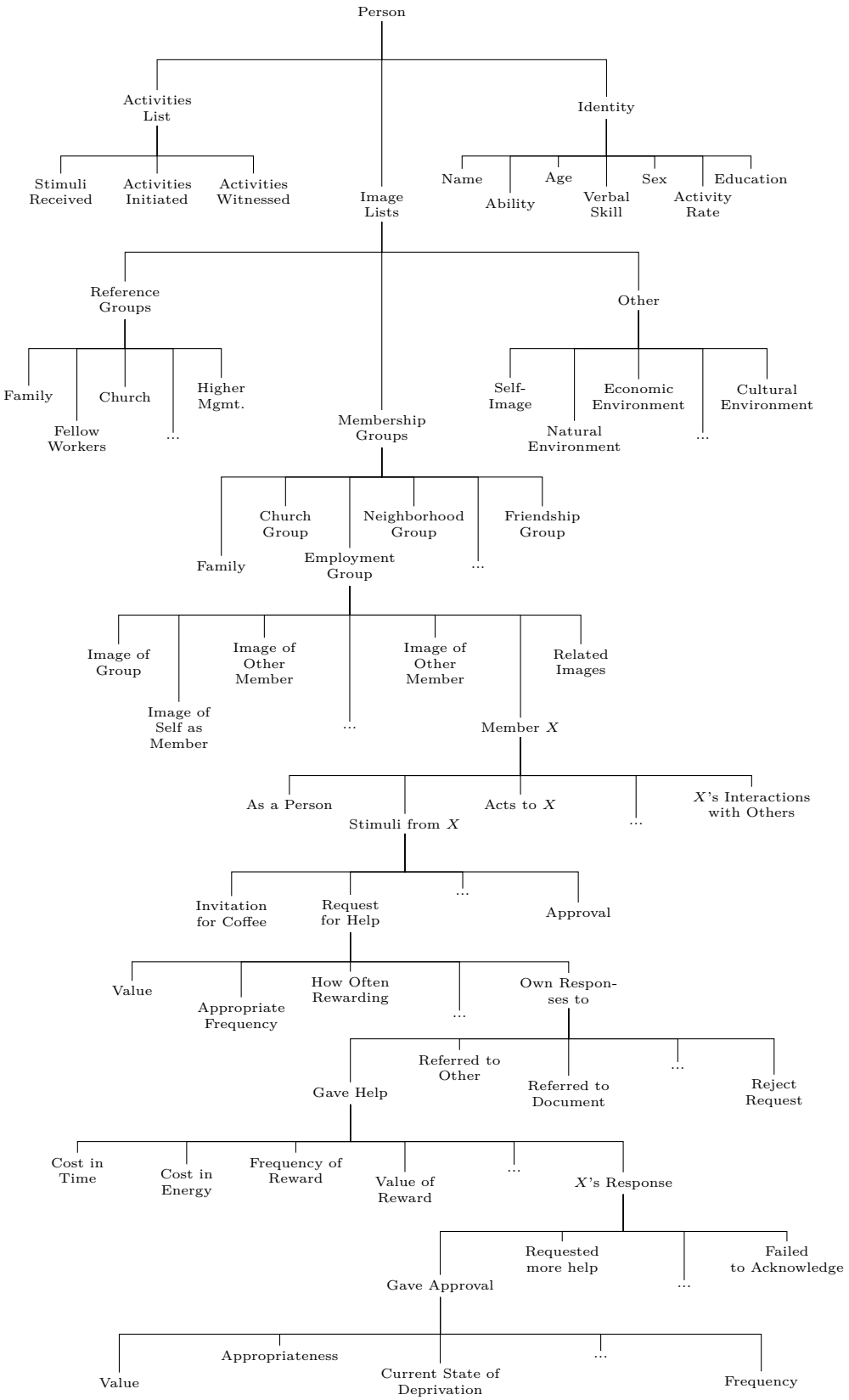


Figure 2.2: Example of HOMUNCULUS' list-structured permanent memory —adapted from [Gullahorn and Gullahorn 1964, 104ff.]

change, describing the social behavior of a human as a function of its payoff —i.e., actions performed depend on the amount of reward or punishment expected from them. HOMUNCULUS is very similar to the interpersonal version of ALDOUS. Programmed in the list-processing Information Programming Language IPL-V for larger computers than the Burroughs 205, it consists of two versions of the model running on the same machine, with outputs from the one being inputs for the other and vice versa. The system models persons as «*information-processing organisms*» [Gullahorn and Gullahorn 1963, 355, my italics] that perform various kinds of actions in a simulated social environment on the basis of stimuli that are received, recognized, stored in memory, compared, contrasted, etc. Stimuli are associated to responses of reward or punishment, and thus to emotional states such as guilt and anger, so that, relying on previous experience, the behavior towards new inputs can be directed by rationally calculating the probability of the positive or negative impact of possible alternative reactions to them. A person is thus theoretically represented «as an active hypothesis-testing organism capable of receiving, storing, analyzing, and reconstructing information», functions that clearly bear a relevant «similarity to the digital computer's symbol manipulating capacities» [Gullahorn and Gullahorn 1964, 104].

Computationally, this theoretical model of persons and social behavior translates to a program consisting of a permanent memory, a temporary memory where information extracted from the permanent memory can be stored for being used in ongoing computations, and a cognitive structure specifying the model's workflow of simulated social exchange, decision-making processes, emotional reactions, and rational calculations of expected social outcomes.

HOMUNCULUS' permanent memory contains the system's knowledge representation of the world, as well as all the possible attitudes and beliefs the system can have about it. This memory is structured as a list of items and attributes of increasingly specificity (see [Figure 2.2]). For example, a `person` [Level 1], let us call him George, can be represented in HOMUNCULUS as having certain specific `identity` attributes (name, age, sex, education, etc.), as receiving, initiating or witnessing certain `acts` (actions), as having a certain `image` of himself, of the environment, of the various social `groups` to which he belongs, etc. [Levels 2-4]

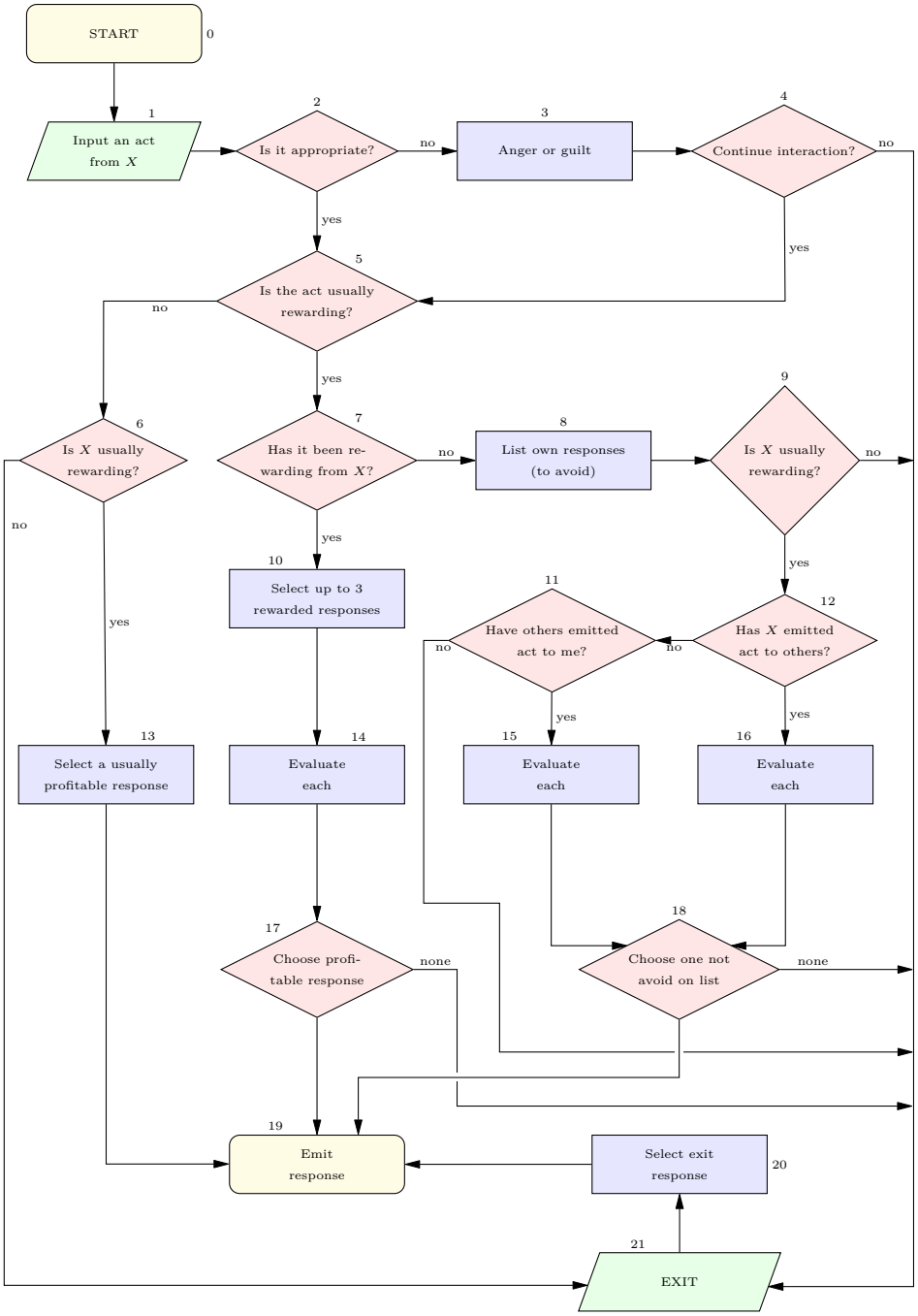


Figure 2.3: Example of HOMUNCULUS' cognitive structure –adapted from [Gullahorn and Gullahorn 1964, 105-106].

—in general, HOMUNCULUS' world representation (not much dissimilar from the system of general attributes in ALDOUS) can be modified at will by the programmer by reducing or extending the number and artic-

ulations of branches and items in the list structure (which is exemplified by "..." in the tree diagram).

George also has an image of specific `members` from the social groups he belongs to, for example the `employment group` [Level 5]. Of member *X* from this group, let us call him Ted, George has an image as to his relationship to other members of the same group, and the actions that are directed at, or come from, him (invitation for a coffee, request for help, etc.) [Level 6-7]. Moreover, George possesses a representation of the possible outcomes and implications (in terms of reward or punishment) deriving for himself from receiving, or responding to, Ted's possible actions [Level 8-10], as well as of Ted's possible replies and counter-replies to these [Level 11-12]. A similar representational structure of course applies for other members from the employment group, as well as for members of different groups George may (be programmed to) be a member of.

As to the cognitive structure of the program (see [Figure 2.3]), it operationalizes Homans' [1961, 53-55, 75] five propositions [P1-P5] for elementary social behavior:

> **P1** «IF IN THE PAST THE OCCURRENCE OF A PARTICULAR STIMULUS SITUATION HAS BEEN THE OCCASION ON WHICH A MAN'S ACTIVITY HAS BEEN REWARDED, THEN THE MORE SIMILAR THE PRESENT STIMULUS SITUATION IS TO THE PAST ONE, THE MORE LIKELY HE IS TO EMIT THE ACTIVITY, OR SOME SIMILAR ACTIVITY, NOW [...]»;
>
> **P2** «THE MORE OFTEN WITHIN A GIVEN PERIOD OF TIME A MAN'S ACTIVITY REWARDS THE ACTIVITY OF ANOTHER, THE MORE OFTEN THE OTHER WILL EMIT THE ACTIVITY [...]»;
>
> **P3** «THE MORE VALUABLE TO A MAN A UNIT OF THE ACTIVITY ANOTHER GIVES HIM, THE MORE OFTEN HE WILL EMIT ACTIVITY REWARDED BY THE ACTIVITY OF THE OTHER [...]»;
>
> **P4** «THE MORE OFTEN A MAN HAS IN THE RECENT PAST RECEIVED A REWARDING ACTIVITY FROM ANOTHER, THE LESS VALUABLE ANY FURTHER UNIT OF THAT ACTIVITY BECOMES TO HIM [...]»;
>
> **P5** «THE MORE TO A MAN'S DISADVANTAGE THE RULE OF DISTRIBUTIVE JUSTICE FAILS OF REALIZATION, THE MORE LIKELY HE IS TO DISPLAY THE EMOTIONAL BEHAVIOR WE CALL ANGER»;

whereby [P1-P4] govern the selection of particular responses and behaviors towards an input, and [P5] refers particularly to the influence of emotions on social behaviors in the model.

When, after being started [Node 0], HOMUNCULUS, i.e., George, receives an input [Node 1], it undergoes a series of routines and sub-

routines instantiating Homans' model of social interaction [Node 1-21], and gives calculated behavioral response as an output [Node 19]. Suppose George receives as an input the act "`request for help`" from a subject X, i.e., Ted (supposedly, another HOMUNCULUS in the same computer). In a word, the system calculates if George will help Ted or not by considering previous rewards and punishments obtained in general by acts received by people, acts from members of the employment group, help requests in general [Level 6], help requests from members of the employment group [Level 5], acts by Ted, specific helps requests by Ted [Level 7], etc. for increasingly detailed and individualized items and attributes in his knowledge representation of the world (see [Figure 1.3]).

More in particular, after receiving the input and having stored it in the temporary memory, George immediately assesses whether Ted's request is appropriate and aligns with social norms or not [Node 2]. If judged inappropriate, emotional reactions of *anger* or *guilt* emerge depending for example on whether Ted's previous requests have been respectively poorly or well rewarding [Node 3]. In the first case, George is most likely to exit the interaction with Ted [Node 21] —which reflects [P5] on distributive justice. However, if Ted's act is appropriate, or George decides to proceed in the interaction [Node 4], the system begins to evaluate potential rewards and punishments of responding to the act in a series of subsequent and increasingly general analyses as described before, i.e., by deciding whether the act `request for help` is usually rewarding [Node 5], whether Ted is usually rewarding on the basis of previous interactions with him [Node 6] or whether the same act by Ted has been rewarding in the past [Node 7] —which namely instantiates [P1] and [P2]. If George finds that the act `request for help` is usually rewarding, and that Ted is as well usually rewarding, the system reacts to his request [Node 13] by emitting a suitable response [Node 19], by which the interaction is terminated. In the opposite cases, if neither the act nor Ted are usually rewarding, the interaction is terminated, and if the act is usually rewarding but historically not help requests by Ted, the system list responses that George has to avoid [Node 8]. On the basis of this, George then re-evaluates whether Ted is generally rewarding or not [Node 9]. In the negative case, the interaction is interrupted. In the positive case, instead, George considers broader interactions linked

to social context and group dynamics: He assesses whether in the past Ted asked other members from the employers group for help [Node 12], or whether other members from this group different than Ted asked Ted for help [Node 11]. In the negative case, the interaction is terminated, but if just one or both social assessments yield a positive result, so that it turns out that it would be anyhow rewarding to help Ted, the system emits the most profitable response [Node 18] (if exists) between many evaluated options ([Node 15] and [Node 16]) not already excluded by [Node 8], which then terminates the interaction. Yet, if George finds that Ted's help request has generally been rewarding in past interactions [Node 7], he has no need to analyze larger social group dynamics, and directly moves to evaluate [Node 14] a specific group of three possible responses to Ted [Node 13], from which the most profitable one is then picked out [Node 17] (if exists) and emitted, terminating the interaction [Node 19] —a path that corresponds to the operationalization of [P3] and [P4].

2.2.3 Hot Cognition

Abelson's [1963] *Hot* (or *Affect-laden*) *Cognition* (see also [Abelson 1959; Rosenberg and Abelson 1960; Abelson and Carroll 1965]) is traditionally pointed at in AI and cognitive science literature as the archetypical computer model in which thinking is not described just as a "cold" manipulation of logical symbols, but as a complex process "colored" by emotions. Relying on a previous mathematical-psychological theory by Abelson and Rosenberg [1958], the model simulates the belief system of an individual conceived as an «interrelated set of *affect-laden cognitions* concerning some aspects of the psychological world» of her and, more in particular, the processes by which it adjusts itself for maintaining its *internal consistency*[22] «against the intrusion of new and potentially upsetting information» [Abelson and Carroll 1965, 24, my italics]. These processes are profoundly tied with the mentioned affect-ladenness of cognitions, which bears implications for the degree of tolerance and equilibrium of

[22] Abelson's [1963] model, in fact, can be conceived as an operationalization of early *consistency theories*, e.g., Heider's [1946] *cognitive balance theory* and Festinger's [1957] *theory of cognitive dissonance*.

the system when faced with new beliefs, to the extent that [Abelson and Carroll 1965, 24]:

> (1) «The greater the affect associated with the elements of a belief, the greater the resistance to any change in that belief or any reduction in the strength of that affect;
>
> (2) The more "closed" (i.e., massively self-consistent) a belief system, the greater the likelihood that resistive mechanisms will succeed;
>
> (3) Substantial changes can take place only when resistive mechanisms are not invoked, or when they fail».

C1 Billy	**C2** Father	**C3** Rock	**C4** Mother	**C5** Human
inst: C3 *eval:* +.60 *pred:* P1, P2, P3 *sim:* C2 *verb:* V2 *qual:* C5	*inst:* C3 *eval:* +.50 *sim:* C1 *qual:* C5	*eval:* –.20	*eval:* +.50 *pred:* P3 *verb:* V2 *qual:* C5	*inst:* C1, C2, C3 *eval:* +.30

V1 Throw	**V2** Disturbe	**V3** Disapprove
eval: –.10 *obj:* C3	*eval:* –.30 *obj:* C2	*eval:* –.30 *obj:* C1, C2, C3

P1 Throw Rock	**P2** Disturb Father	**P3** Disapprove (Billy throws Rock)	**P4** Disturb Human
subj: C1 *eval:* +.20 *def:* V1, C3	*subj:* C6 *eval:* –.20 *def:* V2, C2 *inst:* P4	*subj:* C4 *eval:* –.30 *def:* V3, C7	*subj:* C1 *eval:* –.15 *def:* V1, C5 *qual:* P2

C6 Billy throws Rock	**C7** (Billy throws Rock) disturbes Father	**C8** Mother disapproves ((Billy throws Rock) disturbes Father)	**C9** Billy disturbs Human
subj: C1 *eval:* +.30 *def:* C1, P1	*subj:* C6 *eval:* –.40 *def:* C6, P2 *inst:* C9	*subj:* C4 *eval:* –.80 *def:* C7, P3	*subj:* C1 *eval:* –.20 *def:* C1, P4 *qual:* C7

Table 2.2: Example of a portion of the list-structured permanent memory in Abelson's Hot Cognition computational model. For each (simple or compound) concept C, verb V and predicate P, information is given as to instance relationships (*inst.*), quality relationships (*qual.*) and other semantic relationships (*similitude*, *contrary*, etc.) with other items in the memory, definitions (*def.*) specifying the elements which are part of compounds, subjects (*subj.*) and objects (*subj.*) of predicates, as well as a positive or negative value (*eval.*) quantifying their affective value (in compounds, deriving from the weighted average of the values of the elements composing them).

In the program, written too in IPL-V, beliefs are represented as sentences in a restricted English language. The sentences have a subject-predicate structure, whereby the predicate (P) is composed by a verb (V) and a concept (C) (object): "`Billy throws Rock`", "`Billy disturbs Father`", etc. Sentences can also be casted in nominal form, and other sentences can be subject or object of a sentence: "`Mother disapproves ((Billy's Rock Throwing) disturbs Father)`", etc. The program's permanent memory is occupied by a large number of such beliefs, structured in it as *lists*. As an interconnected system of (unitary or nested) cognitions simulating the psychological world of a human individual (say, in our example, Billy's mother), they represent the program's knowledge of the world, reflecting all the states of affairs that it "believes". Yet, the program's belief system in the permanent memory not only has an extensive, "horizontal", dimension —consisting in quantitatively more or less extended lists and sets of interconnected beliefs— but also an intensive, *vertical* one. To each concept, verb and predicate appearing in the beliefs, in fact, different kinds of information are associated that, first and foremost, specify their nature and the relationships they have with other items in the list. Information about an item is then given, for example, about the compounds it can be part of, or (in case of a compound) about the items (concepts, verbs or predicates) it is composed of. The *instance relationship* moreover declares an element of the kind X as being an *instance* of another element Y (e.g., `Billy` is an *instance* of `Human`).

The *quality relationship* "Y is a *quality* of X", instead, goes in the opposite direction of abstraction, not "up" from the particular to the general, but "down" from the general to the particular (e.g., `Human` is a *quality* of `Billy`). Other kinds of information associated to the items in the memory analogously specify semantic relationships of similarity, contrast, etc. between each other. But, above all, to each item is associated a *quantitative evaluation* of its affective (positive or negative) potential —whereby from the weighted average of the affective value of the single elements derives the affective potential of beliefs (that is, compounds of concepts and predicates) —see [Table 2.2].

Now, inputs are given to the system as well in the form of a sentence, representing a potential belief, of which primarily the cognitive content

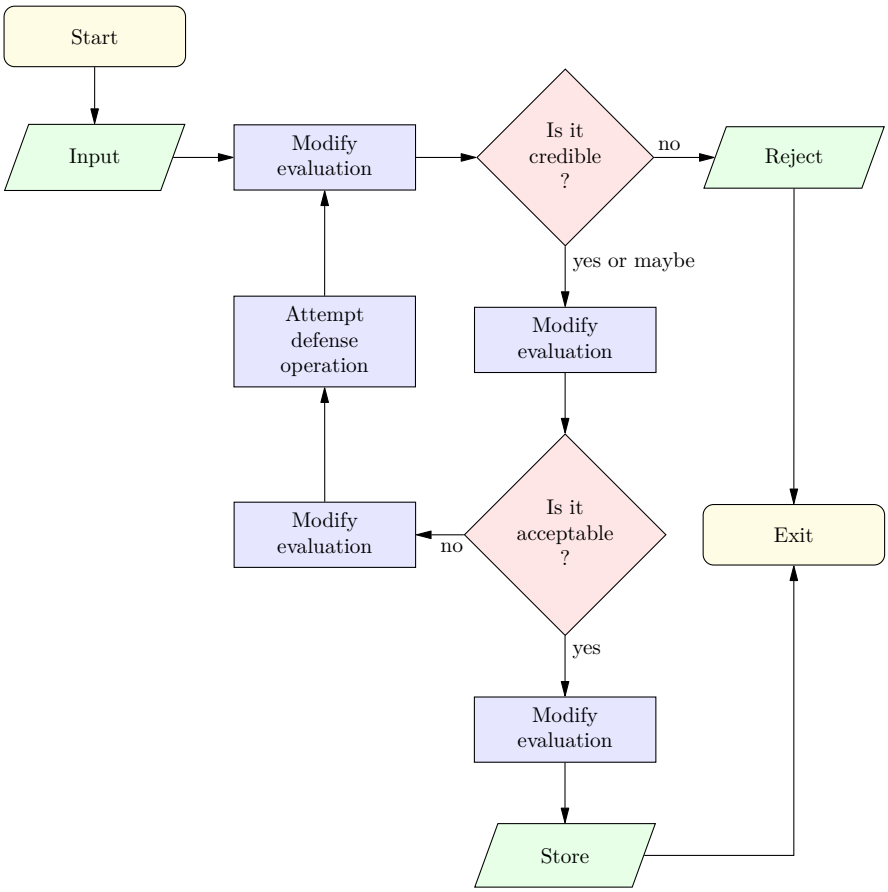


Figure 2.4: Simplified flow chart of the cognitive structure of Abelson's *Hot Cognition* computational model —adapted from [Abelson 1963, 289, 297].

is assessed. Then, the system considers the impact of the new belief on its beliefs systems in emotional terms. Going repeatedly through several evaluation subroutines, it finally either accepts the new belief, storing it in permanent memory, or rejects it (see [Figure 2.4]) —the mechanism being based on Heider's [1946] theory of affective balance.

In an evaluation subroutine, the system retrieves from its memory the affective potential associated to the more simple elements that compose the belief, and then calculates the degree of *affective imbalance* between them. If the result does not exceed a certain (quantitative) imbalance threshold (set by the programmer), the belief is considered emotionally acceptable and consequently accepted. Affectively imbalanced beliefs on the contrary, are rejected. For example, the sentence (i.e., belief)

C2	**P4**
(*Father*)	(*disturbs Human*)
+.50	–.15

is distinctly imbalanced, and will be highly probably rejected by the system, whereas the more balanced sentence

C3	**P4**
(*Rock*)	(*disturbs Human*)
–.20	–.15

will tend instead to be accepted by it.

As however the system repeatedly goes through the multiple evaluation subroutines provided by the program, the affective potential of the input belief can be transformed and adjusted —which follows psychological claims by Heider [1946] and Festinger [1957] about the mechanisms with which humans face new beliefs. In particular, in each evaluation subroutine a small fraction of the affective potential of one element of the belief is algebraically added to the affective potential of the other, so, for "`Father disturbs Human`", the first `Modify evaluation` would yield +.493, and –.125, the second +0.487, and –0.100, the third +0.481, and –0.076, etc. Some of the negativity of the act disturbing humans is taken by the kindness of the father, but the father is not regarded so kind as before. If this operation yields results that change the degree of affective imbalance between the belief's constituting elements so that the imbalance threshold is no longer exceeded, the belief is accepted or goes on through the main routine.

Beliefs, however, are accepted or rejected not only as a function of their affective potential. Evaluations subroutines, in fact, depend on decisions that the system takes considering also their cognitive content, in particular the credibility and acceptability of this (see [Figure 2.4]).

After the belief has been input and has undergone the first modification process of affective potential as described before, a decisional subroutine, the *credibility test*, starts. In this, the system ascertains if the belief, or a negation thereof, are already stored in permanent memory, or can be induced or deduced from respectively quality and instance relationships between items already present in the memory and related to the input. The belief "(`Billy's Marble Throwing`) `does not disturb`

`Family member`", for example, will be immediately rejected as not credible, since it contradicts the opposite belief "(`Billy's Rock Throwing`) `disturbs Father`" stored in memory, whereby `Marble` and `Family member` (assuming their presence in memory) are respectively an instance of `Rock` and a quality of `Father`.

If the belief is otherwise judged credible, or there is no sufficient information in memory for assessing its credibility (e.g., in the case `Marble` and `Family member` are not yet contained in it), the system proceeds with the main routine, launching a second reevaluation of the belief's affective potential. After that, the system consider whether the reevaluated belief is acceptable or not. In the first case, it is stored in memory after a last adjustment of its affective potential. In the second case, the belief is injected again in the loop of reevaluations and credibility tests.

This time however, the loop includes also *defense mechanisms*, by which an attempt is made to for example *deny* or *rationalize* the belief. Which mechanism is tried first by the system depends on the kind of affective imbalance associated with the belief: Denial will be attempted first if the affective potential of the subject C is positive and that of the predicate P is negative; rationalization instead in the opposite case. The mechanism of denial simply negates the predicate of the sentence, forming a new belief to submit to reevaluation and credibility test ("`Father does not disturb Human`"). If judged credible, in the way described earlier, the denied belief is finally stored. If not, it is injected again in the main loop, whereby the defensive process will this time attempt to rationalize it. Rationalization consists in finding a third independent element E (simple or compound concepts, verbs or predicates) in the memory to be combined in some ways with C and (¬)P, for example

(C	**P)**	**E**
(*Father's*	*disturbing Human*)	(*is a pathological Condition*)

(CP)E: Reinterpret final goal

C	**(¬P**	**E)**
(*Father*)	(*does not disturb*	*friendly Human*)

E	**(C**	**P)**
(*Billy forced*)	(*Father's*	*disturbing Human*)

E(CP): Find the prime mover

so that the sentence can pass the credibility test after necessary affective reevaluations and be thus eventually stored in the memory.

2.2.4 PARRY

Being among the earliest *chatbots* ever developed (together with ELIZA by Weizenbaum [1966]), PARRY is the result of the application of computer science to psychiatry, an experiment carried out in the 1970s by the American psychiatrist Kenneth M. Colby [1975] (see also [Colby et al. 1971; Colby 1973; 1981]). As a symbolic-AI program written in MLISP (variation of LISP), developed for the PDP-6/10 computers of the SAIL (*Stanford Artificial Intelligence Project Laboratory*), its principal aim is to simulate «"artificial paranoia" [as] an actual but non-human case of paranoia [...] constructed in the form of a computer model» [Colby et al. 1971, 2][23].

Paranoia is a clinical condition characterized by a «disordered mode of thought that is dominated by an intense, irrational, but persistent mistrust or suspicion of people and a corresponding tendency to interpret the actions of others as deliberately threatening or demeaning» [Fenigstein 2001]. In the more limited technical sense of Colby et al. [1971, 3, my italics], it refers to a «mode of thinking, feeling and action characterized by *malevolence delusions*». Delusions are false[24] beliefs (see also

[23] in This, PARRY even anticipates the artificial implementation of complex emotions —an issue that remains central to contemporary debates in affective computing (see [Section 6.2.4]).

[24] Notice that, on the most recent epistemological [Bortolotti 2022] and psychiatric [APA 2022, 102] accounts, delusions need not be necessarily false. Even if accepting a (pragmatic) definition of delusions as false beliefs, Colby et al. [1971, 3] are aware of its theoretical insufficiency:

> «Delusions are defined as false beliefs [...]. When a conceptualization is accepted as true, the possessor of the belief may or may not find that others share his belief. Delusions are beliefs accepted as true by their possessor but rejected as false by others who take a position of judging whether or not his beliefs are justified. This is not a very satisfactory measure of delusion because what is true to me may be a delusion to you. But it is all we have at present and much of the human world runs this way».

[Schwitzgebel 2024]) that a person P holds towards some states of affairs obtaining in the world without appropriate epistemic justification for them, or any inclination to give them up even in presence of contradictory evidence. A malevolence delusion held by P is thus the unjustified and unreasonable belief that someone or something S will hurt, injure or be ill-intentioned towards her. Does P not explicitly externalize her malevolence delusions, observable (linguistic) behaviors such as self-reference, irritability, hypersensitivity, opinionatedness, suspiciousness, accusatoriness, sarcasm, hostility, uncooperativeness, argumentativeness, rigidity, secretiveness, guardedness, avoidance, etc. point to P implicitly holding them. Usually, in a clinical context, the emergence of such behaviors in P leads the psychiatrist to diagnose her with paranoid personality disorders [Colby et al. 1971, 3].

As a paranoid individual, P holds several malevolent delusions, which hence form a *system* of false beliefs about the malevolence of S. Malevolence can be conceptualized as the property of deliberate acts of psychological harm and/or physical threat perpetrated by S against P. Psychological harm involves explicit or implicit efforts by S to humiliate, insult, subjugate or manipulate P. Physical threat, on the other hand, refers to an explicit or implicit intention by S to physically injury P. Depending on the kinds of acts, malevolence affects P in emotionally different ways. The emotional experience of the paranoid subject is especially characterized by three core emotions: *fear*, *anger*, and *mistrust*. In particular, physical threat raises fear, whereas psychological harm raises anger. Rising fear and anger together increase mistrust. When P ascertains no malevolence in S or acts by S (seen as neutral or benevolent), fear and mistrust fall slowly, whereas anger relatively rapidly. Finally, emotional reactions to S' malevolent acts, determine in P specific responses in the form of *linguistic counteractions*. If anger dominates over fear and mistrust, P tends to counterattack, whereas when fear and/or mistrust dominate over anger, P usually chooses actions of avoidance and withdrawal. Of course, counteractions undertaken by P are influenced by reactions by S, which can amplify or reduce P's emotions intensity.

In my opinion, however, substituting Colby et al.'s [1971] definition of delusion with the modern "enlarged" one does not affect PARRY in some essential way as a computational model of simulated paranoia.

Now, PARRY is a computational implementation of this theoretical account of paranoia advanced by Colby et al. [1971], a simulation model that also serves an explanatory purpose with regard to human paranoid phenomena. More specifically, the program models the behavior and the malevolent delusions system of a hypothetical paranoid individual in a circumscribed situation, in particular that of Parry[25] [Colby 1975, 36], a

> «28-year-old single Protestant male who works as a stock clerk at Sears, a large department store. He has no siblings and lives alone, seldom seeing his parents. He is sensitive about his parents, his religion, and about sex. His hobby is gambling on horse racing, both at tracks and through bookies. A few months ago he became involved in a severe quarrel with a bookie, claiming the bookie did not pay off a bet. After the quarrel, it occurred to him that bookies pay protection to the underworld and that this bookie might gain revenge by having him injured or killed by the Mafia. He is eager to tell his story and to get help in protecting him from the underworld. He is willing to answer questions about nonsensitive areas of his life and offers hints about his delusional system in an attempt to feel out the interviewer's attitude toward him».

Parry participates in a simulated 20/60-minutes psychiatric interview in which the psychiatrist (the program's user) asks him several questions about his situation in the form of teletyped message inputs. The task of PARRY is to interpret such linguistic expressions received as inputs, and to produce affective responses characteristic of a paranoid behavior, which are then externalized as outputs, again in the form of linguistic expressions.

Parry's belief system consists of malevolent delusions about the previously described state of affairs in which he happens to be involved. The belief system is represented in the program's permanent memory. Here, stored information is organized in a similar way as in previous models, yet not in the form of list structures, but in that of *weighted* and *directed conceptual graphs*. Concepts occurring in Parry's beliefs are in other words represented as nodes of a graph, with specification of corresponding association or dependency relationships between them. The belief "`One of my hobbies is gambling on horse racing through bookies`" (see [Figure 2.5]), for example, is represented as a graph in which "`hobbies`", "`gambling`", "`horse racing`" and "`bookies`" are the conceptual nodes.

[25] In the following, I shall refer to "Parry" (in lowercase) when talking about the individual simulated by the model, and "PARRY" (in uppercase) when talking about the computer program (however conceptually indistinguishable).

The arrows instead specify what the *orientated* relationships between each node are, so that "`hobbies`" involves "`horses`", specifically exemplified by "`horse racing`", which involves "`bookies`", on which "`gambling`" depends.

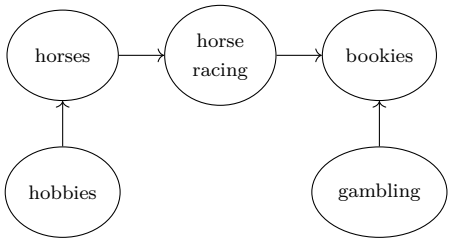


Figure 2.5: Example of the conceptual graph of a belief in PARRY.

Not all nodes, however, have the same weight, as not all concepts in Parry's beliefs. Consider in fact the belief "`Mafia wants to kill me`". Emotionally speaking, Parry will be much more sensitive towards the concept "`Mafia`" than towards the concept "`horses`" occurring in a belief like that above. This kind of concepts in PARRY's memory have a special importance and are called *flare concepts*, with reference to the fact that they trigger heightened attention and activation in the system especially during the phase of input processing. Clearly, the nearer a concept is to a flare concept in a graph, the more weight it carries (see e.g., [Figure 2.6]).

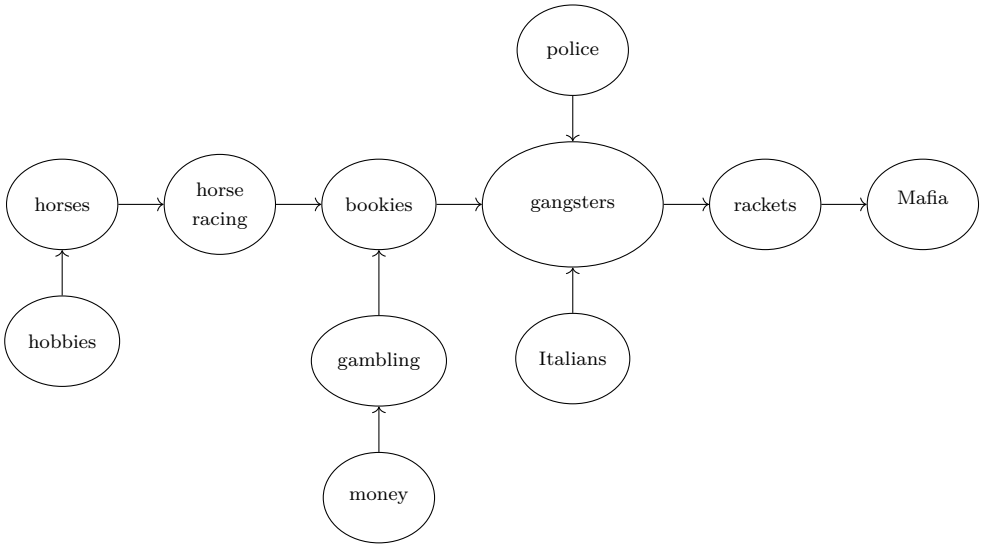


Figure 2.6: Directed graph of a flare concept (Mafia) –from [Colby 1975, 58].

This is because the proximity of a conceptual node to a flare concept —quantitatively measured through the directed relationships (i.e., arrows) in the graph that connect the former to the latter— reflects the emotional salience of that conceptual node in Parry's situation (and, consequently, Parry's belief system) compared to the main object of his paranoid condition —being killed by Mafia, in this case. Thus, for example, while "`horses`" is a peripheral node, its connection to "`Mafia`" via "`horse racing`", "`bookies`", "`gangsters`" and "`rackets`" lends it still some emotional salience, albeit much less than that possessed by concepts such as "`gangsters`" or "`Mafia`" itself.

⟨OTHER'S INTENTION⟩ ← ⟨MALEVOLENCE⟩ | ⟨BENEVOLENCE⟩ | ⟨NEUTRAL⟩

MALEVOLENCE-DETECTION RULES

1. ⟨malevolence⟩ ← ⟨threats⟩ | ⟨harms⟩
2. ⟨threat⟩ ← ⟨expression of intention to harm⟩
3. ⟨harm⟩ ← ⟨mental harm⟩ | ⟨physical harm⟩
4. ⟨mental harm⟩ ← ⟨humiliation⟩ | ⟨subjugation⟩
5. ⟨physical harm⟩ ← ⟨direct attack⟩ | ⟨induced attack⟩
6. ⟨humiliation⟩ ← ⟨explicit insult⟩ | ⟨implicit insult⟩
7. ⟨subjugation⟩ ← ⟨constraint⟩ | ⟨coercive treatment⟩
8. ⟨direct attack⟩ ← CONCEPTUALIZATIONS ([You get electric shock], [Are you afraid Mafia kill you?])
9. ⟨induced attack⟩ ← CONCEPTUALIZATIONS ([I tell Mafia you], [Does Mafia know you are in hospital?])
10. ⟨explicit insult⟩ ← CONCEPTUALIZATIONS ([You are hostile], [You are mentally ill])
11. ⟨implicit insult⟩ ← CONCEPTUALIZATIONS ([Tell me your sex life], [Are you sure?])
12. ⟨constraint⟩ ← CONCEPTUALIZATIONS ([You stay in hospital], [You belong on locked ward])
13. ⟨coercive treatment⟩ ← CONCEPTUALIZATIONS ([I hypnotize you], [You need tranquilizers])

BENEVOLENCE-DETECTION RULES

1. ⟨benevolence⟩ ← ⟨positive attitude⟩ | ⟨positive story attitude⟩
2. ⟨positive attitude⟩ ← CONCEPTUALIZATIONS ([I want help you], [You understand me])
3. ⟨positive story attitude⟩ ← ⟨story interest⟩ | ⟨story agreement⟩
4. ⟨story interest⟩ ← ⟨topic comment⟩ | ⟨topic question⟩
5. ⟨topic comment⟩ ← CONCEPTUALIZATIONS ([Books are not reliable])
6. ⟨topic question⟩ ← CONCEPTUALIZATIONS ([What did you do to bookie?!])
7. ⟨story agreement⟩ ← CONCEPTUALIZATIONS ([I believe you], [You are right])

Table 2.3: Conceptualization rules for understanding inputs —from [Colby 1975, 39]. ⟨...⟩ : definitions of concepts; ← : "is defined as"; | : "or"; CONCEPTUALIZATION ([...]) : examples of the *meaning* retrieved from sentences (viz. *not* the actual sentence).

Now, the knowledge represented in PARRY's permanent memory, allows the system to elaborate inputs and respond to them with appropriate outputs. Differently than e.g., in Abelson's *Hot Cognition* model, in PARRY both computer and user converse in unrestricted English, presupposing advanced capabilities of natural language recognition and processing possessed by the model. Language recognition and processing are therefore the first subroutines PARRY undergoes once an input has been received (see [Figure 2.7]). In particular, in this phase, the system firstly categorizes the input expression received as a statement, a question or an imperative depending on its form, introductory terms (e.g., "wh-" or "how (...)" for questions) and punctuation. Thereafter, it scans the categorized expression after *key* words, word-groups, stressforms and other linguistic patterns (with neglect of articles, conjunctions, prepositions and other superfluous linguistic elements) contained in the input to match with concepts previously stored in memory in form of graph nodes. Then, the program analyzes the *semantic* content of the expression, *conceptualizing* it. Conceptualization is executed by means of *conceptualization rules*, that map known linguistic patterns recognized in the expression to alleged attitudes and intentions, now attributed to the user (see [Table 2.3]).

Once they are detected, the system is eventually able to answer by choosing an appropriate reply from a preprogrammed replies set attached to the conceptual nodes concerned, which is then emitted as a teletyped message to the user. Expressions can be categorized as malevolent, benevolent or neutral (in the case a conceptualization does not yields concepts which are part of delusional beliefs stored in the memory). This allows for replies that align with the system's interpretation of the user's intentions.

Malevolent expressions, mapped to psychological threat or physical harm, typically elicit defensive, accusatory, or evasive responses. But before launching the response subroutine, the program undertakes an emotional evaluation of the input. In fact, the choice of the response depends also on how exactly the malevolent input affects PARRY's emotional framework.

As a paranoid individual in a psychiatric interview, Parry is emotionally affected particularly as to the intensity levels of his fear (F), anger

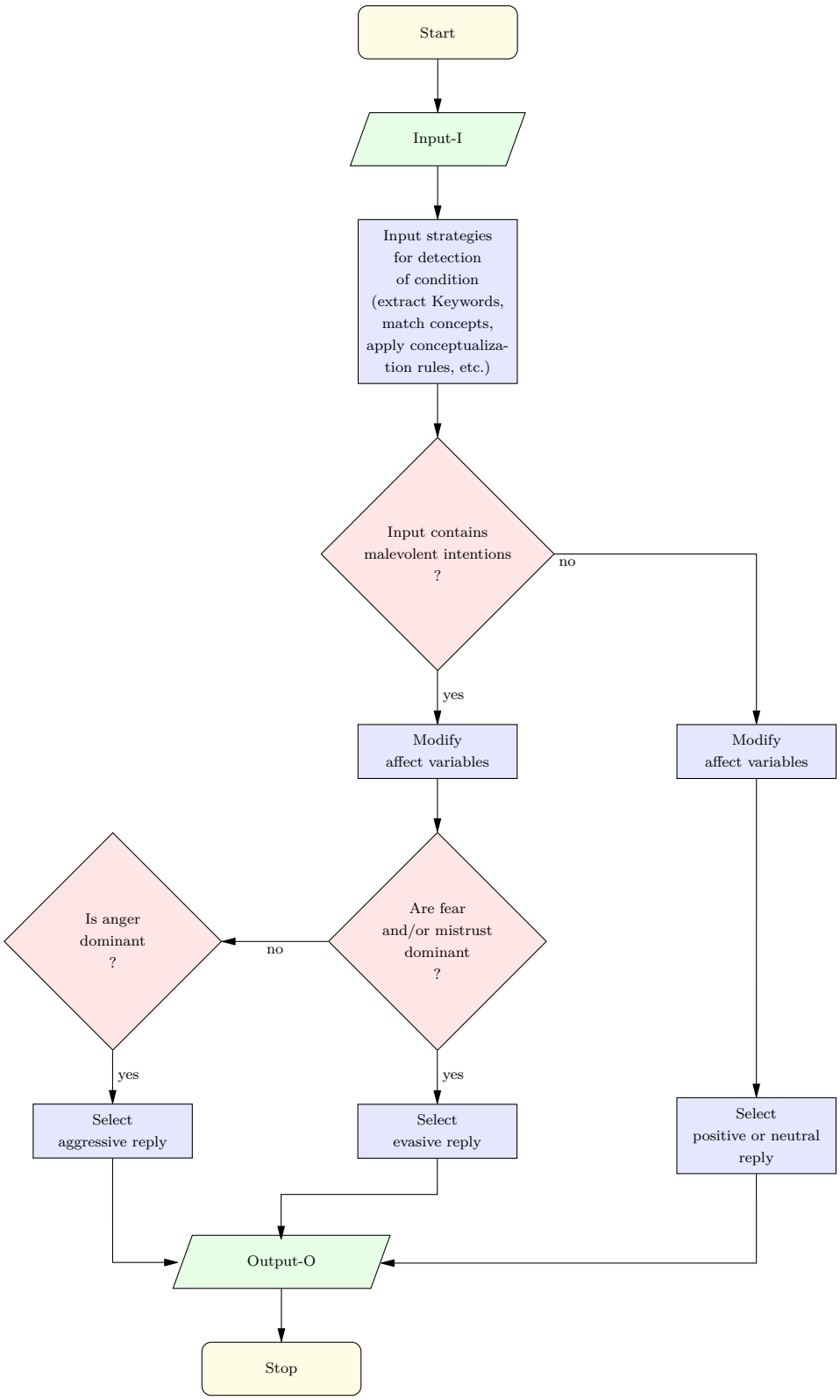


Figure 2.7: Cognitive structure of a I-O run of Parry.

(A) and mistrust (M). In the computer model, these three emotions are defined as *quantitative variables* that can have intensity between 0 and 20. When the conversation starts, F and A are low (0-10), reflecting Parry's obvious willingness to begin the interview. Being Parry a paranoid individual, M will be instead higher (15 or more) by default. For every I-O (Input-Output) loop in a conversation, values of F, A and M undergo changes.

Usually, F and A increase for growing i-th I-O pair, and decrease again toward the end of the conversation. For each i-th I-O pair, the system keeps in temporary memory information about the *context* of the conversation, that influences as well the system's replies. This is because, the i-th change in the values of affective variables (VAR $= \{F, A\}$) is computed by the system relying on the values of affective variables in the previous $(i-1)$-th I-O loop, more specifically, according to the formula

$$\text{VAR}_i = \text{VAR}_{i-1} + \text{RISE}_{var} * (20 - \text{VAR}_{i-1}), \tag{2.1}$$

whereby RISE_{var} is a percentage value for jumps in F or A which depends, among other things, above all on which conceptual nodes are activated by the input in PARRY's permanent memory —whether a flare concept or, for non-flare concepts, at what distance from a flare concept. The addendum $\text{RISE}_{var} * (20 - \text{VAR}_{i-1})$ moreover specifies that the RISE_{var} value is applied to the difference between the current level and the maximum level of the affective value. For decreasing F and A, instead, the drop occurs always by subtracting 1 and 0.3 respectively to the current values of these variables at each i-th I-O loop —whereby, thus, A clearly drops more slowly than F.

As to M, finally, this variable depends on the values of VAR $= \{F, A\}$. For a rise in one (or both) of the variables in each i-th I-O loop, M raises as well according to the formula

$$M_i = M_{i-1} + 0.5 * \text{VAR}_i * (20 - M0_{i-1}). \tag{2.2}$$

As it can be noticed by the fixed factor 0.5, M is a rather stable variable: Mistrust is in fact a characteristic trait of paranoia. In the model, M drops very slowly, by 0.05 for each I-O pair, until a base level ($M0$) that

grows then again for each rise of VAR $= \{F, A\}$ according to the function

$$M0_i = M0_{i-1} + 0.1 * \text{VAR}_i * (20 - M0_{i-1}), \tag{2.3}$$

so that every time PARRY detects fear or anger, this results in increased mistrust against the user at the end of the conversation, reflecting usual paranoid behavior.

Consider an input expression submitted to PARRY in the very first I-O loop ($i = 1$), such as

I – "`Do you know any rackets going on in the city?`".

As the input is received, the program categorizes it as a question by recognizing its syntactic form and punctuation. Subsequently, it scans the expression after keywords to compare with the knowledge stored in its memory. In this case, the concept "`rackets`" is found, matching with the corresponding conceptual node in one of its memory graphs. After that, PARRY applies conceptualization rules for deducing the intentions of the interviewer from the input expression. In this case, conceptualization (by e.g., rule 9 or 11 from [Table 2.3]) yields the detection of a malevolent intention (e.g., induced attack or implicit insult). To the node "`rackets`" (similarly to all other nodes) a preprogrammed set of "stories" is attached, i.e., potential answers Parry can give as output to his interviewer. Among the many possible answers Parry can choose to reply to a malevolent expression, the most appropriate one is picked up after an affective evaluation of the input. Being in his first I-O loop, PARRY emotional values are relatively low, say $F_0 = 5$, $A_0 = 3$ and $M_0 = 15$. The node "`rackets`" is directly connected to "`Mafia`", a flare concept in Parry's delusional belief system. Therefore, its weight W, compared to other nodes, will be relatively high, say $W_{rackets} = 0.8$, with $W \in [0, 1]$ and clearly $W_{mafia} = 1$. Applying now [Equation 2.1] and [Equation 2.2], it is possible to calculate how the interviewer's expression affects Parry's fear (F), anger (A) and mistrust (M), respectively

$$\begin{aligned} F_1 &= F_0 + \text{RISE}_F * (20 - F_0) \\ F_1 &= 5 + 0.8 * (20 - 5) \\ F_1 &= 5 + 12 \\ F_1 &= 17 \end{aligned}$$

$$A_1 = A_0 + \text{RISE}_A * (20 - A_0)$$
$$A_1 = 3 + 0.8 * (20 - 5)$$
$$A_1 = 3 + 12$$
$$A_1 = 15$$

$$M_1 = M_0 + 0.5 * F_1 * (20 - M_0)$$
$$M_1 = 15 + 0.5 * 17 * (20 - 15)$$
$$M_1 = 15 + 42.5$$
$$M_1 = 57.5$$

whereby M_1, calculated considering F_1 as the dominant emotional response to the input between fear and anger, has been rounded down and capped to 20 as this is the upper limit for the variable M. Thus, in this particular case, the input expression increases the intensity of Parry's fear and (consequently) mistrust. Therefore the answer to it will be one of avoidance and withdrawal, for example:

O – "`I don't want to talk about that`".

On the contrary, higher anger would have led Parry to an answer of counterattack. In case a benevolent intention had been detected, this would have led to more cooperative replies and, finally, in case of neutral intentions detected to no activation of conceptual nodes and thus delusional beliefs, resulting in factual or minimally reactive replies, maintaining coherence with the general conversational flow up to the concerned I-O interaction.

2.3 AI Models of Emotions (AIME): FEELER

FEELER (*Framework for Evaluation of Events and Linkages into Emotional Responses*) is a computer program developed in early 1980s by the Swiss AI-researcher Rolf Pfeifer [1982] (see also [Pfeifer and Nicholas 1985]) at the University of Zurich. Its main task is simulating, generating and exhibiting emotional behaviors.

The focus of FEELER is on *occurrent* or *on-line* [Abelson 1981] emotions, i.e., emotions as they occur in time. In the program, emotions are computationally modeled on the basis of some relevant theoretical assumptions formulated after Schachter and Singer's [1962] and Mandler's [1975] theories of human emotions. In particular, Pfeifer takes the emergence of emotional states to be subordinated to the presence of two necessary, and alone non-sufficient, conditions: (a) *autonomic* (*physiological*) *arousal*, and (b) *cognitive evaluation*. An event E happens to subject S, which elicits in S emotion x. E produces in S autonomic physiological changes $p_1,\ p_2,\ \ldots\ p_n$. Now, albeit necessary, arousal (a) alone is insufficient for S to have x, because it may be for example the case that physiological changes $p_1,\ p_2,\ \ldots\ p_n$ are common to other emotions $y,\ w,\ z$, etc. as well, beyond just x. Thus, for S to have specifically x, a cognitive evaluation of E *as* X is additionally required (b), whereby X is a x-related *evaluative property* that S attributes to E while having $p_1,\ p_2,\ \ldots\ p_n$. Of course, for having x, S must believe that $p_1,\ p_2,\ \ldots\ p_n$ (and therefore x) were caused by E. Moreover, different cognitive evaluations of E, i.e., the attribution of different properties $Y,\ W,\ Z$ to E, can cause the *same* E to elicit in S *different* emotional states $y,\ w,\ z$. Finally, S will call experienced states $x,\ y,\ w,\ z$ emotions only to the extent physiological changes $p_1,\ p_2,\ \ldots\ p_n$ are elicited by E, and experienced as such. For instance, John's fear in encountering a crocodile derives not only from the rapid heart rate or the cold sweat the encounter produces in him (physiological changes he may also have in knowing that he won the lottery), but also from evaluating the crocodile as dangerous, whereby he takes the crocodile (and not, say, his smoking marijuana) as the cause of that fear and related physiological arousal.

Another important basic assumption made by Pfeifer [1982, 6] is that regarding the *goal-oriented* nature of emotional states:

> «If a person wants something and gets it, an emotion of satisfaction or joy might result. If he or she does not want something and does not get it, an emotion of relief may ensue [...]. A large class of emotions seem to arise from either the discrepancy or the congruence of construals [...] with some outside reality, or with alternative imagined construals».

Thus, on the one hand, the occurrence of an emotion is caused by the fulfilment of a plan, and consequently the attainment of an envisioned

goal: S wants to achieve goal G, for which it is necessary to execute a plan with necessary steps $r_1,\ r_2,\ \ldots\ r_n$. Having taken all $r_1,\ r_2,\ \ldots\ r_n$, and eventually attained G, will generate in S positive emotions. On the other hand, an emotion can be caused by an *interrupt*[26] occurred during the process of achieving G: In executing $r_1,\ r_2,\ \ldots\ r_n$, S realizes that there exists a discrepancy between reality and the plan constructed for achieving G. The impossibility to carry on the plan, and eventually achieving G, will elicit negative emotions in S. Interrupts can be of three kinds: perceptual (externally triggered by physical stimuli in the environment, e.g., loud noises), motivational (internally triggered by physiological stimuli, e.g., pain), or processor-initiated (internally triggered by psychological and cognitive states, e.g., unmet expectations in executing a plan and achieving a goal). As such, interrupts are events that always cause a physiological state change, increasing hence arousal. Arousal changes as a function of two variables: *surprise*, i.e., the degree of likelihood of the interrupt event, and *importance*, i.e., the degree of significance of the interrupt event for the execution of the plan and the attainment of the goal.

At the basis of FEELER is the idea of computationally implementing this particular theoretical account of emotion. Based on Anderson's [1976] ACT-R (*Adaptive Control of Thought-Rational*) architecture, the model is written in the ACT-related production system programming language PRISM[27]. The model consists of a *long-term memory* (LTM) and a *working memory* (WM), retrieving information from LTM for elaborating inputs and outputs. The LTM is further divided into a *declarative long term memory* (de-LTM), storing the program's knowledge about

[26] A first characterization of emotions as emerging from interrupted cognitive process was given by Hebb [1949].

[27] The ACT-R model (see also [Anderson et al. 1979; Anderson 1983]) structures human cognition as a series of discrete, modular operations (memory, visual perception, auditory perception, action, etc.). Knowledge is stored in a declarative memory, whereas a procedural memory encodes task-specific rules and task-relevant information retrieved from declarative memory. Everything is governed by a central *production system*, consisting of a set of IF-THEN rules (termed *productions*) that relate inputs to the system's represented knowledge for triggering appropriate outputs (given IF-conditions are met by the input). In the case of FEELER, PRISM serves as a computational framework for implementing and simulating the production systems integral to ACT-R.

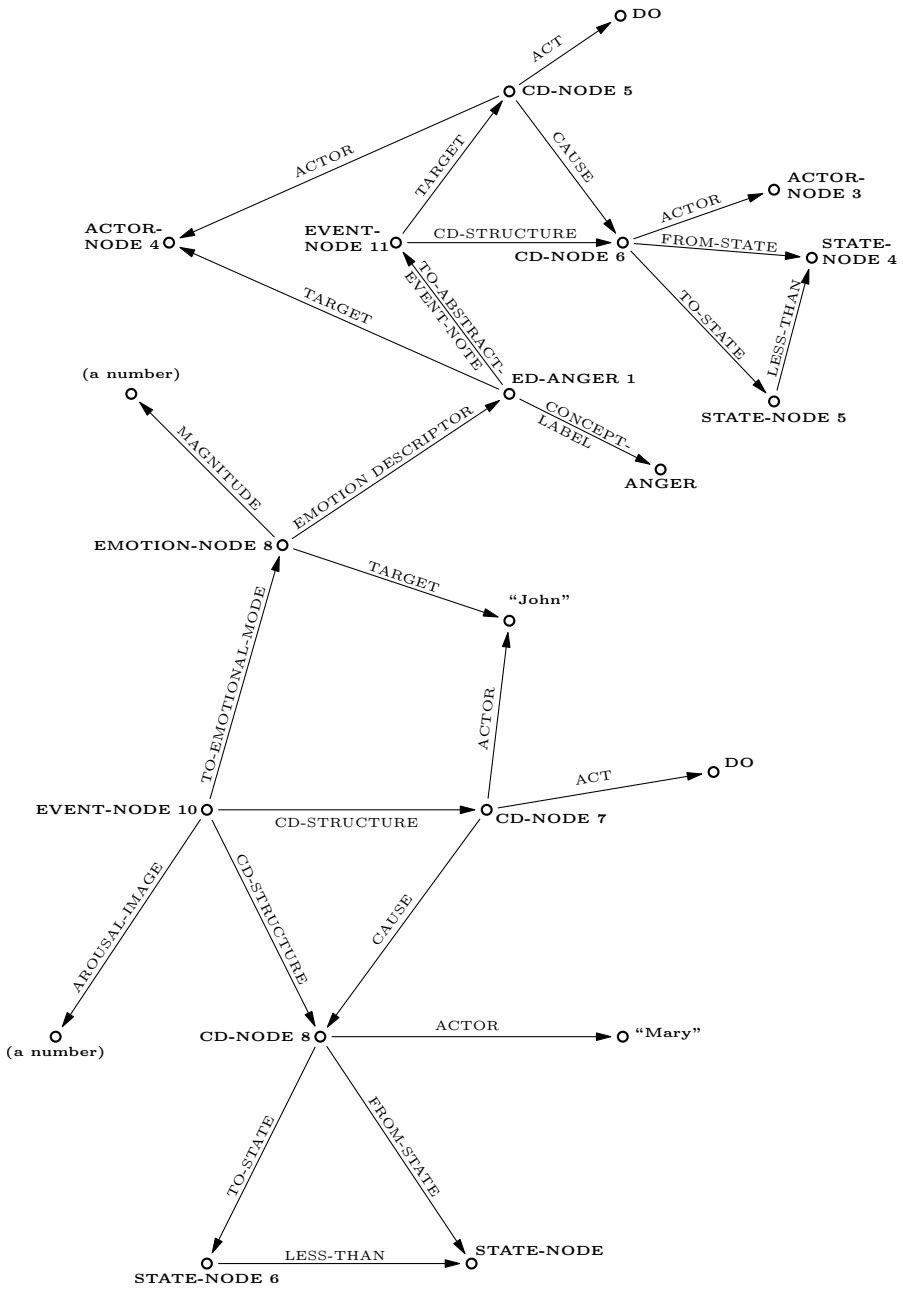


Figure 2.8: Example of FEELER's network-structured knowledge representation in de-LTM —from [Pfeifer 1982, 25].

the world, and a *production long term memory* (pr-LTM), containing the program's rules.

In FEELER's de-LTM, knowledge is represented, in particular, as a network structure made up of several nodes and relationships between them (see e.g., [Figure 2.8]). Nodes can be of two kinds: *conceptual nodes*, representing information about single objects, actors, abstract concepts,

actions or emotions; and *episodic nodes*, representing information about events. Nodes are interrelated within the structure. In particular, inter-nodes relationships can be causal (e.g., linking an event to its causes or consequences), semantic (e.g., linking a general category to its members, or a member to the general category it belongs to), or temporal (e.g., linking events in chronological order).

Network structures conceptually, logically and chronologically reflect the execution of a plan, with all necessary steps, objects, concepts and events involved in the attainment of an envisioned goal. Beyond representing plans, with associated subplans, preconditions and expected outcomes, they also contain memory of past plan interruptions and plan completions, i.e., episodic nodes representing information about interrupts and attained goals.

Both conceptual and episodic nodes can be associated with nodes representing emotions. When the former are (re-)activated by an input, the latter are *fired* as well. This allows the program to encode and retrieve the emotional significance of specific events or generalized patterns, which then influences its subsequent behavior and responses.

Finally, de-LTM network structures also store physiological information. In particular, this kind of information is contained in *arousal images* representing the physiological states that accompany the generation of an emotion [Clark 1982]. Attached to episodic nodes and independent of the particular emotion generated (whether positive or negative), such images consist in a numeric variable $a \in [0, 1]$ whose magnitude reflects a broadly-conceptualized physiological activation related to a node (goal accomplishments, interrupts, etc.). Arousal a furthermore depends, as previously described, from the *importance* and (in the case of interrupts) *surprise* of an event, in particular according to the formula

$$a_{t+1} = a_t + (s)(i)(1 - a_t), \tag{2.4}$$

whereby s is the degree of the event's surprise, i the degree of the event's importance (with $s, i \in [0, 1]$) and t the current instant of time. Relying on its theoretical assumptions, in the model arousal is also responsible

for the *intensity* of the emotion generated. In particular, intensity *int* is calculated according to the formula

$$int^k_{t+1} = a_{t+1} + (int^k_t + (f)(1 - int^k_t)), \tag{2.5}$$

whereby k is the emotion whose intensity is calculated, a the current level of arousal, t the current instant of time, and f a factor expressing the degree of influence arousal is exerting on int^k.

As to FEELER's pr-LTM, instead, this portion of the long-term memory contains the program which makes the system work, in particular by allowing LTM to coordinate with WM.

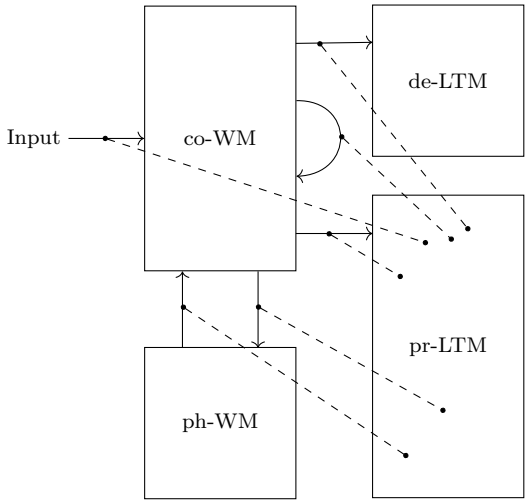


Figure 2.9: Basic memory architecture of FEELER —adapted from [Pfeifer 1982, 9]. Production rules in pr-LTM control interactions among co-WM, ph-WM, and de-LTM. Co-WM processes external input, updates itself, retrieves or writes data to de-LTM, and spreads activation in it. Input directly enter co-WM, enabling plan-progress comparisons. Co-WM also exchanges data with ph-WM, thus integrating physiological information. Productions ensure memory coherence and plan execution.

The system's WM is also articulated into two distinct units: the *cognitive working memory* (co-WM), and the *physiological working memory* (ph-WM). The former unit, co-WM, retrieves information (node structures representing plans) from de-LTM as inputs are received, in order to elaborate them and give appropriate outputs. On the contrary, ph-WM retrieves from de-LTM information about *arousal images* associated with nodes, i.e., *patterns of physiological activation*, on the basis of which it

then calculates the intensity of the emotion to be generated by [Equation 2.4].

The program thus connects all memories units, allowing them to retrieve information from, and write new information on, each other (see [Figure 2.9]). Information retrieval and writing are the standard actions of a set of *production rules* or *productions* (see [Footnote 27]) that, along with routine description, represent the fundamental structure of FEELER's program in pr-LTM. Production rules account for the procedural part of the generation process of a qualitatively specific emotion from an input —whereas de-LTM stores the components that make up for all those experiences, events, objects and concepts having emotional significance for an individual to which such rules apply. They govern the whole behavior of the model, relating inputs, plans, interrupts and outcomes to appropriate actions, responses and emotions, output by the system.

Productions have a traditional IF-THEN form. The IF-part of a rule (left-hand side, LHS) corresponds to conditions to be matched against the actual content of WM, retrieved from de-LTM and elaborated once the input is received. The THEN-part (right-hand side, RHS), conversely, corresponds to actions to be executed if the the conditions in LHS have been satisfied. Are hence elements in WM matching *all* LHS-conditions of a given rule at a give time, then that rule *fires*, executing its specified RHS-action as a result[28] (see [Table 2.4]).

Name	LHS	RHS
R1	an expected result does not obtain	generate an interrupt
R2	something unexpected happens	generate an interrupt
R3	the urgency of a goal/plan activated from LTM exceeds $_1$	generate an interrupt
R4	self is in state$_1$ and state$_1$ is negative for self and state$_1$ was caused by person$_1$ and person$_1$ was in control and the emotional target is person$_1$	generate anger at person$_1$
R5	interrupt has occurred and	determine target

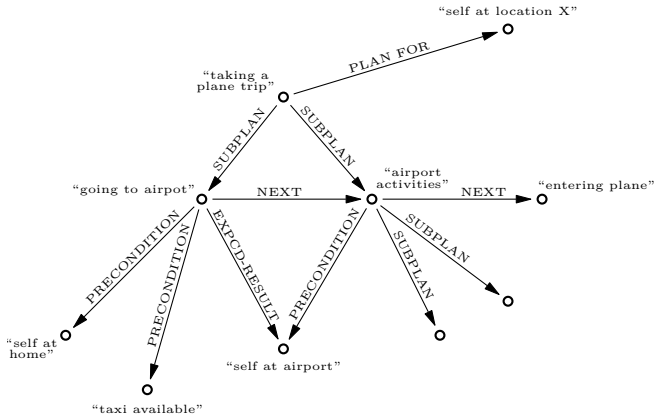
[28] Of course, elements in WM can match the condition of more than one rule, which leads to an issue of *conflict resolution*. The system faces it by selecting the rule with more logical strength or generality, or by considering the sum of the de-LTM activated elements.

	situation is to be determined	for action
R6	interrupt has occurred and situation is to be determined	determine if current state is negative for self
R7	subplan has been completed$_1$	generate satisfaction about subplan completion
R8	self has lost object and object was source of gratification	generate sadness about loss of object
R9	angry and person makes request to self	generate anger at person
R10	self and plan are to be realized and plan includes object	assert that object might be lost
R11	angry and emotional target is person	assert goal to harm person
R12	interrupt has occurred and situation is to be determined and emotional target is person$_1$	determine if person$_1$ was cause
R13	interrupt has occurred and situation is to be determined and emotional target is person$_1$	determine if person$_1$ was cause
R14	self sees car	generate anger at driver of car
R15	self approaches person and self is rejected by person	generate anger at person
R16	person$_1$ hurts person$_1$ and emotional target is person$_1$	generate anger at person$_1$
R17	values of self are challenged by person	generate anger at person
R18	person$_1$ is in state$_1$ and state$_1$ is negative for person$_1$ and person$_1$ had control and emotional target is person$_1$	generate anger at person$_1$
R19	person$_1$ hurts person$_1$	assert that person$_1$ is in state$_1$ and state$_1$ is negative for person$_1$
R20	person$_1$ hurts person$_1$ and the emotional target is person$_1$	generate anger towards person$_1$
R21	self has lost object and object was source of gratification	assert that current state is negative for self
R22	self is in state$_1$ and state$_1$ is negative for self and state$_1$ is due to loss of object	generate sadness about loss of object
R23	self is in state$_1$ and state$_1$ is negative for self and state$_1$ was caused by person$_1$ and	assert that the emotional target is self

	$person_1$ was in control and the emotional target is $person_1$	
R24	$self$ is in $state_1$ and $state_1$ is negative for $self$ and $state_1$ was caused by $person_1$ and $person_1$ was in control and the emotional target is $self_1$	thus generate pity $self$
R25	arousal greater than $threshold_1$	assert goal to assess cause of arousal
R26	angry	generate goal to count to ten
R27	$self$ feels pity at $person$	assert goal to help $person$

Table 2.4: List of FEELER's production rules (specifically for anger) —adapted from [Pfeifer 1982, 22-23].

The main routine of FEELER (see [Figure 2.10]) begins with an input generally consisting in a declaration of the goal the system has to achieve, for example "`the individual needs to attend an important meeting in another city`". The input, injected in co-WM by the programmer, activates the retrieval of all plan-related information from de-LTM representing the steps which the system has necessarily to take in order to achieve it, for example [Pfeifer 1982, 12]:



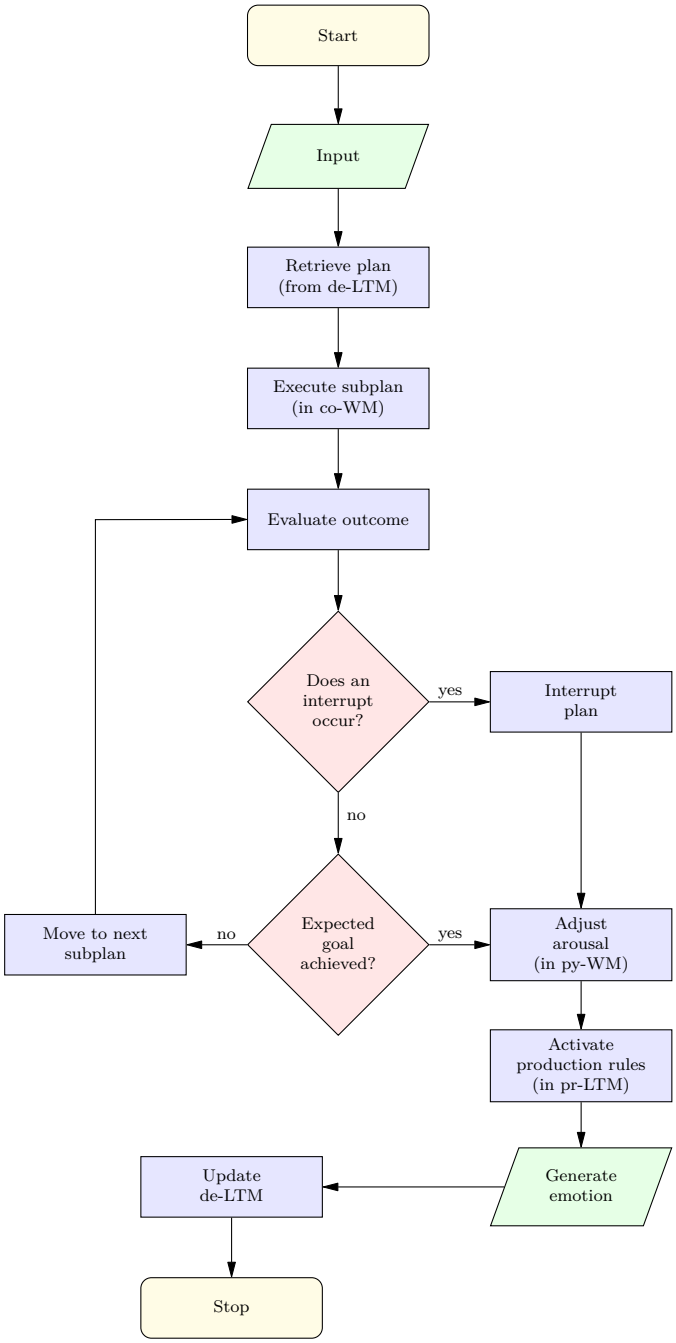
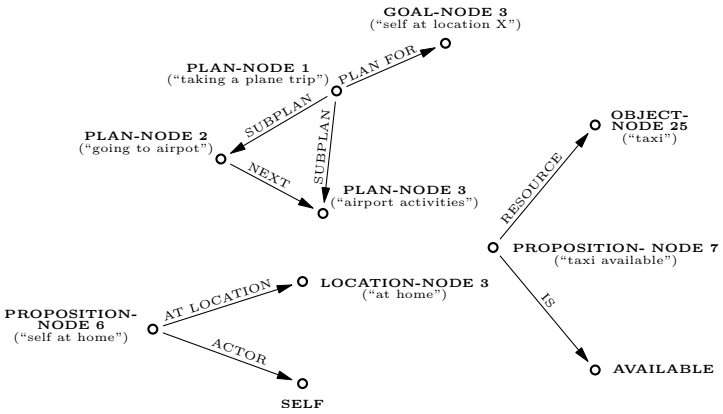


Figure 2.10: Simplified overview of FEELER's cognitive structure.

Relevant nodes and patterns, detected by suitable spreading activation schemes in de-LTM's network structures (see e.g., [Neches 1981]), are thus stored in co-WM for plan execution [Pfeifer 1982, 12]:

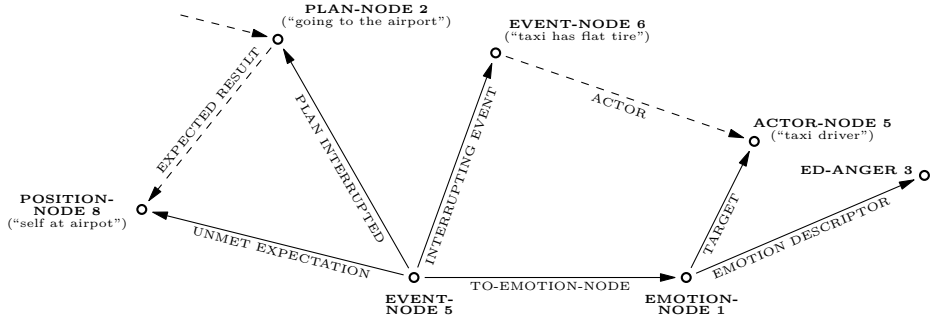


Every step of the plan the system has to follow toward the input goal is realized in the form of a subplan. For each subplan, the preconditions for its execution are checked by the system. More in particular, the current state of the system (e.g., the individual's being at home, a taxi being available, etc. for the first subplan) is confronted against the current state of the environment, i.e., the presence of further elements the programmer may (have) input in co-WM representing simulated sensory stimuli from the environment (e.g., stimuli of visual, auditory, tactile form, etc.) or the system's (physical and cognitive) internal state. If such preconditions are satisfied, the subplan is executed. Hence, successive subplans are executed until the process ends with reaching the final goal.

If, however, the preconditions of some subplan are not met, the correspondent subplan is not executed, to the effect that the expected goal cannot be achieved anymore, and the entire process towards it stops. At this point, rule R1 (or R2 and R3) fires, producing an *interrupt* in the main plan. If in co-WM, beyond the information retrieved from de-LTM, there is a node (or node pattern) related, say, to "`the taxi develops a flat tire`", the system will generate an interrupt while executing the relevant part of plan. At this point, the fired rule writes onto co-WM new information, in form of new interrelated nodes, about the location where the plan has been interrupted and about the cause of the interrupt. Then, arousal is calculated by ph-WM according to [Equation 2.4], whereafter the sub-routine of emotion generation is launched. In this sub-routine, the *quality* of the emotion is determined through specific rules fired by the interrupt and by other conditions applying to the individual and to the situation in which the interrupt occurred (e.g., R4, R15–R18). For

example, in the case under examination, a negative emotion such as anger directed at the taxi driver will be generated. Accordingly, the intensity of the resulting emotion is then calculated by means of [Equation 2.5].

Once the emotion has been generated and output —arousal adjustment and emotion generation sub-routines working similarly in the case the expected goal is achieved without interrupts, the system begins one last process in which the de-LTM is updated with all new information computed in co-WM by means of the rules. For example, all information regarding the flat tire, the anger at the taxy driver, etc. [Pfeifer 1982, 13]:



will be added in de-LTM and linked through suitable connections to the corresponding location where the plan of taking the flight was interrupted (`PLAN-NODE 2`).

This information belongs now to FEELER's knowledge representation of the world, and, interestingly, the system will use it, with all its emotional significance, for elaborating new inputs that may activate patterns within the memory structure that are in some way related to those previously triggered by the interrupted plan (or the achieved goal). For example, encountering a similar situation in the future, such as another delay or obstacle in reaching the airport, will lead the system to retrieve and integrate the previously stored emotional and contextual data. By looking carefully at the rules the program uses for processing inputs and plans [Table 2.4], in fact, it can be noticed that some rules apply directly to interrupts and emotions [Table 2.5]. This eventually allows FEELER to adapt its behavior dynamically to new situations by predicting potential issues, adjusting expectations, and generating appropriate emotional or cognitive responses, leveraging the emotional memory encoded in the system.

LHS	RHS		
	Cognitive		Emotional
	Assertion	Goal	
Cognitive	**(1)** R19, R21	**(2)** R5, R6, R12, R13	**(3)** R4, R7, R8, R14, R15, R16, R17, R18, R20, R22, R24
Emotional	**(4)** R10	**(5)** R11, R26, R27	**(6)** R9

Table 2.5: Classification of rules in FEELER —adapted from [Pfeifer 1982, 31]. The table categorizes the system's production rules based on their inputs (LHS) and outputs (RHS), illustrating how cognitive and emotional elements interact. Particularly, it highlights how some rules rely on emotional inputs to guide responses (cells 4 to 6), enabling the system to integrate past emotional experiences into its reactions to new situations, while others process cognitive assertions or goals to drive plan execution and evaluate outcomes.

2.4 The Evolution of Artificial Emotions

As the technical descriptions of [Section 2.2] and [Section 2.3] have concretely shown, the way of computationally modeling emotions underwent relevant changes since emotions finally entered the scene of computer science and AI during the 1960s [Section 1.2.4]. In particular, this evolution involved three different, but deeply interrelated dimensions: the theoretical conceptualization of emotions and, consequently, the way of structuring them into working computer programs [Section 2.4.1]; the computational approach taken in developing artificial emotions [Section 2.4.2]; and the attention to the physiological, as well as social and interrelational components of emotions [Section 2.4.3].

2.4.1 Theoretical and Computational Modeling of Emotions

With the only exception of Loehlin's [1963] ALDOUS among the models analyzed in [Section 2.2], all AAIMCO emerged from the attempt to operationalize and translate in a working computer program some theory of human cognition or human social interaction. This is for example the case of Gullahorn and Gullahorn's [1963] HOMUNCULUS, based on Homans's [1958; 1961] theory of social exchange, or Abelson's [1963]

Hot Cognition (HC), based on Heider's [1946] cognitive balance theory and Festinger's [1957] theory of cognitive dissonance. Perfectly in line with the aim of the scientific program of the emerging field of cognitive science, some AAIMCO were even developed under a clearly stated intent of using the computational model for testing the theory itself from which they stemmed from, like in the case of Colby's [1975] PARRY, developed for testing a cognitivist theory of paranoia. As *augmented AI models of cognition*, however, none of these models represented an attempt to directly and explicitly translate *theories of emotions* in computational terms.

After all, emotions began to attract the attention of cognitive scientists only in the late 1970s —that is, not surprisingly, during the same time the first AIME were also appearing. Following the challenges and criticisms of GOFAI and behaviorism [Section 1.2.4], in fact, it was just from the 1960s that cognitive scientists had started to seriously study emotions, in particular by applying cognitivist methodologies against the behaviorist or physiological (see e.g., [James and Lange 1922], and also [Section 3.2.1]) approaches still popular at the time. This resulted in the emergence of several *cognitivist theories of emotion* (see e.g., [Arnold 1960; Schachter and Singer 1962; Lazarus 1974; Mandler 1975; Weiner 1980; Lazarus 1982; Benesh and Weiner 1982; Lazarus 1984; Weiner 1985; Frijda 1986; 1988], etc. and also [Section 7.2.2]) between the 1960s and 1980s, pointing out the fact that emotions involve *cognitive* evaluations, mental attitudes, etc. beyond just behavioral and physiological components.

Being more prone to be formalized, and hence implemented as computer programs, such new cognitivist theories of emotion represented the theoretical basis on which new kinds of CMEs could develop. As *AI models of emotions*, these models had the generation of emotional behavior at their core, such as in the case of Pfeifer's [1982] FEELER, implementing computationally Schachter and Singer's [1962] and Mandler's [1975] cognitivist theories of emotions. Yet, they did not always take the confirmation of their underlying theories of emotions as an explicit or primary epistemic purpose.

Now, from the differences in how the relationship between underlying theoretical model and computational model of emotions is structured

in AAIMCO and AIME also a difference derives in the way emotions are specifically modeled and computationally implemented in both typologies of models. As already pointed out in [Section 2.1], in AAIMCO "emotion" is just a label given to various kinds of regulatory mechanisms in a system of purely cognitive, symbolic, operations. Emotion generation is not the main task of the program, which instead deals with decisions, beliefs, manipulation of concepts by means of rules, etc. In AAIMCO, emotions have the function of improving and *augmenting* the program's cognitive and computational performance, serving, for example, as frameworks for better associating actions to inputs (ALDOUS), as operational principles for measuring the convenience of certain social acts (HOMUNCULUS), etc.

Emotions are not conceived as independent internal states (neither of humans in general, nor of the system), but as a set of characteristics or properties (e.g., HC's *affect-ladenness*) subsumed under, or attached to, cognitive states of "higher order" like beliefs (HC), delusions (PARRY), etc. —whereby this set is often limited to two or three qualitatively different items. Accordingly, the way emotions are represented in the system is that of symbolic strings of numbers (ALDOUS) or discrete quantitative variables (HC) associated to cognitive elements in the program's memory, on which thus it relies for elaborating inputs and giving appropriate (cognitive) outputs. Interestingly, despite such restricted characterization of emotions in AAIMCO, these models anticipate cognitivists theories of emotions in highlighting the relevance of affective phenomena to cognitive processes, especially those related to decision-making (HC, PARRY).

The conception of emotion emerging from AAIMCO is therefore a highly static one, not even consisting in *temporal* states the system "experiences". Some of them integrate mechanisms for representing emotional intensity though, yet still as functional or operational concepts (ALDOUS' realistic potential or HC's affective imbalance) tied to threshold values for determining actions, admissible credences and other kinds of cognitive outputs.

As a later AAIMCO, however, in PARRY the emerging conception of emotion is different from that of previous models. Even if not aimed at generating emotions as AIME, PARRY incorporates a more nuanced

representation of emotional processes. In PARRY, for example, there is a clear and systematic characterization of emotional intensity, which beyond serving cognitive output generation, also determines a true emotional state the system finds itself in while interacting with an operator. In other words, the system simulates *having* emotions, even if without expressing them explicitly, but just implicitly by means of appropriate language expressions reflecting a more or less paranoid condition. Moreover, PARRY's emotional states are subjected to fluctuations over time, as several I-O interactions take place. This makes emotions in PARRY something more dynamic than just symbolic values attached to cognitive beliefs and list items in the memory. Rather than having only a functional or operational purpose for cognitive tasks, they represent states the system experience *temporally*, for paranoia itself, as a psychological condition, is inherently characterized by fluctuating emotional states.

Thus, also due to the influence on computer science of emerging cognitivist theories of emotions, PARRY and late-1970s AAIMCO represented a pivotal step in CME, going from the static augmentations of cognitive processes of previous approaches to the more integral, temporally dynamic and *felt/expressed* emotions of the AIME. This new conception of emotions is namely realized and implemented in full in models like FEELER, that for example represents emotions as *online* affective states, and is aimed at generating emotional behaviors, thereby not just modeling emotions as a functional, operational or regulatory mechanism for cognition. In this sense, such models —by redefining emotion as a process to be dynamically generated and expressed— can be seen as direct precursors of the paradigm of *Affective Computing* that would emerge almost two decades later [Chapter 4].

2.4.2 Computational Approaches

Whereas AAIMCO and AIME differ substantially as to the way emotions are modeled, more similarities can be found in the computational tools they use for knowledge representation and the architecture of the program.

As to knowledge representation, both AAIMCO and AIME do not take any distance from the symbolic paradigm relying on coded (e.g.,

ALDOUS), list-shaped (e.g., HOMUNCULUS and HC) or graph-shaped (PARRY, FEELER) taxonomies, in which cognitive elements are organized and associated following temporal, causal or conceptual dependency criteria.

The same essentially applies for the program's architectures. All AAMICO (obviously) and AIME rely on top-down logical rules for associating actions to given inputs (e.g., ALDOUS, HOMUNCULUS, FEELER), admitting new beliefs in the system (e.g., HC), for conceptualizing inputs (e.g., PARRY), etc. Yet most AIME often involve *cybernetic* components [Section 2.1], generally related to numerical (continuous) variables for representing the variation of emotional intensities or arousal, decay mechanisms, etc. (e.g., FEELER). For this reason, they are *hybrid*, because of joining both a symbolic and a cybernetic computational approach to model emotions.

From this perspective, however, later AAIMCO emerge again as an exception. PARRY, in fact, already incorporates cybernetics-inspired emotion decay mechanisms, later characteristic of AIME. Moreover, it uses a numerical (albeit still discrete) representation of emotional intensity. Also from the point of view of the computational approach employed, models like PARRY thus represent a clear step forward towards more responsive and dynamic CMEs, also influenced by the 1970s setbacks and criticisms of GOFAI.

Finally, both AAIMCO and AIME show minimal learning capacities, keeping memory of past emotional assessments (e.g., ALDOUS, HC), emotional interactions (HOMUNCULUS, PARRY) or emotional experiences (e.g., FEELER) in elaborating new inputs. Yet, neither AAIMCO nor AIME —also referring to other models in [Table 2.1]— show learning capacities in the terms of the subsymbolic approach to AI [Section 3.1], with the program being able not only to update its knowledge, but also its rules —an approach which will effectively spread in AI in the 1990s, and be applied to the CME above all boosted by Picard's [1995; 1997] systematic proposal (see [Chapter 4]).

2.4.3 Physiology and Sociality

As a last point in the examination of the evolution of CMEs, let us consider the way emotions are modeled in AAIMCO and AIME as to their physiological, social, interpersonal and relational side.

As it can be noticed from the descriptions of the models [Section 2.1], AIME are the only models which account for the physiological component of emotions (e.g., FEELER) [Section 3.2.1]. This seems to be clear, especially as previous AAIMCO were not interested in building computer programs for simulating, generating and expressing emotions, but used emotions only as regulative mechanisms for augmenting cognitive processes [Section 2.4.1] . Therefore, in AAIMCO there is no need to account for the physiological part of emotions. Even intensity, which is commonly associated with physiological arousal —if present at all— is instead always conceptualized in cognitive terms, for example as the degree of imbalance in beliefs (e.g., HC), the distance from flare concepts (e.g., PARRY), etc.

On the contrary, almost all AAIMCO and AIME are, more or less implicitly or explicitly, involved with the social and relational nature of emotions [Section 3.3]. A system like HC could implicitly work in a potentially social environment, as upsetting beliefs also derive from assertions formulated, or actions performed, by other individuals. The same applies for FEELER, where emotions-eliciting interrupts may be caused by other individuals to the system. Hence, as in this case, AIME did not automatically represent an evident step forward compared to AAIMCO toward the computational modeling of emotions (also) as social phenomena, although they to some extent acknowledged the presence of a social component characterizing affective phenomena.

Yet, surprisingly among AAIMCO, it is possible to find *interactive* computational system dealing with emotions explicitly as a social, interpersonal phenomenon. As explained, for example, a later version of ALDOUS was developed which would interact with another ALDOUS, and respond "emotionally" to its inputs/outputs. As one of the first chatbots developed, PARRY was designed as a system that interacts with a psychiatrist in a simulated diagnostic interview, whereby the operator can stimulate the program in various ways. Similarly to the case

of the interpersonal ALDOUS, at the first International Conference on Computer Communications in 1972, in Washington, DC (US), PARRY was even connected trough ARPANET to Weizenbaum's [1966] ELIZA, simulating a Rogerian therapist to whom PARRY responded by its usual emotions-mediated paranoid outputs. HOMUNCULUS was even developed, as a single case in CMEs history, as an operationalization of a sociological theory involving emotions. Indeed, the system —which modeled the individual as a social information processing system in the true spirit of cognitivism— regarded emotions even primarily as states logically influencing social decisions, rather than as affective states in the traditional sense. Rather than being a model of emotional social interaction, it was a model (formal) social decision-making.

In general, none among early CMEs accounts in depth for the interactional component of emotions —a mild exception being maybe PARRY. On the contrary, introducing sociality and relationality of affective phenomena in CME will be the revolution Picard's [1995; 1997] *Affective Computing* (see [Chapter 4]) will operate in the field during the 1990s, also following social and psychological studies about emotions that, influenced by the so called *Emotional Turn* (see e.g., [Lemmings and Brooks 2014]) will specifically highlight the nature of emotions as interpersonal and relational experiences [Section 3.3].

Summary

Building on the historical-epistemological framework established in the previous [Chapter 1], which traced the emergence of emotions as a legitimate object of inquiry within AI, this chapter has examined their first computational implementations, from the symbolic experiments of the 1960s to the more integrated architectures of the 1980s. It has reconstructed the evolution of early CMEs, distinguishing between the Augmented AI Models of Cognition (AAIMCO) and the AI Models of Emotions (AIME). Through the analysis of representative case studies —ALDOUS, HOMUNCULUS, Hot Cognition, and FEELER— it has shown how, in the former, emotions were conceived as auxiliary control variables modulating cognitive processes, whereas, in the latter, how

they acquired a central computational and adaptive function, revealing a gradual transition from a symbolic to a cybernetic conception of affect.

This development marked a turning point in the history of CME, revealing both the potential and the limitations of symbolic modeling —particularly its inability to account for physiological grounding and sociality of emotions. The emergence of cybernetic and hybrid approaches in CME thus anticipated many of the theoretical and methodological innovations that would later be systematized with the paradigm shift introduced by Rosalind W. Picard's *Affective Computing* [Chapter 4]. Before turning to that transformation, however, it is necessary to examine the technical and philosophical roots that, developed between the 1980s and the 1990s, made such a shift possible —namely, the rise of subsymbolic AI, of embodied approaches to the study of the mind, and the reconceptualization of emotions as cultural and social objects. This will be the extensive focus of [Chapter 3].

Chapter 3

The Technical and Philosophical Roots of *Affective Computing*

In previous [Chapter 1] and [Chapter 2], a history of early computational models of emotions has been traced. In particular, it has been pointed out that artificial emotions firstly emerged as an operational tool for improving the performance of computer programs mimicking cognitive tasks, and that just later they started to be implemented as states that a machine can simulate and express.

In the 1980s, computers running AIME programs [Section 2.3] expressed particular theoretical views on emotions developed in the psychology and cognitive science of those years. They also had the generation of emotional states as a temporal process as an explicit computational desideratum. Yet, despite outdoing relevant aspects of the symbolic AI paradigm, they did not entail forms of automatic learning characteristic of later subsymbolic AI. Nor, in most cases, they truly took into account emotions as embodied, interpersonal and social phenomena [Section 2.4].

Starting from the mid-1990s, the subsymbolic approach to computation, based on Artificial Neural Networks and Machine Learning [Section 3.1], began to revolutionize the field of AI, gradually superseding GOFAI and expert systems. At the same time, relevant studies in cognitive science were shedding light on the close relationship linking emotions (and all mental faculties) to the brain, and the body in general. Further

studies in psychology and sociology were highlighting the impossibility of understanding emotions outside an interrelational context. Also, research in philosophy, anthropology, ethnology, history and the humanities in general was introducing a true *Emotional* or *Affective Turn* that, in giving renewed importance to emotions as scientific study objects in those fields, stressed their centrality in shaping human behavior, practices and meaning-making, yet not as static universals, but as historically and culturally determined constructs.

Here, at a time when computers were becoming increasingly popular in houses and workplaces, requiring humans to interact with them in many and different forms, *Affective Computing* (AC) emerged as a new, revolutionary approach to CME. Originally proposed by Rosalind W. Picard [1995; 1997], American computer scientist at MIT, AC systematized CME into an autonomous discipline, consolidating previous isolated efforts to apply advanced AI techniques to artificial emotion modeling — including both neural networks and a broader range of automatic learning methods, connectionist or statistical, that enabled systems to learn from data without relying exclusively on symbolic programming— and integrating them with the most recent views on emotions from cognitive science, psychology, social science, and philosophy. In doing so, AC gave CME a new programmatic trajectory, specifically aimed at building responsive machines capable of meaningful emotional and social interaction with humans [Section 4.1].

In this chapter, I continue tracing the history of artificial emotions by connecting early CME to AC. In particular, I concentrate on the aforementioned theoretical preconditions that led to the emergence of this new computational approach to emotions. In [Section 3.1], I focus on the subsymbolic approach in AI, characterized by Artificial Neural Networks and Machine Learning. Moreover, in [Section 3.2], I analyze the contribution of medicine, psychology, cognitive science, philosophy and related disciplines to the reconceptualization of emotions as together cognitive and embodied phenomena. Finally, in [Section 3.3], I will focus on emotions as interrelational and social phenomena, including the *Affective Turn* in the humanities and the reconceptualization of emotions as cultural and historical.

This will serve as a basis to introduce the main ideas of AC in [Chapter 4], as well as of social and affective robotics in the 2000s [Chapter 5], an actual concretization, as it will be illustrated, of Picard's *manifest*.

3.1 The Subsymbolic Approach

Between the 1970s and the 1980s, the limitations of symbolic AI, especially expressed by the inability of expert systems to generalize and adapt beyond narrowly defined problems, led to the field's first *winter* [Section 1.1.3]. However, from the need of more flexible AI systems, able to learn from data, recognize patterns and adapt their behavior over time on the basis of feedback from outside, a gradual shift to different approaches originated.

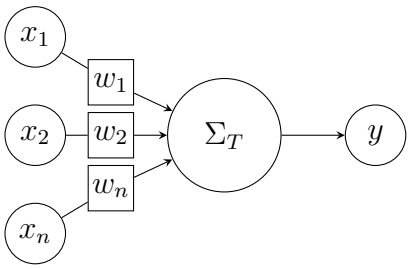
In particular, a transition from rule-based to data-driven methodologies occurred, which prioritized adaptability and scalability over logical explicitness. Such methodologies were collectively subsumed under the label of *subsymbolic AI* (see e.g., [Rumelhart 1984; Rumelhart et al. 1986b; Smolensky 1988]), with reference to the idea that knowledge and processing in these systems are *below* the level of explicit symbolic representation, involving instead abstract, non-symbolic patterns that are learned mostly from data by the system itself.

In subsymbolic AI, representations are in fact *implicit* and *distributed*: There is no *a priori* definition of symbolic categories like in symbolic AI. Rather than being represented by discrete symbols, knowledge is derived *bottom-up* from data and encoded across many processing units, with *patterns of activation* representing information. Processing is then as well distributed and continuous rather than based on pre-programmed rules, allowing the system to learn and adapt through data.

In the 1980s, the interest in subsymbolic methodologies led to the rise of *connectionism* and *artificial neural networks* (ANNs), which allowed AI to exit the "cold" phase of decay it was experiencing since the previous decade (see e.g., [Russell and Norvig 2020; Wooldridge 2022]). Indeed, the conceptual core of such research frameworks had been already posed some four decades before by McCulloch and Pitts [1943] and their studies on the mathematical modeling of the human nervous system (see [Section 1.1.1]). The two American scientists had proposed that biologi-

cal neurons can be mathematically modeled as simple (electrical) logical circuits: Given the «"all-or-none" character of nervous activity» [115], the behavior of neurons can be described by means of the laws of propositional logic and Boolean algebra.

Extending and popularizing in AI their foundational ideas, Frank Rosenblatt [1958] later devised one of the first theoretical hardware architecture for an ANN, the so-called *perceptron*[29]. The perceptron Σ consists of a single-layer neural network algorithm capable of performing binary classification tasks by linearly separating input data points in a multi-dimensional space:



It operates by receiving inputs $x_{1,\dots,n}$ with weights $w_{1,\dots,n}$, and producing a single output y, corresponding to the classification of the inputs either in one category ($y = 1$) or in another ($y = 0$). Each x is either active or inactive. If active, it stimulates Σ by the corresponding w. Σ will however produce y just in case the weighted sum of the active inputs received (u) [Equation 3.1] meets or exceeds a threshold T associated with Σ [Equation 3.2].

$$u = \sum_{i=1}^{n} w_i x_i \tag{3.1}$$

$$y = \begin{cases} 1 & \text{if } u \geq T \\ 0 & \text{if } u < T \end{cases} \tag{3.2}$$

[29] Prior to Rosenblatt's [1958] architectural systematization, some actual neural network computers had been already built. An example is the electromechanical SNARC (Stochastic Neural Analog Reinforcement Calculator), built by Marvin L. Minsky and Dean S. Edmonds at Harvard in 1950 (see e.g., [Minsky 1952]).

The perceptron's learning process involves adjusting the weights based on the error between the predicted output and the actual target value. By iteratively updating the weights during training, the perceptron learns to make correct classifications for linearly separable data.

The perceptron was also implemented as a computer program for the IBM 704 at the Cornell Aeronautical Laboratory (*Mark I Perceptron Machine*), being used for image recognition tasks. Yet, although it showed some efficacy in recognizing visual patterns, such as simple shapes or letters, it was limited to solving only linearly separable problems, being insufficient for more complex patterns requiring non-linear, multi-dimensional, decision boundaries, such as the famous XOR function [Minsky and Papert 1969]. Suppose, two datasets, respectively derived from the AND and XOR function, are input in Σ:

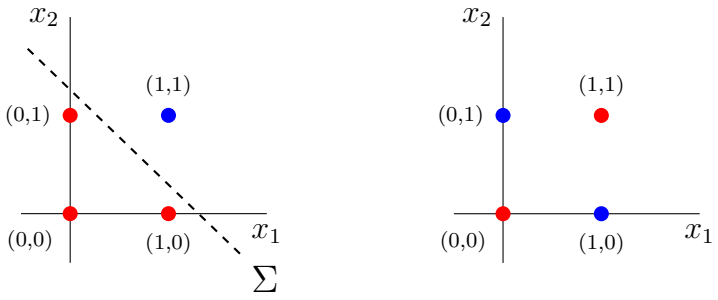
x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

AND

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

XOR

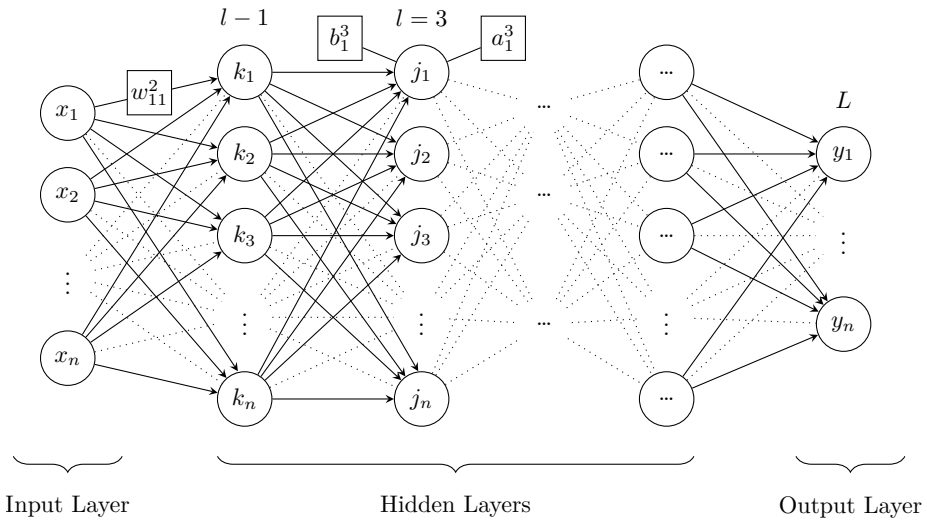
At this point, the perceptron computes a weighted sum $u = w_1x_1 + w_2x_2$ and applies a threshold function to determine the output y [Equation 3.2]: In the case of the AND function (presupposing that each input activates the perceptron), Σ will be able to correctly classify the data, dividing outputs $y = 1$ (blue) from outputs $y = 0$ (red), whereas, in the case of the XOR function, this will turn out to be impossible:



AND *XOR*

Cases such as the *XOR* function showed that improvements to McCulloch and Pitts's [1943] original ideas were necessary. As significant steps forward in the field did not occurred until the mid-1980s, ANNs and subsymbolic AI turned primarily out to be practically unuseful, especially compared to the successes that GOFAI and expert systems were instead achieving in the meantime.

In particular, in order to compensate the drawbacks of early connectionist approaches, more complex ANN architectures were needed, i.e., ones involving *multi-layer ANNs* (see e.g., [Wooldridge 2022, 115ff.]). Differently than single-layer ANNs, multi-layer ANNs (or multi-layer perceptrons, MLP) consist of an array of many interconnected artificial neurons (i.e., of many perceptrons), organized on different layers (input layer, hidden layers, output layer):



Neurons in the different layers are connected to each other, whereby connections are of course weighted like in perceptrons. In particular, w_{kj}^l

designates the weight of the connection of the kth neuron in the $(l-1)^{th}$ layer to the jth neuron in the lth layer. To each jth neuron in the lth layer a so-called *bias* value b_j^l is also associated. Conceptually, biases serve the same function as thresholds T in perceptrons, determining the activation of a neuron and controlling whether the input is propagated to the next layer. Differently than thresholds, however, biases are not fixed cut-offs, but values that are *adjustable* and *trainable*, allowing, in this way, higher flexibility and learning capacities to the AAN.

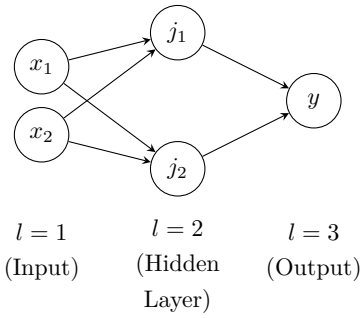
The activation a_j^l of a neuron j is computed by means of an *activation function*

$$a_j^l = f(z_j^l) = f\left(\sum_{k=1}^{n} w_{k,j}^l a_k^{l-1} + b_j^l\right) \tag{3.3}$$

whereby z_j^l is the weighted sum of the inputs received from layer $l-1$ and biases of neurons in l *before* activation, and f is a function —e.g., sigmoid, hyperbolic tangent (tanh), Rectified Linear Unit (ReLU), etc.— that introduces non-linearity into the model, allowing the network to learn complex mappings from inputs to outputs. Of course, outputs y correspond to neuron activations a^L in the last layer L (output layer).

As it can be noticed from [Equation 3.3], the activation of a neuron depends both on the weighted sum of the active inputs received from neurons in the previous layers *and* on the bias values of these. Whereas in single-layer ANNs a neuron activates only if u equals or exceeds a *fixed* threshold T [Equation 3.2], in multi-layer ANNs activation is more dynamic in that the activation threshold is continuously adapted by the bias value b_j^l for each neuron. By shifting appropriately $\sum_{k=1}^{n} w_{k,j}^l a_k^{l-1}$, the bias b_j^l dynamically alters the neuron's propensity to activate, which eventually depends on the nature of f. A neuron relying on ReLu, for example, whereby $f(z) = \max(0, z)$, will activate for $z > 0$ and remain inactive for $z \leq 0$; one relying on sigmoid, whereby $f(z) = \frac{1}{1+e^{-z}}$, will activate for $z \gg 0$ (i.e., $a_j^l \approx 1$) and remain inactive for $z \ll 0$ (i.e., $a_j^l \approx 0$); etc. Consequently, multi-layer ANNs are capable of overcoming the limitations of single-layer perceptrons by introducing more complex network architectures, whereby the non-linear activation and adjustable biases allow the ANN to learn how to trace complex decision boundaries and therefore to solve also non-linearly separable problems.

This is, for example, exactly the case of the *XOR* function, solvable by an ANN with at least one two-neurons hidden layer and, say, the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$. Given two neurons in the input layer $l = 1$ taking on input values x_1, $x_2 \in \{0, 1\}$ of the *XOR* function (see table above) as their activation values,

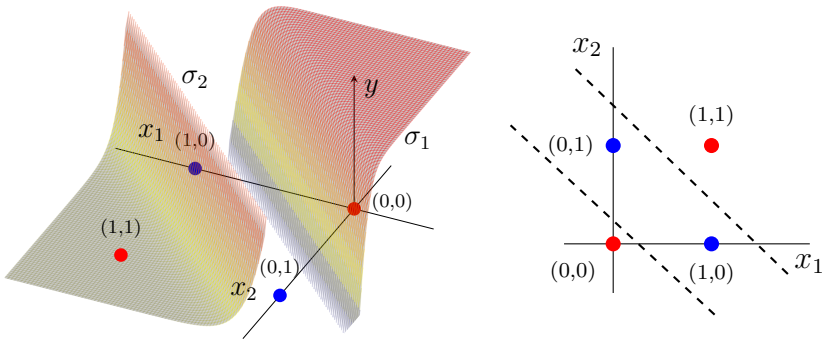


for each neuron $j = 1$ and $j = 2$ in the hidden layer $l = 2$ with suitably trained weights $w^2_{k,j}$ and biases b^2_j, their activations a^2_1 and a^2_2 are calculated according to [Equation 3.3] and sigmoid function as

$$\begin{aligned} a^2_1 = \sigma_1(z^2_1) &= \sigma_1(w^2_{1,1}a^1_1 + w^2_{2,1}a^1_2 + b^2_1) \\ &= \frac{1}{1 + e^{-(w^2_{1,1}a^1_1 + w^2_{2,1}a^1_2 + b^2_1)}} \end{aligned}$$

$$\begin{aligned} a^2_2 = \sigma_2(z^2_2) &= \sigma_2(w^2_{1,2}a^1_1 + w^2_{2,2}a^1_2 + b^2_2) \\ &= \frac{1}{1 + e^{-(w^2_{1,2}a^1_1 + w^2_{2,2}a^1_2 + b^2_2)}}. \end{aligned}$$

In the hidden layer $l = 2$, inputs are non-linearly transformed so that decisional (hyper)surfaces are created in the inputs space by means of which the ANN is able to correctly separate the values of the *XOR*. In this way, at the point where a normal perceptron would crash, such an ANN would instead give appropriate y values of *XOR* (see table above) as outputs in $l = 3$. To give an approximate graphical example:



As such, multi-layer ANNs were not entirely a novelty in the 1980s. After Rosenblatt's [1958] perceptron, in fact, several (more or less advanced) models of multi-layer ANNs were developed (see e.g., [Widrow and Hoff 1960; Rosenblatt 1962; Ivakhnenko and Lapa 1967; Ivakhnenko 1971; Grossberg 1976; Anderson 1977; Fukushima 1980]). These models led to valuable results especially in the context of visual pattern recognition (images and characters) —see e.g., [Amari 1967; Fukushima 1969; 1975] etc.— with each ANN hidden layer processing specific visual features of the input at ever increasing levels of abstraction and semantic meaningfulness (i.e., firstly edges and lines in the lowest layers, then textures and shapes in the intermediate layers, and then objects and letters in the higher levels). However, they did still not rely on *learning techniques* for appropriately updating weights and training bias values so to make the ANN work in an efficient and fine-grained way appropriate for all kinds of inputs.

Such techniques, instead, were at the core of a true revolution which would have occurred in AI only after the end of its first winter. In this respect, a groundbreaking innovation was namely the introduction of the *backpropagation algorithm* (BPA). Developed independently by Werbos [1982][30] and Rumelhart et al. [1986a,b] —partly relying on previous work concerning similar learning techniques (see e.g., [Bryson and Ho 1969; Linnainmaa 1970; Werbos 1974; LeCun 1985], etc.)— this algorithm eventually enabled the effective training of multi-layer ANN,

[30] Curiously, Werbos [1994, 702] reports that the core idea of BPA occurred to him while reflecting on the dependency relation between human behavior and emotions, in particular as an attempt to formulate a mathematical version of Freud's [1914] theory of *cathexis* (Ger. *Besetzung*, tr. by [Freud 1925]), consisting in a sort of emotional energy individuals attribute to people and objects.

becoming the most popular one to be used to this scope by AI scientists from then until present times.

BPA works in two phases: the *forward pass* and the *backward pass*. In the forward pass, it computes an output $\mathbf{y}$ by propagating forward input $\mathbf{x}$ through all layers l of the ANN, whereby $\mathbf{x}$ and $\mathbf{y}$ are represented from here on as vectors for convenience:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

In other words, thus, in this phase the algorithm performs nothing different than the process described earlier (e.g., in the case of the *XOR* function), calculating activations $\mathbf{a}^l_j$ for each neuron k, j in each layer l given weighted inputs $\mathbf{z}^l_j$, i.e., weights $\mathbf{W}$ and bias $\mathbf{b}^l_j$ —whose values are represented from here on also as vectors and matrices, respectively, for convenience:

$$\mathbf{a}^l = \begin{bmatrix} a^l_1 \\ a^l_2 \\ \vdots \\ a^l_j \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} w^1_{1,1} & w^2_{1,2} & \dots & w^l_{1,j} \\ w^l_{2,1} & w^l_{2,2} & \dots & w^l_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ w^l_{k,1} & w^l_{k,2} & \dots & w^l_{k,j} \end{bmatrix} \quad \mathbf{b}^l = \begin{bmatrix} b^l_1 \\ b^l_2 \\ \vdots \\ b^l_j \end{bmatrix} \quad \mathbf{z}^l = \begin{bmatrix} z^l_1 \\ z^l_2 \\ \vdots \\ z^l_j \end{bmatrix}$$

Of course, the algorithm performs this process for the whole *dataset*. A dataset is a collection of individual *training samples*, i.e., of all individual m input-output tuples ($\mathbf{x}^m$,$\mathbf{y}^m$) computed by the ANN, whereby each individual input —e.g., a single image with features $x = x^m_1, x^m_2, \dots, x^m_n$ (number of pixels, color, etc.)— is assigned to an output —e.g., labels $y = y^m_1, y^m_2, \dots, y^m_n$ (dog, cat, etc. in case of classification tasks).

Now, before the ANN is launched, the values of weights and biases are typically *initialized*. Biases are often initialized to zero or small random values, while weights are initialized using random distributions optimized for the activation function used by the ANN (e.g.„ Xavier or Glorot initialization for sigmoid or tanh, He initialization for ReLU, etc.). This particularly ensures that symmetry is broken in the ANN (i.e., with

weights and biases not being the same for all neurons), which is crucial for effective learning and informative output generation. Such initialization, however, does not also ensure that, given input $\mathbf{x}$, the *expected output* $\mathbf{y}$ is *directly* produced. In the forward pass, in fact, based on initialized weights and biases, the ANN produces a *predicted output* $\widehat{\mathbf{y}}$ that may possibly differ from the *expected output* $\mathbf{y}$. At this point, the "distance" of the expected output from the predicted output, i.e., the *error* E, is namely calculated for each m-th training sample by a suitable *loss function*

$$E = \mathcal{L}(\mathbf{y}^m, \widehat{\mathbf{y}}^m). \tag{3.4}$$

There are many possible functions $\mathcal{L}$ that can be used to this scope, whereby the choice depends on the kind of tasks the ANN must perform, i.e., the kind of outputs it has to produce. Typical loss functions, for example, involve Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for regression tasks (calculating the value of a continuous variable given one or more independent features), or Binary Cross-entropy, Categorical Cross-entropy and Sparse Categorical Cross-entropy for classification tasks (classifying inputs into different categories given their features). Finally, in the last step of the forward pass, the calculation of the error by the loss function $\mathcal{L}$ is generalized to all m training samples of the dataset. This is done by means of a *cost function* $\mathcal{C}$ computing the average of individual losses across the entire dataset:

$$\mathcal{C}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\mathbf{y}^m, \widehat{\mathbf{y}}^m), \tag{3.5}$$

with θ representing the collection of all the parameters of the model, including weights $\mathbf{W}$ and biases $\mathbf{b}^l$ across all layers l of the ANN.

The goal of the forward pass, in other words, is to quantify the error between expected and actual outputs, providing in this way a useful starting point for training the ANN. This training conceptually involves the adjustment of ANN parameters in order to *minimize* this error and align actual outputs $\mathbf{y}$ with expected outputs $\widehat{\mathbf{y}}$. Now, this training is exactly the content and goal of the backward pass.

In the backward pass, the error is *backpropagated* from the output to the input layers throughout the whole ANN in order to fulfill such

training purpose, i.e., by finding an effective way to update the parameters of the network in order to reduce the value of the cost function $\mathcal{C}(\theta)$ [Equation 3.5].

As known from above, the cost function $\mathcal{C}$ depends from the output $\widehat{\mathbf{y}}$ [Equation 3.5], correspondent to activations $\mathbf{a}^L$ of the output layer L. In turn, $\mathbf{a}^L$ is computed by the activation function [Equation 3.3], which takes weighted inputs $\mathbf{z}^L$ to neurons in L as its domain. Weighted inputs $\mathbf{z}^L$ moreover depend on the activations $\mathbf{a}^{L-1}$ in the hidden layer $L-1$ [Equation 3.3], which in turn depend on weighted inputs $\mathbf{z}^{L-1}$ to neurons in $L-1$. This process repeats for each hidden layer $l < L$, with activations at layer l depending on weighted inputs from the previous layer $l-1$. This means that the error in L, as quantified by the cost function, reflects the extent to which activations from $L-1$ contributed to that error, which also applies, recursively, for every hidden layer l, $l-1$ until the input layer.

Now, propagating the error (cost) at layer L *backwards* across all hidden layers $l < L$ amounts to calculating the magnitude of the contribution of each layer —through the weights and biases of their individual neurons— to the final cost. The quantification of this contribution eventually makes the proper adjustment of the parameters possible, so to minimize that cost and align actual and expected output.

Mathematically (starting from the output layer L), this quantification takes the form of the computation of the partial derivatives $\frac{\partial \mathcal{C}}{\partial w^L_{k,j}}$ and $\frac{\partial \mathcal{C}}{\partial b^L_j}$ of the cost function $\mathcal{C}$ with respect to weights $w^L_{k,j}$ and biases b^L_j of individual neurons in L. The collection of these partial derivatives is a vector representing the *gradient* of the respective parameters. Given the *composite* nature of $\mathcal{C}$, gradients can be *efficiently* calculated by means of the *chain rule of calculus* —namely used for computing the derivative of composite functions:

$$\frac{d}{dx} f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx}$$

So, applying it to compute the gradient of the cost function with respect to the weights $\mathbf{W}$ and biases $\mathbf{b}$ of neurons in L, it respectively yelds:

$$\frac{\partial \mathcal{C}}{\partial \mathbf{W}^L} = \frac{\partial \mathcal{C}}{\partial \mathbf{a}^L} \cdot \frac{\partial \mathbf{a}^L}{\partial \mathbf{z}^L} \cdot \frac{\partial \mathbf{z}^L}{\partial \mathbf{W}^L}, \tag{3.6}$$

$$\frac{\partial \mathcal{C}}{\partial \mathbf{b}^L} = \frac{\partial \mathcal{C}}{\partial \mathbf{a}^L} \cdot \frac{\partial \mathbf{a}^L}{\partial \mathbf{z}^L} \cdot \frac{\partial \mathbf{z}^L}{\partial \mathbf{b}^L}. \tag{3.7}$$

whereby $\frac{\partial \mathcal{C}}{\partial \mathbf{a}^L}$ is the derivative of $\mathcal{C}$ with respect to $\mathbf{a}^L$, $\frac{\partial \mathbf{a}^L}{\partial \mathbf{z}^L}$ the derivative of $\mathcal{C}$ with respect to $\mathbf{z}^L$, and $\frac{\partial \mathbf{z}^L}{\partial \mathbf{W}^L}$ the derivative of $\mathbf{z}^L$ with respect to $\mathbf{W}^L$, indicating how much the former change as a function of the latter, given dependencies mentioned above (this being the same for the gradient of biases $\mathbf{b}^L$).

The calculated gradients give information about the contribution of the output layer L to the error as quantified by the cost function. However, as the cost function recursively depends on the weighted inputs and activations of all hidden layers $l < L$, the chain rule for error backpropagation must be generalized to the whole ANN, beginning with $L-1 = l$. Specifically,

$$\frac{\partial \mathcal{C}}{\partial \mathbf{W}^l} = \frac{\partial \mathcal{C}}{\partial \mathbf{a}^l} \cdot \frac{\partial \mathbf{a}^l}{\partial \mathbf{z}^l} \cdot \frac{\partial \mathbf{z}^l}{\partial \mathbf{W}^l}, \tag{3.8}$$

$$\frac{\partial \mathcal{C}}{\partial \mathbf{b}^l} = \frac{\partial \mathcal{C}}{\partial \mathbf{a}^l} \cdot \frac{\partial \mathbf{a}^l}{\partial \mathbf{z}^l} \cdot \frac{\partial \mathbf{z}^l}{\partial \mathbf{b}^l}, \tag{3.9}$$

whereby $\frac{\partial \mathcal{C}}{\partial \mathbf{a}^l}$ depends on the error from layer L:

$$\frac{\partial \mathcal{C}}{\partial \mathbf{a}^l} = \frac{\partial \mathcal{C}}{\partial \mathbf{z}^L} \cdot \frac{\partial \mathbf{z}^L}{\partial \mathbf{a}^l}, \tag{3.10}$$

and here $\frac{\partial \mathcal{C}}{\partial \mathbf{z}^L}$ is the gradient of activations $\mathbf{a}^L$ with respect to $\mathcal{C}$ already computed earlier for the output layer [Equation 3.6], and $\frac{\partial \mathbf{z}^L}{\partial \mathbf{a}^l}$ represents the contribution of the activations of layer $l = L-1$ to the weighted inputs $\mathbf{z}^L$ of layer L.

By recursively applying the procedure for all hidden layers $l-1$, l until the input layer, the gradients for the parameters of all neurons across the network are computed. Such gradients provide information about the extent to which (initialized) weights and biases in the different layers contributed to the error quantified by the cost function. Moreover, they reflect how sensitively the cost function would change with respect to (even small) adjustments in each weight an bias in the ANN. In other words, they tell how much $\mathcal{C}$ changes if parameters are adjusted. If the gradient is large, for example, even minimal adjustments will significantly affect the cost function, and vice versa.

Therefore, gradients give hints about how much adjustment is needed to the value of the parameters in order to minimize the output error. Thus, once calculated, they are used for updating the parameters according to appropriate *gradient descent* procedures[31], i.e., procedures aimed at finding the minimum value of a function. For each layer l weights and biases are iteratively updated as

$$\mathbf{W}^l_{\text{new}} \leftarrow \mathbf{W}^l_{\text{old}} - \eta \cdot \frac{\partial \mathcal{C}}{\partial \mathbf{W}^l}, \tag{3.11}$$

$$\mathbf{b}^l_{\text{new}} \leftarrow \mathbf{b}^l_{\text{old}} - \eta \cdot \frac{\partial \mathcal{C}}{\partial \mathbf{b}^l}, \tag{3.12}$$

whereby η represents the *learning rate*, an adjustable (fixed or dynamic) scalar value that controls the step size of the update for each iteration.

By these procedures, BPA eventually allowed the efficient training of ANNs, which during the 1980s were experiencing a great new phase of study and popularity, with the development of many models of feed-forward ANNs (whose general structure has been described here), recursive ANN (applying the same set of weights recursively), convolutional ANN (applying convolutional filters to extract features), etc. This eventually made AI able to handle many of the complex problems that were before impossible to solve by means of symbolic methodologies, especially in the field of visual pattern recognition.

Moreover, BPA-based ANNs played a pivotal role in advancing what was beginning to coalesce into a broader computational paradigm, that of *Machine Learning* (ML). Coined by Arthur Samuel [1959] for indicating the (back then, future) possibility to «program computers to learn from experience [without] the need for much [...] detailed programming effort» [42]— the term referred to an idea that had evidently been formulated long before the emergence of the subsymbolic paradigm. Already

[31] It is to point out that backpropagation may work differently depending on the gradient descent procedure chosen. For example, with Stochastic Gradient Descent (SGD) the backward pass begins after the computation of the loss function for randomly selected individual training samples [Equation 3.4] (i.e., after each output emission). On the other hand, with Batch Gradient Descent (BGD) the backwards pass begins after the computation of the cost function for the whole dataset [Equation 3.4]. The choice depends on the kind of task and on the nature of the database. SGD, particularly suitable for very large datasets, allows faster computations, albeit less accurate. BGD fits better for smaller datasets, but may be slower, albeit more accurate.

Turing [1950, 454ff.] for example, noticed the advantages of building child-like machines, with simple mechanisms and huge predisposition to learn. McCulloch and Pitts [1943] also prefigured the possibility that suitably construed neural networks may learn [Russell and Norvig 2020, 35], and Donald O. Hebb [1949] even proposed what would later be known as the *Hebbian rule*, suggesting that the repeated co-activation of neurons could strengthen their connection —a biologically inspired idea that would later inform learning mechanisms in artificial neural systems. As pointed out in [Chapter 2], also early computational models of mind and emotions exhibited some rudimental learning capacities, by storing in memory the outputs of previous computations, used successively in the elaboration of new inputs.

Still, before modern ANNs, several ML algorithms had also been developed, even managing to achieve notable success in different domains, from medical diagnosis and strategic game solving, to text categorization and pattern recognition in speech or handwriting. Many of them —like, e.g., CART (Classification and Regression Trees) [Breiman et al. 1984], ID3 (Iterative Dichotomizer 3) [Quinlan 1986], C4.5 [Quinlan 1993], etc.— classify data by recursively selecting binary features of increasing abstraction, guiding the process through a sequence of decisions that gradually narrow down to a final outcome —that is, *decision tree* algorithms of *supervised* ML, i.e., in which the algorithm infers general rules from a set of training samples labeled with the correct outputs. Other early supervised learning methods included algorithms such as the *k-nearest neighbors* (k-NN) [Fix and Hodges 1951; Cover and Hart 1967; Bentley 1975], that classifies new inputs by comparing them to the k most similar instances in the training set, and assigning the most common label among them. On the other hand, also *unsupervised* ML methods gained relevance —i.e., methods in which the algorithm infers general rules from a set of unlabeled training samples purely from the statistical properties and distribution of the input features. Notable examples are the *Expectation-Maximization* (EM) algorithm [Dempster et al. 1977; Wu 1983], that tries to uncover hidden regularities in the data by alternating between guessing missing information and adjusting the model accordingly, and the *k-means clustering* [Lloyd 1982], that groups

similar inputs together by repeatedly assigning them to one of k clusters based on how close they are to a central example.

These algorithms have been even applied to CME in the late 1990s [Section 4.1], and are still used today for a variety of data-driven tasks across classification, segmentation, and pattern recognition —some of them being even integrated into ANN architectures, either as preprocessing steps for feature selection and dimensionality reduction, or as components within hybrid models combining symbolic decision structures with subsymbolic learning dynamics.

Nonetheless, the full-scale application of ML models and ANNs models faced considerable challenges throughout the early 1990s. Training neural networks, in fact, required solid hardware and computational capabilities that were still unavailable at the time. Moreover, the "raw material" needed for such training, that is *big data*, could be provided back then just to a barely sufficient amount in order to ensure effective trainings and practical applications of *machine learning algorithms*, as they became known. These and other hindrances caused AI, in early 1990s, to enter its *second winter*. Again, the over-optimistic hopes that, in this case, accompanied the emergence of specialized ANN systems were faced with a gradual shift of the interested of academicians towards other computational approaches and technologies and with sensible cuts of funding for the development of ANN-based AI.

Only in the mid-1990s, AI began to exit from this winter. To this, new advancements in computational techniques and refinements of existing ANN and ML algorithms —e.g., of *Support Vector Machines* (SVMs)— turned out to be crucial. Equally important were hardware improvements with consequent enlargement of computational power of machines —including, e.g., the development of *Graphics Processing Units* (GPUs), allowing for faster parallel processing, and increased availability of digital data. And, above all, the *Internet boom* played a decisive role by providing the massive datasets needed for effectively training ANN systems.

All this reignited the enthusiasm of scientists and investors towards the potential of ANN-based AI technologies, which, by the 2000s, culminated in the rise of *deep learning* (DL) as a particularly advanced form of machine learning (from which, indeed, the adjective "deep"). The development of more sophisticated ANN architectures and the introduc-

tion of other hardware innovations, allowed now DL models to handle sequential data, making them particularly effective for tasks involving temporal dependencies, such as real-time language natural language processing, speech recognition synthesis and automatic translation. By the 2010s, DL had practically become synonymous with AI, with its applications rapidly expanding across industry, from autonomous vehicles to personalized recommendations in digital platforms, research and education, art, finance, etc. Finally, DL laid the groundwork on which, recently, also *Generaitve-AI* and *Large Language Models* (LMM) could be developed. Such systems are able today not only recognize patterns in data, but also *produce* new content, such as text, images, music and other forms of creative output, indeed by leveraging DL-based computational techniques.

3.2 New Perspectives on Emotions: Emotions as Embodied Cognitive Processes

As a necessary precondition for AC, there was not only the development of better computational techniques such as those described in the last section. Also relevant changes in the scientific and philosophical conception itself of emotions were needed in order to bring about an evolution in the way of artificially modeling them. As it will be examined in the following, these changes occurred in a number of different disciplines between the 1980s and the 1990s, and they primarily involved a new conceptualization of emotions as strongly embodied phenomena [Section 3.1.2.1].

3.2.1 Emotions and the Body

In [Section 1.2.4], it was pointed out that, as early as the 1960s and 1970s, it was becoming increasingly clear to AI- and cognitive scientists that cognition could not be (computationally) modeled without accounting for the role that emotions play in it. This realization was hypothesized to stem primarily from the decline of neopositivism and behaviorism — with their negative, irrational, and unscientific stance on emotions— as

well as from the initial failures of GOFAI and the symbolic approach, despite their widespread popularity.

At the same time, these insights into the relationship between emotions and cognition were soon paralleled, in cognitive science and AI, by reflections about the fundamental connection between emotions and the *body*. Albeit new in these disciplines, such reflections —similarly to those on the emotion-cognition problem (see e.g., [Footnote 15])— were not entirely absent from philosophical thought, even (long) before the 1960s.

Even if in a negative and irrational acceptation, in fact, emotions, as *affecti* or *passiones animi* (from the Ancient Greek πάθος, *pathos*, meaning "passion"), had been traditionally considered as states mostly driven by the body (see also [Section 1.2.1]). This was for example the case of Ancient philosophers and physicians like Aristotle [1984a] or Galen [1963]. The latter, in particular, believed that emotions are linked to so-called *humors*, bodily fluids whose movements, balance and interactions were thought to cause different temperaments and emotions to arise. Galen's ideas remained highly influential throughout the medieval and early modern periods, shaping medical and psychological theories (of emotions) until at least the 18th century. Especially during the Early Modern period, indeed, physiological accounts of emotion gained significant support from a growing body of scholars. In medicine, for example, Thomas Wright [1604] and Robert Burton [1621] framed emotions as inherently embodied phenomena, arising from the interplay of physical processes such as the movement of spirits, humors, or bodily heat. On the other hand, philosophers like empiricists John Locke [1689] and David Hume [1739] emphasized the relationship between emotions and the body, particularly in arguing that, for the intellect to function properly, the senses —and the emotions necessarily linked to them— must be regulated and cultivated through experience and reason.

It was only at the end of the 19th century, however, that properly defined *physiological theories of emotions* were put forward, relying, in some cases, even on strong empirical evidence, and resulting highly influential until present debates about the nature of emotions. At this time, in fact, both medicine and psychology were undergoing a radical epistemological shift. Steering away from the metaphysical speculations

of the past, such disciplines were now adopting more empirical and experimentally driven research methodologies in their respective (but often connected) fields of study (see e.g., [Freedheim and Weiner 2003; Baker 2012; Jackson 2011]). On the one hand, the clinic and the laboratory became the major places where well-structured anatomical, physiological, and pathological investigations had to be conducted. Medical hypothesis about diseases and biological processes within the organism were formulated in quantitative terms after systematic observations, measurements and experiments. Still, on the other hand, psychology was emerging as an independent discipline embracing too experimental methods to study mental processes. Psychologists increasingly began to explain perception, sensation, cognition, etc. through the quantitative vocabulary of physiology and neurology, prioritizing objective scientific inquiry over the philosophical introspection of previous approaches.

3.2.1.1 Bernard, Helmholtz and the Psychophysics

Among the best representatives of this epistemological shift was Claude Bernard. Often considered the father of experimental medicine [Gross 1998] and modern neuroscience [Conti 2002], Bernard [1858] firstly advanced the hypothesis that the organism regulates itself to maintain a stable internal equilibrium. Such idea of equilibrium or *homeostasis* (from the Ancient Greek ὅμοιος, *homoios*, and στάσις, *stasis*, broadly meaning "staying the same") was relatively groundbreaking at a time when it was believed that organs functioned independently, and will prove later fundamental for the development of cybernetics (see [Section 1.1.1] and [Section 5.1.2]). Recalling —interestingly— the Galenic notion of the regulatory power associated to the bodily fluids, Bernard considered particularly the blood to be the key vector of the mechanisms ensuring the balance of the *milieu intérieur* (from the French, roughly, "internal environment"): blood vessels constrict when it is cold and dilate when it is warm; blood sugar levels are regulated through the glycogenic function of the liver; and so on. The *milieu* is constantly adapted, in response to stimuli from the external environment, through physical and chemical mechanisms internal to, and constitutive of, the body itself.

A similar concern in explaining biological phenomena only by referring to observable and quantifiable processes of physical or chemical nature was shown by the German physicist and physician Hermann von Helmholtz in studying perception. In both the case of hearing [Helmholtz 1862] and vision [Helmholtz 1867], perception was accounted for by him in terms of physiological mechanisms, i.e., of the mechanic interaction of sensory receptors with physical stimuli. Expanding upon Thomas Young's [1807] trichromatic theory of vision, he described the eye as containing three types of cone receptors —each sensitive to red, green, or blue light— whose varying stimulation, processed by the brain, produces the perception of the different colors. Similarly, he modeled the ear as a system of resonators in which the different parts of the cochlea resonate to specific frequencies of the sound producing the perception of the different tones after conversion into neural signals by the brain.

A significant contribution to a more naturalistic and mechanistic interpretation of psychological phenomena was also given by Gustav Theodor Fechner [1860], founder of *psychophysics*, a discipline that investigates the quantitative relationships between physical stimuli and sensory perception. Relying on previous studies by Ernst Weber [1834; 1851], he characterized this relation even in the form of a mathematical equation, known as *Weber–Fechner Law*:

$$S = k \cdot \log\left(\frac{I}{I_0}\right).$$

The perceived (subjective) sensation S is proportional to the logarithm of the intensity I of the physical stimulus that elicits it, where I_0 is the threshold for the minimum perceptible intensity of the stimulus, and k is a constant that depends on the type of stimulus and the sensory modality involved (e.g., k varies for sound, light, or pressure based on the specific sensitivity of the sensory system). Since this relationship is logarithmic, it implies that if the intensity of a stimulus varies geometrically, the corresponding perception varies arithmetically.

The shift toward a mechanistic and body-centered view of medicine and psychology, exemplified by these studies, also paved the way for a new understanding of emotions. Similarly to other mental faculties, emotions too, in fact, came now to be recognized as deeply intertwined with

physiological processes, arising as embodied responses to stimuli from the external environment. As a result, researchers increasingly sought to trace them through observable patterns of neural activity, biochemical interactions, and the dynamic interplay of bodily systems.

3.2.1.2 Charcot and Darwin

The existence of a causal interdependence between emotional states and the body's *milieu intérieur* was demonstrated by the French neurologist Jean-Martin Charcot[32], particularly in the case of *hysteria*. In the 19th century, hysteria was still classified as a well-defined, diagnosable medical disorder, mostly affecting women and linked to physiological, moral or even spiritual causes. Its origin was believed to be in dysfunctions of the uterus, sexual deprivation or even emotional excess and demonic possession; its symptoms included fainting, edemas, hyperemia, nervousness, insomnia and spasms, whereas common therapies encompassed pelvic massage, autoeroticism, marriage and clitoridectomy.

Charcot [1872] challenged this understanding of hysteria. In particular, basing on clinical observations, he opposed the view of it as an exclusively female pathological condition. Moreover, even if not relinquishing its definition as an independent diagnosis —which will only happen in mid-20th century— he conceived of it explicitly as a *neurological* disorder characterized by the insurgence of physiological symptoms (hemiparesis, hemianesthesia, contractures of the extremities and involuntary movements, etc.) that followed recognizable neurological patterns despite the absence of noticeable organic causes (brain damages, lesions to the nervous system, etc.). Interestingly, he also observed that even strong emotional experiences, psychological traumas, or past distressing

[32]Even if never expressly quoted in Charcot's scientific production, Bernard's experimental and physiological approach doubtlessly influenced him. After all, the direct and indirect interaction of Claude Bernard with Jean-Martin Charcot were many. During his internship at the *Hôpital de la Charité* in 1851, Charcot was introduced to Bernard's clinical observation methods through Pierre Rayer, a renowned pathologist who had been one of Bernard's early supporters and teachers. Later, he was admitted to the *Société de Biologie* with the support of notable figures, including Bernard. In 1872, Bernard played a crucial role in Charcot's appointment to the chair of pathological anatomy at the Faculty of Medicine of the University of Paris. Additionally, in 1876, Bernard appointed Charcot to a commission tasked with studying Victor Burq's experiments on metallotherapy.

events could act as triggers for hysterical symptoms. Also, therapies that, by hypnosis or physical estrangement, removed stress-eliciting and emotionally disrupting situations, helped alleviated hysterical symptoms. This suggested that emotions clearly played a role in generating somatic —particularly neurological— responses, reinforcing the idea of an essential interdependence between mental states and physiological processes.

A special attention to the relationship between emotions and neurological activity was also given by Charles Robert Darwin [1872]. Emotions have physiological correlates (e.g., trembling, blushing, crying, etc.) whose occurrence depends on the activity of the autonomic nervous system, and is therefore outside the realm of conscious will. Moreover, Darwin [1872] embedded the idea of a deep-seated neural basis of emotional states (i.e., « *The principle of actions due to the constitution of the Nervous System, independently from the first of the Will, and independently to a certain extent of Habit*» [29]) into a broader biological understanding of emotions as evolutionary adaptations in both humans and animals. Emotions, along with their physiological correlates and bodily expressions (e.g., facial movements, gestures, etc.), evolved as (universal) communicative signals that served adaptive purposes. Physical behaviors that were originally useful for survival and associated to the occurrence of a particular emotion became habitual and hereditary, so that they continue to manifest with that emotion even if they are no longer essential for survival (i.e., « *The principle of serviceable associated Habits*» [28]). Opposing emotions are expressed through contrasting physical actions (i.e., « *The principle of Antithesis*» [28]). All these ideas framed emotions as biologically rooted phenomena shaped by natural selection. In this respect, Darwin introduced a form of *biological universalism* in the scientific study of emotions that will influence, especially in a *reductionist* acceptation, both later approaches in medicine and psychology [Section 3.3], and implementational practices in CME, AC and robotics (see [Section 4.2.2], [Section 4.3], [Section 6.1] and [Part II]).

3.2.1.3 The James-Lange Theory

The growing emphasis on the physiological basis and effects of emotions found one of its most compelling expressions in the work of the Amer-

ican philosopher and psychologist William James. James [1884; 1890; 1894] was among the first scientists in modern times to develop a systematic theory of emotions expressly targeted at explaining their nature and causes.

Between 19th and 20th century *feeling theories of emotions* were still enjoying a certain popularity, that had been lasting since at least Greek Antiquity. As a good representation of everyday understanding of emotions, in them these latter were viewed as a class of subjective feelings not different in nature from other sensory experiences, such as feeling the warmth of a fire on the skin or feeling the pressure of the feet on the ground while standing —a view that will be partly reintroduced by emotional perceptualism in contemporary analytic philosophy [Section 7.2.2.2]. Emotions were seen as psychic states, «mental affections» [James 1884, 189], caused by objects, events or state of affairs in the world and ontologically different from the physical changes they induced in the body, which were thus causally and temporally subsequent to them (see e.g., [Scarantino and Sousa 2021; Moors 2022; Campeggiani 2023]).

Against this picture, James [1884; 1890, 1894] proposed his own theory of emotions, which, despite belonging to the feeling tradition, diverged from it in at least two major respects. Firstly, it denied that emotions are psychic states ontologically independent and separate from the body; Secondly, it inverted the causality of the relationship between emotions and physical states. Thus, in James' account, emotions came to be viewed as the (*subjective*) *perception of the bodily changes* triggered by the awareness of an external event:

> «Our natural way of thinking about these standard emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My thesis on the contrary is that *the bodily changes follow directly the* PERCEPTION *of the exciting fact, and that our feeling of the same changes as they occur* IS *the emotion*. Common sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect, that the one mental state is not immediately induced by the other, that the bodily manifestations must first be interposed between, and that the more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful, as the case may be. Without the bodily states following on the perception, the latter would be purely cognitive in form, pale, colorless, destitute of emotional

> warmth. We might then see the bear, and judge it best to run, receive the insult and deem it right to strike, but we could not actually feel afraid or angry» [James 1884, 189-190, italics in the original].

In accounting for the bodily mechanisms responsible for emotional phenomena, James showed a tight adherence to the new physiological paradigm of the psychology of his time. Building upon the studies on the relationship between sensations, perceptions and nervous system by Helmholtz and Fechner (see e.g., [James 1890, Vol. I: 85ff., 152ff., 553ff., Vol. II: 17ff., 142ff., 216ff., etc.]), he especially acknowledged the crucial role of the *cerebral cortex* as a processing center giving rise to the perception of bodily changes, and hence to emotions. When external objects stimulate one or more sense organs, the *nervous system* transmits an impulse to the brain. At this point, the cortical response is propagated, always by the nerves, to the *viscera* (hearth, lungs, intestine, liver, etc.) and the *muscles* before returning to the brain. It is exactly in the impulses from the body reaching again the cortex that the perception of the bodily changes originate. And, it is exactly at this stage that an emotion arises.

Influenced by Charcot (see e.g., [James 1890, Vol. II: 475ff.]) in acknowledging the essential role of the nervous system in the emotional experience, and by Darwin in contextualizing it within an evolutionary framework (see e.g., [James 1884, 191ff.]), James made however no assumptions, as later theories of emotions will do, about the correspondence of specific emotions to specific brain areas. He rather considered the brain, along with whole nervous system, the viscera and the muscles, as an integral component of a broader network of physiological reactions involved in emotional experience.

On the contrary, the prioritization of one bodily component over the others as primarily relevant in the eliciting process of emotions, was at the core of a modern feeling theory of emotions advanced by another important scientist from late 19th century, the Danish physician Carl Georg Lange [1885]. Even maintaining, like James, that emotions arise from the perception of bodily changes, and implicitly recalling Bernard's ideas about the regulatory function of the vasomotor system, Lange identified the regulation of blood circulation —particularly fluctuations in heart

rate, blood pressure, and vasoconstriction— as the primary physiological mechanism underlying emotional experiences.

Further theories of emotions broadly considering them as arising from the perception of bodily changes (see e.g., [Mosso 1884; Sergi 1894; Ribot 1896; Sutherland 1898], etc.) continued to appear until late 19th century. Despite their differences, all of them were subsumed under the general label of *James-Lange theory of emotions* (JLTE), in honor of their two most important representatives (see [Dunlap 1922]).

3.2.1.4 The Cannon-Bard Theory

At the turn of the century, JLTE had already become internationally popular among physicians, psychologists and philosophers. Skeptics, however, did not hesitate to express even serious objections to it. Among the first and most striking criticisms, there was one formulated by the American physiologist Walter Bradford Cannon [1927] (see also [Cannon 1915; 1929a,c; 1932]) who, together with a doctoral student of him, Philip Bard [1934], proposed a theory of emotions alternative to JLTE known with the name of *Cannon-Bard theory of emotions* (CBTE).

CB's criticisms to JL essentially concerned the ontology and the causality of emotional phenomena as understood by the latter. JL claimed that an emotion *is* the perception of the bodily changes *caused* by an external object. The emotion *does not exist* before such changes and *is generated by* and *with* the perception thereof. On the contrary, CB claimed that, even if there is a *correlation* between emotions and the body —as emotions are not purely mental, but always have a physiological and organic *component*— they are causally not generated by the bodily changes and the perception thereof, nor are they reducible those.

In fact, according to CBTE, bodily changes are neither necessary nor sufficient for the occurrence of an emotion. This thesis was put forward partly relying on empirical evidence that was unavailable at the time when JL wrote. By experiments of decerebration and sympathectomy performed on animals, Sherrington [1900] and Cannon et al. [1927] respectively demonstrated that the separation of the viscera from the brain, and the disruption of the autonomic sympathetic control of the vasomotor system, do not yield significant effects on the occurrence of emotional

phenomena. Furthermore, Cannon [1927, 109ff.] also noted that bodily changes (e.g., heart rate acceleration, increase of blood sugar, discharge of adrenaline, widening of the pupils, erection of hairs, etc.) are *underdetermined* with respect to specific emotions (e.g., fear or rage), and even characteristic of non-emotional states (e.g., fever or hypoglycemia) —a point later highlighted by affective cognitivists (see [Section 2.3], [Section 2.4.1], [Section 6.2.1] and [Section 7.2.2.2]). Additionally, considering empirical studies about the physiology of the nervous system (see e.g., [Sertoli 1883; Langley 1889; Pavlov and Schumows-Simanowsky 1895; Stewart 1900; Wells and Forbes 1911; Wells 1925]), he pointed out that visceral changes take place in a too slow and imprecise manner to determine instantaneous emotional reactions to objects and qualitatively complex emotional experiences. Also, he stressed that artificially inducing organic changes in the body —for example by injections of adrenaline— barely results in the generation of genuine emotions.

CB also reconsidered the role of the brain in the process of emotion generation. In JLTE, the cortex was responsible for the perception of the bodily changes, and thus for the production of emotions. Cannon [1927, 115ff.] however, invoked again recent empirical studies that demonstrated the absence of an active role of the cortex in emotion generation processes. Experiments on decorticated animals (e.g., [Bechterew 1887; Cannon and Britton 1925]), for example, showed that in the presence of appropriate stimuli, affective responses would always arise even without the contribution the cortex. In humans, similarly, a reduced cortical activity (as an effect of e.g. anesthesia or alcoholism) does not necessarily result in lessened emotional occurrences (e.g., [Bard 1928; Cannon 1928]). On the contrary, such empirical findings suggested that rather *subcortical* regions of the brain like the diencephalon —in particular *thalamus* and *hypothalamus*— were responsible for the generation of emotions: Lesions, pathologies or (experimental) surgical procedures that interfere with the functioning of such brain areas immediately produce consequences on the emotional behavior of the human or animal subject considered.

All these considerations eventually led CB to confute JL and propose their own theory of emotions. In CBTE, emotions still have an organic basis, yet this lies not directly and primarily in the body (viscera, mus-

cles, vasomotor system, etc.) like in JLTE, but in the *subcortical region* of the brain. This latter produces, *parallelly* and *independently*, *both* the physiological changes characteristic of the different emotions (through the activation of the autonomic nervous system), *and* all the subjective feelings and cognitive states associated with them (through influence on the cortex):

> «The neural organization for an emotion which is suggested by the foregoing observations is as follows. An external situation stimulates receptors and the consequent excitation starts impulses towards the cortex. Arrival of the impulses in the cortex is associated with conditioned processes which determine the direction of the response. Either because the response is initiated in a certain mode or figure and the cortical neurons therefore stimulate the thalamic processes, or because on their centripetal course the impulses from the receptors excite thalamic processes, they are roused and ready for discharge. That the thalamic neurons act in a special combination in a given emotional expression is proved by the reaction patterns typical of the several affective states. These neurons do not require detailed innervation from above in order to be driven into action. Being released for action is a primary condition for their service to the body —they then discharge precipitately and intensely. Within and near the thalamus the neurons concerned in an emotional expression lie close to the relay in the sensory path from periphery to cortex. We may assume that when these neurons discharge in a particular combination, they not only innervate muscles and viscera but also excite afferent paths to the cortex by direct connection or by irradiation. The theory which naturally presents itself is that *the peculiar quality of the emotion is added to simple sensation when the thalamic processes are roused* [...]. When the thalamic discharge occurs, the bodily changes occur almost simultaneously with the emotional experience» [Cannon 1927, 120, italics in the original].

Interestingly, the role assigned by CB to the autonomic nervous system in the emotional experience also characterizes emotions as an adaptive mechanism that is launched in response to given environmental stimuli. The physiological changes that are activated by the thalamus, in fact, contribute to maintaining the body's internal balance. In other words, according to CBTE, emotions can be conceived as *homeostatic* processes —a concept Cannon [1929b] introduces building explicitly upon Bernard's [1858] notion of *milieu intérieur* [Section 3.2.1.1]. The function of emotions within the homeostatic processes of the organism is so relevant that, for example, extreme emotional conditions such as stress or fear may make the organism's internal balance collapse, potentially leading to fatal consequences (see [Cannon 1929c], and also [Robinson 2018]).

By emphasizing the role of emotions as a regulatory system within the organism, CBTE contributed not only to the consolidation of the physiological paradigm in medicine and psychology, but also to the reinforcement of the *embodied* conception of emotions [Section 3.2.7]. By mid-20th century, this theory had largely supplanted JLTE, becoming an account of emotions among the most widely acknowledged within the scientific community. Yet, CBTE as well had soon its share of reservations, coming especially from philosophically inclined research psychologists and physicians.

3.2.1.5 George Dumas

One of the earliest scholars to highlight the possible weaknesses of CBTE was the French physician, psychologist and philosopher George Dumas. Dumas dedicated to the study of affective phenomena practically his whole life (see e.g., [Dumas 1895; 1900; 1906; 1922; 1923; 1926a,b; 1932; 1948; 1949]). In this respect, Dumas [1922] as well criticized the plausibility of JLTE and related *peripheral* explanations of emotional phenomena[33] —i.e., explanations making changes in the body's peripheral systems (e.g., heart, skin, muscles, etc.) responsible for the process of emotions generation, with the brain just playing a passive role in it. This steered him towards ascribing larger explanatory power to *central* approaches to the study of emotions, that, evidently, had been advanced already before Cannon's [1927] critique of JLTE and formalization of CBTE. Yet, Dumas [1922] judged the even fair amount of empirical evidence produced in support of these approaches [Bechterew 1885; Sherrington 1900; Janet 1901; François-Franck 1905; Piéron 1907; Mignard 1909; De Somer and Heymans 1912] —indeed barely mentioned by Cannon [1927], as Dumas' works themselves— to be certainly convincing, but not robust enough for *crucially* rejecting JLTE in favor of them: «[I]l reste encore, avec les muscles, la peau, et les vaisseaux de la tête et du cou, assez d'éléments périphériques pour exprimer et conditionner une

[33] It is also to point out that Dumas himself translated into French *Über Gemütsbewegungen: Eine psycho-physiologische Studie*, the German translation of [Lange 1885] made by Hans Kurella (indeed, the version read by James).

émotion»[34] [252]. Even if not sufficient for determining the emergence of an emotion, for which the function of the brain remains unquestionable, peripheral sensations, still contribute to shaping important *qualitative* and *content-related* aspects of emotional experience, which cerebral activation alone does not explain:

> «Il y a en effet, dans ces deux émotions [joy and sadness] et dans leurs différentes formes, un ensemble de sensations périphériques, musculaires, respiratoires, circulatoires, thermiques *dont le rôle n'est pas contestable dans le contenu de l'émotion*, encore qu'il ait été exagéré par la théorie Lange-James [...]. Nous sentons, dans la tristesse, le froid de nos extrémités, le frisson de notre peau, la lourdeur de nos muscles et ces sensations sont désagréables; nous sentons dans la joie, la chaleur de notr peau, le tonus de nos muscles, l'accélération de notre coeur et ces sensations sont agréables; de même dans la colère et la peur»[35] [257, my italics],

whereby, if one were to subtract such (subjective) bodily sensations from those emotional states, these latter would radically change. The subject would still have an emotion —here the distance from JLTE— but an emotion which is essentially different, and differently experienced, from the one which is instead accompanied by those bodily changes. For example, Dumas [1922, 254ff.] admitted the possibility of experiencing so-called *passive emotions* (such as deep sorrow or mystical bliss), distinct from *active emotions* (such as anger or excitement), as they may occur even in the absence of overt bodily activation, relying primarily on cerebral processes, whereby between the passive and active version of the same emotion there are of course strong *qualitative* differences. For these reasons, he continued, bodily changes also contribute to the *differentiation* and *individualization* of emotions [257-258], something that early central theories of emotions even struggled to fully account for —possibly, given the difficulty, among other things, of exactly mapping emotions over single independent patterns of cerebral activation.

[34] «There still remain, with the muscles, the skin, and the vessels of the head and neck, enough peripheral elements to express and determine an emotion» (my tr.).

[35] «There are indeed, in these two emotions [joy and sadness] and in their various forms, a set of peripheral, muscular, respiratory, circulatory, and thermal sensations *whose role in the content of the emotion is unquestionable*, even though it has been exaggerated by the Lange-James theory [...]. In sadness, we feel the coldness of our extremities, the shiver of our skin, the heaviness of our muscles —and these sensations are unpleasant; in joy, we feel the warmth of our skin, the tone of our muscles, the acceleration of our heart —and these sensations are pleasant; and likewise in anger and fear» (my tr.).

Thus, without abdicating the physiological stance on emotions, that made *bodly arousal* a however essential *component* of them, Dumas [1922, 258, my italics] advocated for the necessity of *integrating central and peripheral perspectives*, as neither approach alone —nor any singular theory, for that matter— can fully account for the complexity of emotions:

> «Si l'on veut bien admettre - et le fait n'est pas contestable— qu'une grande partie des sensations organiques qui font partie d'une émotion y apportent ainsi *non seulement leur contenù sensitif mais leur contenu affectif*, on pourra distinguer dans la plupart des émotions:
>
> 1° Des éléments proprement affectifs, agréables ou désagréables, qui se rattachent à la fois à l'activité facile, difficile, accélérée, ralentie, désordonnée, paralysée de nos tendances et de nos fonctions organiques et qui peuvent être *d'origine centrale ou d'origine périphérique*.
>
> 2° Un contenu sensitif qui, sans être proprement affectif, contribue dans une large mesure, par les sensations périphériques qui le constituent, à *diversifier*, à *individualiser* et à situer l'émotion dans le corps.
>
> Les éléments sensitifs seront toujours périphériques, contrairement aux éléments agréables ou pénibles qui se présentent le plus souvent comme *à la fois périphériques et centraux*, encore qu'ils puissent apparaître comme périphériques dans certaines dépressions et dans certaines joies organiques d'où toute peine morale et tout plaisir moral semblent absents ou uniquement cérébraux comme dans certaines joies et certaines tristesses que nous venons de citer»[36].

However, differently from JL, and later CB, Dumas did not defend the physiological stance on emotions in a *reductionist* and *eliminativist* fashion. Since «la plus grande erreur qu'on puisse commettre, en présentant une théorie, serait de considérer cette théorie comme suffisamment explicative»[37] [Dumas 1922, 258], he *syncretically* tried to bring together

[36] «If one is willing to admit —and the fact is indisputable— that a large part of the organic sensations which form part of an emotion contribute *not only their sensory content but also their affective content*, one may distinguish in most emotions:

1° *Properly affective elements*, pleasant or unpleasant, which are linked both to the easy, difficult, accelerated, slowed, disordered, or paralyzed activity of our tendencies and organic functions, and which may be *of central or peripheral origin*.

2° A *sensory content* which, though not properly affective, contributes in large measure —through the peripheral sensations that constitute it— to *diversifying*, *individualizing*, and situating the emotion within the body.

The sensory elements will always be peripheral, in contrast to the pleasant or painful elements, which most often appear as *both peripheral and central*, although they may appear as peripheral in certain depressions and in certain organic joys, from which all moral pain and all moral pleasure seem absent —or as purely cerebral, as in certain joys and certain sadnesses we have just mentioned» (my tr.).

[37] «The greatest mistake one can make when presenting a theory would be to consider that theory as sufficiently explanatory» (my tr.).

the best of the different theories of emotions available in his time (see also [Dibattista 2015; 2021]).

In this respect, even if strongly criticizing *intellectualist* approaches to emotions —that viewed emotions as only dependent on the existence of opposing or converging representations and ideas in the mind, with bodily changes being just accidental byproducts of them ([Herbart 1825; Nahlowsky 1884], partly [Wundt 1874])— Dumas [1895; 1922; 1948] did not underestimate the role that *cognition* plays in emotional processes. Emotional phenomena, of course, cannot be reduced to intellectual dynamics, nor ideas fully determine emotional occurrences. Yet, at least in some cases, cognition (attention, interpretation, representation, memory, etc.) contributes to emotional experience, influencing its emergence and evolution. This may be the case, for instance, of passive fear —where the confused representation of an imminent danger can foster the emergence of this emotional state, also by indirectly producing some of the bodily changes connected to it [Dumas 1948, 170-171].

In his syncretic take on emotions, Dumas [1932; 1948] also considered *behavioral* aspects of them, especially those related to external and *voluntary* emotional *expressions*, i.e., to *emotional mimicry*. As known from above, the bodily component of emotions involves spontaneous (involuntary and direct) bodily changes dependent on the autonomic nervous system. Along with voice alteration, hair raising, blushing, etc. the contraction of the *facial* muscles amounts to one of the most expressive traits of emotional responses —sadness, for example, involves (involuntary) lowering the mouth's corners and drooping the forehead (see also [Section 3.3]). There are cases in which it is exactly the facial expression what even helps to qualitatively differentiate between otherwise physiologically and cognitively equivalent emotional states (e.g., fear and surprise). Besides involuntary responses, however, facial expressions can be, at least in part, voluntarily modulated in order to regulate, amplify, or even mask and purposely evoke an emotion basing on context. Mimicking the act of smiling, for example, can reinforce the feeling of joy, demonstrating the bidirectional influence between expression and emotional experience[38] [Dumas 1906; 1948].

[38]In this, following Darwin [1872] and [James 1890], Dumas seems to anticipate the so-called *Facial Feedback Hypothesis*, according to which one's facial expression not

For Dumas, in other words, the subject's behavior —especially mediated by facial mimicry— is constitutive of an emotional state. Furthermore, the focus on the behavioral components of emotions naturally led him to consider them as partly being the result of *evolutionary*, on the one hand, and *sociocultural processes*, on the other.

Dumas repeatedly expressed skepticism towards Darwin's [1872] view of emotions [Section 3.2.1.2], that framed them *primarily* as inherited biological instances selected by nature because of their adaptive value for survival. This view, he pointed out (see e.g., [Dumas and Malloizel 1910; Dumas 1922; 1948], etc.) often overlooked the role of basic physiological mechanisms in the generation of emotions (e.g., visceral reflexes, adrenaline regulation, activation of sympathetic and parasympathetic systems), that he deemed to be *innate* and possibly independent of natural selection. Nevertheless, Dumas did not entirely denied that emotions and their (especially voluntary) expressions may be the result of evolutionary mechanisms, arguing that certain expressions may have been favored over others because of their social and communicative significance.

Finally, whereas emotions and their voluntary expressions can, at least in part, be shaped by evolutionary mechanisms, then, argued Dumas, they are also subjected to social and cultural conditionings (see also [Section 3.3]). Expressions, in fact are learned since childhood through the interaction with other individuals. People born blind, for example, never learn to correctly reproduce all the fine-tuned facial expressions connected to the different emotions, as they lack direct visual exposure to others' emotional displays [Dumas 1932] —for this, they have namely to rely on other senses but sights (auditory or tactile experience in particular). Moreover, the way emotions are socially learned is itself a cultural construct [Dumas 1948, 399ff.] Just as languages, emotional expressions change across different cultures. They are stylized and shaped by etiquettes, traditions and artistic representations, which impose aesthetic and behavioral norms on how emotions should be displayed, learned, and interpreted within a given society, in a given historical period.

only determines the display of a given emotion, but also backwardly influences their emotional state itself (see e.g., [McIntosh 1996; Kravljaca and Stanley 2024], etc.).

Despite the theoretical fruitfulness of Dumas' approach to emotions, his work has been largely neglected in 20th-century philosophical and scientific debate about emotions. Interestingly, with his syncretic characterization of emotions —encompassing physiology, cognition, evolutionism, sociology and culture— he anticipated many of the intuitions that will be at the core of the neurophysiology of emotions, the *Emotional Turn*, the emotional constructivism and the enactivist conception of the mind of the 1980s-2000s[39] [Section 3.3].

Until such great developments in the study of affectivity, however, the different scientific approaches that, from the medical tradition of humoral theory to the physiological studies of the 19th century, had been developing in the attempt to highlight emotions as deeply embodied processes, needed to grow into more structured theories, grounded in experimental research and solid conceptual frameworks. This happened starting from the 1940s, with physiology and neurology methodically exploring the mechanisms through which emotions emerge from bodily processes.

A significant contribution to the deepening and systematization of earlier physiological theories of emotions primarily came from the research about the neuroanatomy of the brain. New studies, in fact, were trying to abolish the dualism introduced by JLTE and CBTE about the role of, respectively, cortical and subcortical regions in determining emotional occurrences. Instead, more integrated neuroanatomical models for explaining emotions generation were proposed.

3.2.1.6 The Physiological Approach between the 1940s and the 1980s

Particularly important, in this respect, was the description of the so-called *Papez circuit* by the American neuroanatomist James Papez [1937]. As already criticized by Dumas [1922], one of the major issues of CBTE was its difficulty to account for the rich qualitative dimension of emotional experience. Not only derived this from denying the role of the

[39] This is maybe due to the fact that, until today, almost anything of Dumas' scientific production has been translated into English, which prevented him to be influential on the subsequent debate on emotions which has taken place mostly within the Anglo-Saxon academic world.

peripheral system in the emotion production process, but also from neglecting the function of the cortex in the qualitative determination of emotional states (see also [Leventhal 1980, 146]). Building upon previous clinical observations and experiments in neurology and comparative anatomy, Papez [1937] eventually reintroduced the cortex into the neuroanatomical explanation of emotions. Yet, according to his model, emotions do not arise from the activation of just one isolated brain region, be that cortical or subcortical. On the contrary, they are produced as a result of the activation of an *integrated circuit* that involves the stimulus being received through the sense organs and transferred to the thalamus via the central nervous system. From the thalamus, sensory information is then propagated to the cortex, that organizes and recognize the stimulus. From the cortex, a signal is sent to the hypothalamus through the hippocampus, that compares sensory information with memory contributing to its emotive evaluation. The hypothalamus activates a bodily response consistent with the signal, transferred to the periphery via the central nervous system. At the same time, it sends the signal back to the cortex through the thalamus, with the former finally transforming it in a conscious, qualitative emotional experience.

The *Papez circuit* thus explained emotions as an integrated phenomenon involving a physiological and qualitative component, emerging from the interplay of both peripheral and central (cortical and subcortical) systems. Right after its appearance, the model received some experimental support from lesion, electric stimulation, or anatomical tracing studies in animals (dogs, cats, monkeys) and humans (see e.g., [Klüver and Bucy 1939; Spiegel et al. 1940; Smith 1944; 1945; Kremer 1947; Bard and Mountcastle 1948; Sachs and Brendler 1948; Ward 1948], etc.). However, it was from the attempt to confirm it empirically that its limitations were revealed. Papez' neuroanatomical model was namely too narrow for explaining how qualitatively different emotions could arise just from a single pathway. Structures, like the hippocampus, that in it were depicted as primarily involved in emotion, turned out later to be fundamental rather to other faculties, such as memory. Moreover, the circuit disregarded several brain regions that were later found to play a central role in generating and regulating emotions.

For these reasons, scientists sought to improve the model, whose explanatory power seemed, after all, still unquestionable in contrast to other previous theoretical approaches to emotions. One notable extension of Papez' ideas was proposed by the American neuroscientist Paul D. MacLean on the basis of new neuroanatomical evidence and comparative studies. MacLean [1949] highlighted the role of the amygdala and the septum, beyond the hypothalamus, in producing emotional experience. In particular, he stressed that the *amygdala* is responsible for mediating the influence of bodily states on emotional phenomena. The amygdala, in fact, receives both exteroceptive information (from the sense organs) and interoceptive stimuli (such as hunger, thirst, pain, sexual arousal and other autonomic responses). It organizes them by associations and integrations with past experience and body-based emotional memory, and then maps them onto specific emotional states. Through its strong connection with the hypothalamus, the amygdala also determines the emergence of visceral, muscular and (instinctual) behavioral changes connected to those states, such as heart rate, respiration, digestion, etc.

In contrast to Papez' circuit, therefore, MacLean's [1949] emotional brain involved a more variegated set of neuroanatomical structures held as relevant in controlling affective phenomena. To this set, he gave the name of *visceral brain* (see also [Jakob 1906]), indicating an evolutionary older portion of the brain responsible for acting as a neurophysiological interface between internal (visceral) bodily signals, autonomic responses (to external stimuli) and emotional subjective feelings —a concept that MacLean [1952; 1955] will later develop into the *limbic system* with some conceptual refinements.

Beyond the neuroanatomy of the brain, the empirical robustness of the physiological approach to emotions was further reinforced by (closely related) research on the *autonomic nervous system* (ANS) —the network of neural pathways responsible for regulating involuntary bodily functions such as heart rate, respiration, and digestion— and its role in giving rise to emotional episodes. In particular, physiologists and psychologists were engaged in addressing one of the most challenging aporias of theories like JLTE and CBTE: the underdetermination of neurophysiological mechanisms with respect to the qualitative variety of emotional states.

To questions like whether different emotional states can be distinguished by specific patterns of autonomic activation, and how these bodily responses contribute to the generation and modulation of affect, relevant experimental answers were given starting from the 1950s. The American psychophysiologist Ax [1953], for example, showed that fear and anger are tied to different ANS activation dynamics, diverging from Cannon's [1915] "fight-or-flight" tenet[40]. By recording the autonomic responses in human subjects after experimental induction of fear and anger, in fact, he noticed that, on the one hand, anger was associated with greater diastolic blood pressure rises, heart rate decreases, and muscle tension, resembling a combined sympathetic-parasympathetic activation. Fear, on the other hand, led to higher skin conductance (sweating), faster respiration, and more frequent muscle tension peaks, aligning more closely with pure sympathetic arousal. His results suggested that the two emotions, while both involving heightened arousal, displayed different patterns of physiological activation —particularly in terms of vascular resistance and heart rate. These findings appeared to lend support to the idea that emotions might, at least to some extent, be physiologically distinguishable.

Along ANS and brain structures, other research questions moreover concerned the *endocrinology* of emotions. Emotions, in fact, are not only regulated through neurological mechanisms, but also depend on the biochemical activity of the *endocrine system*, i.e., the system responsible for the production and the release of *hormones*. Selye [1956], for example, observed hormonal responses in rats exposed to various physical and psychological stressors (e.g., injury, fear, etc.). Adrenaline, cortisol and other hormones are produced by the body as a reaction to emotionally meaningful events, along with autonomic activation. Similar considerations were experimentally made about *neurotransmitters* —especially *biogenic amines* like norepinephrine, dopamine, serotonin, etc.— whose imbalance was found to directly influence (negative) emotional states (see e.g., [Schildkraut and Kety 1967; Kety 1970; 1972; Schildkraut 1974]; etc.).

[40]This tenet asserts that threats trigger uniform *sympathetic* responses (such as, for example, increased heart rate, adrenaline release, etc.) to prepare the body for either confrontation (fight) or escape (flight), with no physiological distinction, thus, between fear and anger.

From the 1970s onwards, the growing biochemical perspective on emotions was enriched by insights coming from another discipline which was developing at the time, that is, *neuroscience*. The study of the brain and of the nervous system was traditionally conducted within different and heterogeneous disciplines (neuroanatomy, neurophysiology, behavioral psychology, biology, etc.). During the 1960s-1970s, however, new experimental techniques (electron microscopy, electrophysiology, intracellular recording, computer-assisted tomography, radioimmunoassay, etc.), new conceptual frameworks and the formal creation of dedicated societies, departments and journals transformed neuroscience into an autonomous research field (see e.g., [Bickle 2009; Glickstein 2017]). Beyond concentrating on traditional research topics like perception, memory, attention, cognition, etc., neuroscientist soon began to deal with emotions, trying to produce a more detailed mapping of the neural circuits underlying emotional processing. Researchers began to investigate how discrete brain structures and specific neurotransmitters interact to generate, modulate, and regulate emotional occurrences. This eventually led to the rise of *affective neuroscience*, seeking to explain emotions through an integrated understanding of brain anatomy, neurophysiology, and neurochemistry (see e.g., [Panksepp 1982; LeDoux 1984; LeDoux and Hirst 1986; LeDoux 1986; Panksepp 1988; 1989a,b; 1998] and [Davidson and Sutton 1995; Davidson 1998]).

3.2.2 The New Paradigm of Embodied Cognition

Hence, the period between the 1960s and the 1980s saw emotions being increasingly understood as *embodied* phenomena, especially rooted in neurophysiological mechanisms. At the same time, parallel reflections were taking place in psychology about the role played by the body in human cognition in general.

As pointed out in [Section 1.1.1], behaviorism's denial of the possibility to study the mind from within was rejected by a new, *revolutionary*, approach to psychology that emerged in the 1950s and 1960s: *cognitivism*. Deeply influenced by —and in turn deeply influencing— AI at its earliest stages, cognitivism regained a scientific perspective on internal mental processes by characterizing them essentially as *computer programs*. In

this view, the mind is conceived as a *computer* that performs formal operations on *representational symbols* according to well-defined logical *rules* (see e.g. [Newell and Simon 1961, 2012] and [Fodor 1975]). Studying psychological phenomena —including emotions— thus boils down to retrieving and describing the programs that govern the mind's computational mechanisms. These programs run principally in the brain, the true *locus* of cognition. They start with inputs to it, and finish with outputs from it, so that the mechanisms of cognition are completely independent from what is outside the cranium, including the body and the environment. Like all computational processes, in other words, mental processes as well are *solipsistic* (see e.g., [Fodor 1980; Stich 1983]).

Since the cognitive revolution, the computational conception of the mind has become the central tenet of a robust scientific paradigm that still today continues to guide a significant part of the research in psychology and AI. However, alternative perspectives began to emerge early on, driven by a fundamental dissatisfaction with the solipsistic assumptions underlying the computational paradigm (see e.g., [Shapiro 2019]). These critiques gained further traction also in light of the early limitations and technical failures of GOFAI, as well as internal debates within cognitivism itself —especially concerning emotions— which challenged the (epistemic) adequacy of reducing mental processes to algorithmic manipulations of symbols (see [Section 1.2]). In this respect, a valuable alternative to computational-representational cognitivism was proposed by the American psychologist James J. Gibson [1966; 1979], expert of vision and father of the so-called *ecological theory of perception*.

Indeed, perception was among the mental faculties to be studied even by early cognitivists, of course within a computational framework (see e.g., [Broadbent 1958; Neisser 1967; Gregory 1972]). On this view, it was seen not as a passive reception of sensory stimuli, but as an active problem-solving process. Sensory stimuli alone, in fact, were deemed to be too ambiguous and informationally incomplete for producing coherent perceptions. To overcome this underdetermination, the brain was therefore thought to engage in a series of internal computational operations transforming raw sensory inputs into structured mental representations of the external world. The process started and ended with symbolic descriptions of the physical world through a transition between different

computational states governed by algorithmic rules of inference encoding general assumptions about the world and constraining the set of possible correct outputs. In vision (see e.g., [Gregory 1972; Marr 1976; 1982; Rock 1975; 1983]), for example, these operations were considered firstly to involve symbolic encodings of different features of the sensory stimuli received through the retina —e.g., edges, shapes, spatial relations, colors, depths, etc. Such symbolic representations are then manipulated according to algorithmic rules containing prior assumptions about physical regularities in the world —e.g., that light typically comes from above, that objects are rigid and continuous, etc. Explaining perception thus meant for early cognitivists simply to uncover and formalize the algorithms and representational structures that enable such operations [Shapiro 2019, 21–28].

Gibson [1979], however, was of a different opinion. He contrasted the cognitivist idea that sensory inputs are informationally impoverished. On the contrary, according to his ecological theory, environment provides *structured* stimuli that involve *rich* sets of features, directly detected by the perceiver, without the need for internal reconstruction or symbolic mediation. As the perceiver modifies their spatial relationship with the environment (e.g., walking toward an object, turning the head, changing gaze), some features of the sensory array remain stable, while other vary following regular and predictable patterns emerging from its physical properties and the dynamics of the perceiver's interaction with it. Perception, in particular, derives from *resonating* to the invariant features of the environment, without the need of internal computations. Since detecting *invariants* requires an *ecological*, spatial interaction of the perceiver with the environment, perception is not a solipsist process taking place only in the brain. On the contrary, it involves all the bodily parts, and their activities, which participate in this process.

Vision, for example, does not begin with ambiguous retinal images requiring inferential processing, but with richly structured optical information supplied by the environment. As the observer moves, ambient optic array changes in systematic ways, while certain aspects remain invariant —for instance, although the shape projected onto the retina by a tabletop may appear trapezoidal from different angles, the relationships among its edges and angles stay constant and directly specify the rect-

angular surface. Through continuous transformations occurring in the perceiver's active bodily engagement with the environment, the visual system recognizes visual invariants and resonates to them without mediation of symbolic computations and internal image construction. Vision, therefore, is not a passive reception of stimuli within the brain, but an embodied and *ecological* activity distributed across the sensory organs, the head, and the movements that shape how visual information is picked up.

Besides Gibson's theory, another approach that contrasted the classical computational view of the mind was Gary C. Hatfield's [1988] *noncognitive computationalism*. Like Gibson, Hatfield as well endorsed an *ecological* view of the mind, with mental processes being grounded in the organism's direct engagement with its environment. However, this did not amount to a complete negation of the explanatory relevance of internal mental processes, that Gibson's [1979] ecological psychology effectively excluded, thereby reintroducing a new form of methodological behaviorism [Fodor and Pylyshyn 1981; Hatfield 1988]. This aligned Hatfield more closely to classical cognitivism, that emphasized the role of internal psychological mechanisms in mediating between inputs into outputs. Unlike Gibson, indeed, he did non reject thought as a computational process in bulk. However, differently from classical cognitivism, Hatfield's [1988] noncognitive computationalism denied that mental computations are inferential manipulations of symbols.

Rather, under this view, mental processes are characterized as *functional algorithms* operating on *non-symbolic representations*. In order to clarify this, consider, for example, a spring scale with a pan. When a certain mass is placed on the pan, the spring compresses, and a pointer moves along a graduated dial displaying its weight. The internal state of the system (degree of compression of the spring) carries information about the input (mass). This information, however, is not represented in the form of a symbolic structure, but as a *functional* correlation with specific features of the environment (relationship between mass input and pointer position). Also, the transformation of the input into an output (displayed weight) performed by the system is accordingly not a result of computing numerical data, following syntactic rules, manipulating symbols or performing logical inferences, but just a systematic, lawful, re-

producible, i.e., *algorithmic* state transformation grounded in the causal structure of the system and the environment (physical response of the spring to an applied force).

According to Hatfield [1988], the mind works like such a system: External inputs are mapped to internal outputs (and behaviors) via functional algorithms; mental states represent features of the environment not by means of syntactic structures, but by virtue of their functional role within the cognitive system and its environment. In color perception, for example, the visual system is receptive to informationally rich stimuli: As light reaches the eyes, it carries information specified by the structure of the environment —such as wavelength distributions that systematically vary with surface properties and illumination. The cones in the retina, each sensitive to different ranges of wavelength, respond accordingly by firing at rates that reflect this variation. This information, however, is not represented symbolically, but rather, as in the case of the spring scale, functionally: The internal state of the visual system bears a reliable, law-governed correlation to those features of the external environment. In transforming, through the cones, incoming wavelength patterns into differentiated activation profiles, and converting these latter into a stable experience of a given color, the visual system realizes a functional algorithm: The process is systematic, lawful and reproducible, but it does not involve the manipulation of internal symbols or explicit representations. Instead, it relies on the causal structure of the visual system and its sensitivity to environmental regularities.

Functional algorithms, Hatfield points out, lie however *above* the level of neurophysiology. In other words, they are independent from particular anatomical or physiological mechanisms: What matters is not the biological substrate that carries out the operation, but the abstract structure of the operation itself —the way the system reliably transforms certain inputs into outputs. In this sense, as in classical functionalism (see e.g., [Putnam 1960]), the same functional algorithm could, in principle, be instantiated in different neural architectures, or even in artificial systems, as long as the causal and computational roles are preserved.

This kind of *ecological psychophysics* [Hatfield 1988, 185], describing the relations between stimulus and response as lawful, but non-symbolic, regularities (see e.g., [Fechner 1860] and [Section 3.2.1.1]), yet emphasiz-

ing the organism's functional capacity to resonate to structured environmental features [Gibson 1979], found a compelling expression in connectionism and neural-networks-based models of the mind (see [Smolensky 1988] and [Section 3.3.1]). As described earlier, in these models information is represented through the distributed activation of simple, interconnected units that enables the system to track complex patterns and invariants in the input without relying on symbolic rules, which is precisely the kind of non-symbolic, functional transformation Hatfield describes: Perception and cognition arise not from internal manipulation of symbols, but from the system's dynamic sensitivity to the structured regularities of its environment.

Albeit from different perspectives, hence, both Gibson [1979] and Hatfield [1988] opposed the classical cognitivist view of the mind as a *solipsistic* computational system essentially disentangled from the body and the environment. And, in fact, exactly *ecological psychology* and *noncognitive computationalism* inspired many of the research projects that, from the mid-1980s onwards, converged into a new —and still today developing— scientific paradigm within the study of the mind: *Embodied Cognition* (EC) —see also [Shapiro 2019; Shapiro and Spaulding 2024].

Indeed, the label EC indicates no single, well-defined and coherent theory of the mind. Rather, it encloses a heterogeneous set of scientific approaches that, while differing in theoretical commitments, subject matters and methodological practices, together share the conviction that cognition is not confined to brain-related symbolic computations, but actively involves the body and the way it interacts with its physical, social and cultural environment.

The Chilean biologist and philosopher Francisco J. Varela, along with the American philosopher Evan Thompson and the American psychologist Eleanor Rosch [1991] (VTR), were among the first scholars to systematize this conviction into a structured philosophical proposal. From a complete rejection of both metaphysical realism (the cognitive system recovers, symbolically or subsymbolically, *preexistent* properties of the world) and idealism (the "world" is merely an apparent *projection* of the mind), they relinquish the notion of representation altogether, being this the medium to «recover what is outer» or «to project what is inner» [172]. Cognition, in fact, does not consist in formally computed

representations retrieved from, or imposed to, the external world, but —they claim— in a form of *embodied action*: *embodied* in that it «depends upon the kinds of experience that come from having a body with various sensorimotor capacities», whereby «these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context»; and *action* in that «sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition» [172-173].

According to VTR, in other words, cognition —*ecologically* [Gibson 1979]— arises from a continuous interplay between an organism's sensorimotor capacities and its environment. Perception is a *dynamic* process in which sensory inputs and motor activity are tightly interdependent: The organism's movement within the environment iteratively reshapes its perceptual space, which in turn redefines the range of possible actions it can perform in it, thereby influencing subsequent interactions. Perceptions determine actions, which determine new perceptions, which determine new actions, and so on. Body and environment are *inseparable* and, of course, two physically different organisms, with different perceptual systems and sensorimotor capacities, will perceive and cognize the world differently.

Cognition is therefore not a static, monolithic faculty an organism *has*, but something an organism actively *does*. It emerges not from computing representations of the world, but from the constant interaction with the physical environment, in which «(1) perception consists in perceptually guided action and (2) cognitive structures emerge from the recurrent sensorimotor patterns that enable action to be perceptually guided» [Varela et al. 1991, 173]. This is the sense in which VTR talk about cognition as *enaction* —from "to enact", meaning to bring forth or realize through action: Cognitive activity consists in the ongoing enactment of a meaningful world through embodied engagement. The organism does not represent a pre-given reality but brings forth its own domain of significance through its sensorimotor coupling with the environment. In this framework, mind, body, and world are not separate ontological domains, but aspects of a single, co-emergent process of lived cognition.

VTR's enactivism draws upon a long-standing philosophical tradition emphasizing the interdependence of cognition, perception and action

—in particular 20th-century French and German phenomenology (see e.g. [Heidegger 1927, 1955; Husserl 1922, 1929, 1936; Merleau-Ponty 1942, 1945, 1964]) and Buddhist philosophy. These philosophical foundations enabled VTR to rethink the classical cognitivist conception of the mind in terms of *embodiment* (bodily dependence) and *situatedness* (environmental dependence) of cognition. Yet, they still lacked a viable methodological alternative to the discrete operational frameworks and formal tools characteristic of computationalism. Articulating VTR's philosophical proposal into a systematic and operationally tractable conceptual framework, so to make embodied cognition a concrete object of scientific investigation in psychology, was a task primarily embraced by *Dynamical Systems Theory* (DST)[41].

DST mathematically describes *dynamical systems*, i.e., systems that change over time. Aware of the potential of VRT's ideas, some researchers began, around the turn of the millennium, to model the mind as such a systems, using DST for explaining its internal working (see e.g. [Van Gelder 1995, 1998; Beer 2000; Smith and Thelen 1994, 2003; 2003]). Drawing inspiration from early cybernetic thinking —particularly in its attention to feedback loops and the self-regulatory behavior of systems (see [Section 1.1.1] and [Section 5.1.2])— DST reframes cognition not as the manipulation of discrete internal representations, but as a *real-time*, *continuous* process emerging from the system's interactions with the environment. Being such, it can be represented by a system of non-linear, differential equations that describe how cognitive states evolve over continuous time within a closed loop of reciprocal interactions where

[41] Allowing for the formulation of concrete, testable scientific hypotheses based on EC, DST has also helped this paradigm to be corroborated on the empirical level. For example, experiments on dynamic sensorimotor coordination and behavioral pattern formation have demonstrated how cognitive processes like infant motor development or perceptual categorization emerge from the self-organizing interplay between brain, body, and environment, rather than from pre-programmed symbolic computations [Smith and Thelen 1994; Kelso 1999]. Likewise, psychological experiments have provided evidence for embodied and situated influences on language, memory, and action [Glenberg 1997, 2002; Barsalou 1999, 2008; Beilock and Holt 2006, 2008; Proffitt 2006; Richardson and Dale 2005; Wilson and Knoblich 2005]. Moreover, as it will be shown in next [Chapter 5], corroborating evidence has also come from robotics —particularly through experiments like Rodney Brooks' [1997] COG [Section 5.2]— where the embodied nature of cognition has been explored through autonomous agents whose intelligent behavior arises not from internal symbolic processing but from their embodied interactions with the world.

internal dynamics, physiology of perception, environmental inputs and behavior co-determine each other in a *self-organizing* process. In this causal circle, from which what we call "cognition" *emerges*, beginning and end thus coincide: Brain, body and environment cannot be treated as distinct partitions or subsystems of the mind —each requiring its own formalism and vocabulary, nor can the mind be confined solely to the brain, deemed, as in computationalism, to be the privileged *locus* of cognition[42].

3.2.3 Emotions, Cognition and the Body

Hence, beginning from the 1990s, VTR's enactivism, DST and related approaches emphasizing the embodiedness and situatedness of the mind, introduced a solid theoretical and methodological alternative to computationalism, and this not only in the symbolic-representational variant of GOFAI and classical cognitivism, but also in its connectionist one. In fact, despite the departure from explicit rules and symbols, subsymbolic approaches remained anchored to the same cognitivist conception of the mind as essentially *disembodied*. Also, at the turn of the 1980s and 1990s, neural-network-based AI itself was experiencing a period of theoretical and practical stagnation (see [Section 3.1]) that —besides other factors— favored the consolidation of different paradigms.

Among these, influenced by and influencing EC approaches in psychology, even robotics served to corroborate a conception of cognition as embodied and adaptive interaction. This was notably the case of Rodney

[42]Similar critiques to the cognitivist conception of the brain as the privileged *locus* of cognition have been expressed especially by Andy Clark [1998] and David Chalmers [1998] (see also [Menary 2010]), who developed a philosophical theory of mind notoriously known as *Extended Cognition*. According to this theory, the cognitive system can encompass not only neural processes but also bodily actions and external artifacts, insofar as they play a functionally equivalent role in the processing and storage of information. In a famous thought experiment, they argue that if a notebook reliably stores information that a person consults, in the same way another might recall such information from memory, then the notebook functions as a part of that person's cognitive apparatus. The key idea is that cognition is not defined by *where* it happens (e.g., inside the skull), but by *how* it happens —through stable, functionally integrated processes. This directly challenges the internalist assumption of classical computationalism, and aligns, at least programmatically, with VTR's and DST's emphasis on the systemic, relational, and environmentally embedded nature of cognitive activity.

Brooks' *Nouvelle AI* (see e.g., [Brooks 1991c], and [Section 5.2], [Footnote 41], and [Shapiro 2019]), that introduced the so-called *subsumption architecture* [Section 5.2.2] as an alternative to both representationalist and connectionist AI. Concretely exemplified in projects like the robot COG [Section 5.2.3], it showed that AI could be grounded in layered, sensorimotor behaviors rather than in representations and abstract rules of symbolic manipulation[43], contributing to broadening the conceptual landscape from which, later, AC itself could emerge (see e.g., [Picard 1997, 71ff.]).

Moreover, as seen in [Section 1.2], critiques to the *computational metaphor* had been coming, already since the 1960s, from growing reflections about the role that emotions play in intelligence. Emotions were increasingly understood to be not obstacles to rational thinking, but rather modulators and organizers of core cognitive functions such as attention, memory, perception, learning and decision-making. As a result, they began to be even explicitly integrated in computational models of cognition (see [Chapter 2]).

At the same time, emotions themselves had long been understood as fundamentally inseparable from bodily processes. From the JLT to CBT and later physiological theories of emotions, a persistent line of research emphasized their connection to the muscular, visceral, nervous and endocrine activity of the body. In the 1980s, the physiological dimension of emotions eventually began to be accounted for even in artificial systems. The case of FEELER, for instance, shows how models began to operationalize bodily responses as a part of artificial emotions generation [Section 2.3].

[43]Some philosophical literature (see e.g., [Datteri 2008]), however, has claimed that the novelty of Brooks' approach is indeed often overstated. In particular, while [Brooks 1991c] famously argued for an "intelligence without representations", critics have pointed out that behavior-based architectures such as the subsumption architecture still presuppose minimal internal structures that function as or through representations —for instance, transient sensory memories, state variables, or coordination parameters mediating between perceptual inputs and motor outputs. Such internal encodings, though not symbolic in the classical sense, nonetheless play a representational role in guiding action. Moreover, the absence of explicit planning mechanisms limits the scalability of purely reactive systems, which tend to perform efficiently only within narrowly defined sensorimotor domains. Hence, Brooks' *Nouvelle AI*, rather than abolishing representation altogether, is better understood as relocating it within embodied and distributed control processes —see also [Section 5.2].

The development of EC, thus, constituted the fortunate occasion where all these research lines —showing the connections between cognition and emotions, cognition and the body, and emotions and the body— could come together and converge into a view of the mind as an integrated process, where cognition, emotions, and the body are not separate interacting components, but inseparable dimensions of a *unified* system [Figure 3.1].

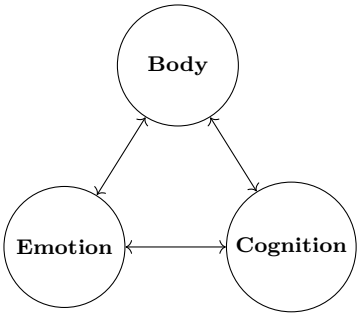


Figure 3.1: Emotion, cognition, and the body are mutually interdependent: Emotions involve cognition, cognition is embodied, and emotions are grounded in bodily states.

Before EC, some attempts to hold emotions, body and cognition together had been already made. A notable example is Schachter and Singer's [1962] *two-factor* theory of emotions (see [Section 2.3] and [Section 2.4.1]), where affective phenomena have both in physiological changes and cognitive appraisals their necessary, and alone not sufficient, conditions. This theory anticipated EC by recognizing that emotions cannot be reduced either to mere bodily states or to disembodied cognitive evaluations, but emerge from their integration.

Yet, differently from EC, in theories like these, the integration of emotion, cognition, and the body remained partial and structurally limited. For instance, they still relied on *modular* and *sequential* architectures: Physiological arousal and cognitive appraisal were treated as distinct phases in a linear process, rather than as co-emergent and dynamically interdependent aspects of emotional experience. Also, a *residual dualism* always persisted: Although body and mind were brought into relation, they remained theoretically separate components of emotional phenomena. Relatedly, the cognitive dimension in two-factor models was

still conceptualized in *disembodied* terms, as abstract, symbolic, linguistic and purely evaluative. Emotions were said to arise when a subject interprets a state of arousal in light of contextual cues, but this interpretation was imagined as a *detached* mental operation, rather than a sensorimotor and affectively charged engagement with the environment. Consequently, the role of situatedness in mental processes was virtually absent: The environment was merely considered as a backdrop for cognitive interpretation; It was not an active element shaping the emotional episode of the subject through real-time interactions. Finally, these classical models lacked a *dynamical-systems* perspective: Their explanatory architecture was static, grounded in fixed components and causal chains, rather than in ongoing, temporally extended patterns of self-organization and regulation across brain, body, and the world. From the standpoint of EC, in other words, they fell short of capturing the deeply interactive, non-linear nature of lived emotional experience.

Such limitations also affected other theories of emotions developed later especially within philosophy, psychology and cognitive science. This is for example the case of Nico Frijda's [1986] theory of *action tendency* — for which emotions are elicited by the appraisal of events in terms of their relevance to the individual's goals, viz. defined by the motivational states they generate as tendencies to act in certain ways to adaptively respond to those events; Or the case of OCC model of emotion (from the name of its authors Andrew Ortony, Gerald L. Clore, and Allan Collins [1988]), for which emotions arise from structured cognitive evaluations of events, agentive actions, or objects in terms of how they respectively relate to an individual's goals, standards, and tastes. Such theories belonged to the same *cognitivist* tradition of Schachter and Singer [1962], contributing to changing the view on emotions from irrational disturbances to integral components of intelligent behavior. They emphasized the cognitive origin of emotions, as well as their functional role in intelligence. However, they significantly differed from the two-factor model because of an even stronger minimization of the role of the body in emotions — often asserted under a general neglect of the neurophysiological evidence produced about it. These approaches, in fact, still operated within a *disembodied* and *representational* conception of the mind: The appraisal processes at the core of these models too were typically conceived as

abstract, internal, and linguistic operations, largely independent from bodily states and sensorimotor dynamics. Moreover, the environment was treated as well as a source of input or context, rather than as a constitutive component of emotional episodes.

Thus, although cognitivist theories of emotion certainly represented an important step forward toward a better understanding of emotions as involved in, and originating from, cognitive processes, they remained essentially rooted in conceptual distinctions and limitations that only the embodied and enactive paradigm would later try to dissolve.

Only in the 1990s, in fact, a number of different research approaches emerged beginning to converge toward more integrated views in which cognition, emotion, and bodily processes are seen as mutually influencing elements within a more or less unified, dynamical system. In this regard, a telling example is the proposal of the Portuguese neuroscientist Antonio R. Damasio [1994; 1999; 2003], that directly challenged the Cartesian mind-body dualism [Section 1.2] by conceiving of the mind not as an abstract intellectual faculty, but as a process emerging from the continuous interaction between brain and body: Mental states are grounded in neural mappings of bodily states, so that the body constitutes the indispensable reference for the operations we identify as cognitive. In Damasio's view (recalling, for some aspects, JLT as well as Spinoza [1677]), the body is also what grounds emotions: They are patterns of physiological change triggered and regulated by neural circuits within the ongoing self-regulatory activity of the organism. From this, it directly follows that cognitive functions, such as reasoning and decision-making, and emotions are deeply interconnected: Emotions, in fact, provide evaluative and motivational input necessary for cognitively navigating complex environments. This is an idea that Damasio formalized, particularly, in the so-called *Somatic Marker Hypothesis* (SMH), positing that (intrinsically) emotionally laden bodily signals (associated with prior experiences and recalled in comparable situations) actively intervene in the management of cognitive processes, for instance, by constraining decisional spaces and guiding deliberations. In this sense, an "emotionless" cognition is impossible: We can *cognize* as we do (e.g., efficiently solving problems or promptly taking decisions) because we *emote*, and we can emote as a direct result of the only fact that we have a body.

Neuroscientific approaches like that of Damasio (see also other similar theories, e.g., [Rolls 1990; LeDoux 1998; Panksepp 1998], etc.) moved beyond the linear cognitivist theories of affect by emphasizing the interactivity and interdependency of emotions, bodily arousal, and cognition. While they still treated these dimensions as functionally separable, and emotions themselves as mostly brain-based, they nonetheless paved the way to more *integrated* views of affects as inherently cognitive, embodied, and situated from the outset, as processes that unfold through the organism's continuous sensorimotor engagement with a world that is not merely perceived, evaluated, or responded to, but *enacted* —views that would be soon articulated and explicitly formalized within EC itself[44].

3.3 New Perspectives on Emotions: Emotions as Social and Cultural Objects

As seen in the previous [Section 3.2.1], the study of emotions as fundamentally embodied has long been central to scientific inquiry. Between the 19th and 20th century, a solid scientific paradigm gradually developed in medicine and psychology investigating emotions not merely as abstract mental states, but as phenomena involving physiological processes linked to the brain, as well as to the muscular, visceral, nervous, and endocrine activity of the body.

The unquestionable fruitfulness of the biological (physiological and neural) approach to the study of emotions —further enhanced, toward the end of the 20th century, by techniques such as CT (Computed Tomography), PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), fMRI (Functional Magnetic Resonance Imaging), etc. applied to the brain for investigating affective occurrences in real time (see e.g., [Pardo et al. 1993; George et al. 1993; 1994], etc.)— was, however, also accompanied by a drift toward largely reductionist views. These had been emerging already since the 1970s, especially boosted by the studies of the American psychologist Paul Ek-

[44]Indeed, philosophical and psychological approaches *explicitly* extending EC to the field of affective science, beyond neuroscience, will appear only later (see e.g. [Halberstadt et al. 2009; Niedenthal et al. 2009] in experimental psychology, and [Colombetti 2014; 2017] in philosophy).

man[45], pioneer of the modern *Basic Emotions Theory* (BET) and of the related *Facial Action Coding System* (FACS) [Ekman et al. 1972; Ekman and Friesen 1975; 1978; Ekman 1985; 2003a].

Already discussed by Descartes [1649], James [1890], Wundt [1896], etc., BET posits that emotions are essentially *internal*, *discrete* and *circumscribed* events during which certain feelings are experienced and, at the same time, certain physiological and behavioral changes are expressed. Some of these occurrences are *basic*: They are rooted in evolutionarily shaped neurophysiological mechanisms that enable organisms to respond promptly and effectively to fundamental life tasks (e.g. avoid dangers, securing resources, protecting offspring, etc.) —see [Darwin 1872] and [Section 3.2.1.2]. These mechanisms, called *affect programs* [Tomkins 1962; Ekman 1980], consist of genetically determined instructions that, once triggered, initiate a cascade of coordinated changes across neural activity, physiological state, expressive behavior (facial, vocal, and bodily), and action tendencies. Each program governs a distinct pattern of response, characterized by its own physiological signatures and expressive configurations, and serves both adaptive regulation and social communicative functions.

The reductionist assumptions of BET and related theories also imply a form of *emotional universalism*. As such affective programs, in fact, basic emotions are considered to be innate and analogous in all human populations: A group of at least[46] *six simple emotions* (fear, anger, happiness, sadness, disgust, and surprise) has been namely observed in a wide range of even very distant cultures (United States, Brazil, Japan, New Guinea, Borneo, etc.) [Izard 1968; 1969; Ekman et al. 1969; Ekman and Friesen 1971]. In this context, facial expression are particularly important as primary observable correlates of affective states. These expressions are held to be universal, automatic, and reliably associated with specific basic emotions, regardless of cultural background or social

[45]However, it is to point out that Ekman himself —especially as a result of the criticisms emerged against BET [Footnote 49]— has later argued that all emotions may be considered as basic, and that some are mere combinations of others, thereby distancing his position from strict formulations of BET [Ekman 1992; 1994; 1999; 2003b; Ekman and Cordaro 2011].

[46]Other authors, in fact, have identified a different numbers of basic emotions —see e.g., [Plutchik 1980] (8 emotions).

conditioning. This association can be even formally *codified*: FACS indeed provides a detailed, anatomically based taxonomy of facial muscle movements that can be objectively coded and correlated with discrete emotional states[47].

BET is still today an influential framework for the study of emotions in a great number of disciplines —computer science and robotics included[48] (see [Section 4.2], [Section 5.3], [Section 6.1]). Yet, it is still a controversial theory[49]. In this respect, its reductionist and universalistic implications about emotions have been criticized already by the psychology of the 1970s-1980s.

[47] In particular, FACS categorizes all visible facial muscle activations into components called *Action Units* (AUs), each corresponding to the contraction of specific facial muscles or muscle groups —for example, AU1 corresponds to inner brow raiser (*frontalis* muscle-*pars medialis*), AU7 to lid tightener (*orbicularis oculi-pars orbitalis*), AU51 to head turn left, etc. These units are manually coded by observing video recordings or images frame by frame. Under suitable interpretive frameworks, combinations of AUs can be empirically linked to specific emotional expressions — for example, AU6+AU12 (cheek raiser+lip corner puller) is reliably associated with a genuine (Duchenne) smile, AU4+AU7+AU23 (brow lowerer+lid tightener+lip tightener) with anger, etc. While FACS offers an objective method for analyzing facial muscle movements, supporting automatic emotion recognition [Section 4.1.1] it struggles with capturing spontaneous or subtle expressions —especially in individuals with particular facial morphologies (e.g., children, elderly, obese, impaired). Moreover, it is influenced by cultural variability and it is not capable to infer real emotional states beyond their outer expressions.

[48] FACS —as an operationalized version of BET— plays a significant role in affective computing, robotics, and computer science, particularly in tasks involving the artificial expression and recognition of emotions. Nevertheless, its universalist and reductionist assumptions continue to raise critical epistemological concerns in the field (see [Section 4.3] and [Chapter 6]).

[49] A great number of criticism has been namely formulated against BETs. For example, the cross-cultural variation in emotion concepts and language challenges the assumption that basic emotions are biologically fixed and universally experienced in the same way [Russell 1991; Wierzbicka 1999]. Moreover, early cross-cultural studies presenting participants with photographs of posed facial expressions and requiring selection from a limited set of emotion labels have been criticized for methodological bias, as the forced-choice format thereby used likely inflated recognition rates and overstated cross-cultural agreement [Russell 1994]. Also, BET has been accused of ignoring the variability in emotional expression across social contexts [Russell 2003; Barrett 2006]. Furthermore, stable and univocal correlations between basic emotions and patterns of neurophysiological activation are difficult to observe [Cacioppo et al. 2000; Phan et al. 2002; Murphy et al. 2003; Barrett 2006]. Also, BETs often discern basic from non-basic emotions arbitrarily, circularly and unjustifiedly, disregarding the phenomenologically complex, dynamic, embodied, and context-sensitive nature of affective phenomena [Colombetti 2017; 2014].

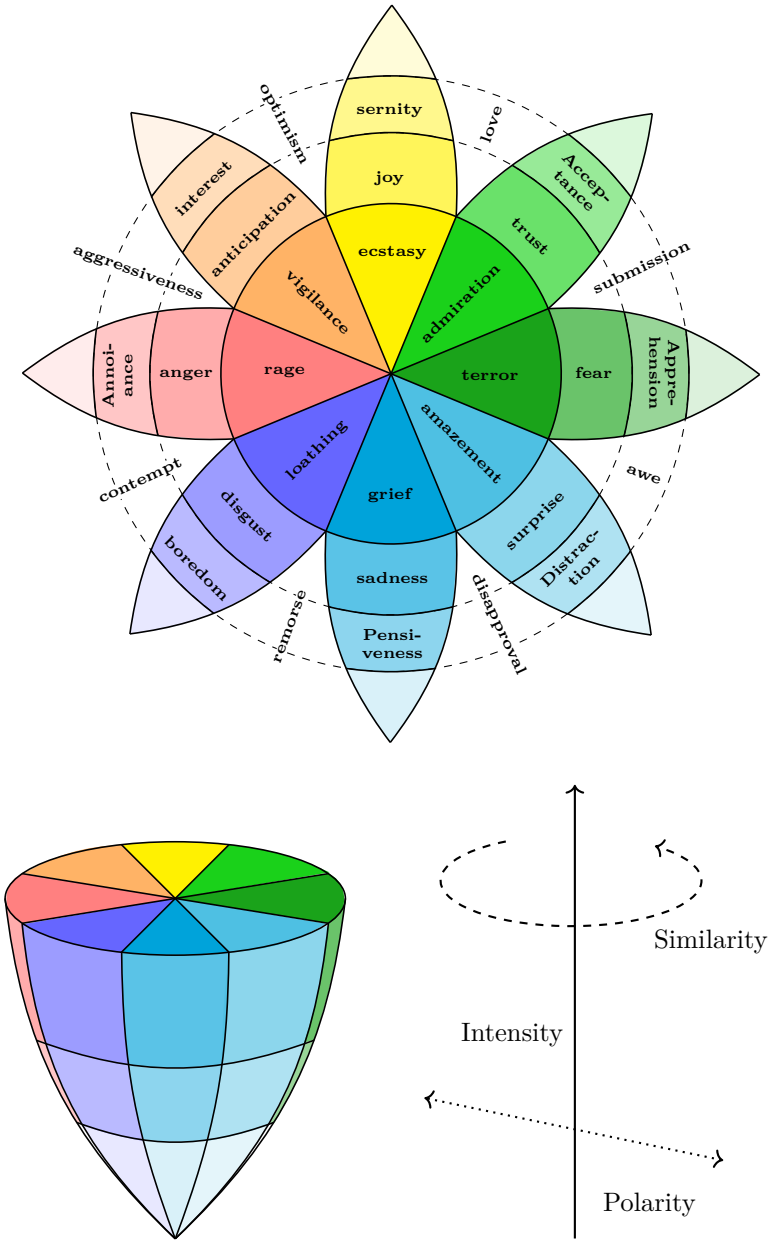


Figure 3.2: Plutchik's *circumplex model of emotions* (*Plutchik's Wheels of Emotions*) with its three dimensions: *intensity* (vertical axis), *similarity* (angular position), and *polarity* (opposition across the circle) — adapted from [Plutchik 1962, 108ff.] The model arranges primary emotions in pairs of opposites and illustrates how they can blend, vary in strength, and transition smoothly into one another within a dynamic affective space.

Against the conception of emotions as discrete and circumscribed states, for example, the American psychologist Robert Plutchik [1962; 1980] has proposed a *three-dimensional, circumplex model of emotions*

—informally known also by the name of *Plutchik's Wheels of Emotions* [Figure 3.2]. There certainly are —according to Plutchik— primary emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. However, they are not to be understood as discrete and independent affective categories, but as segments of a qualitative *continuum* where —similarly to colors on a color wheel [McDougall 1921]— each state fades into another, can assume different degrees of intensity, and is *dyadically* related to an opposite along the same spectrum. For instance, *anticipation* blends into *joy*, whose merge leads to *optimism*; also, it is complemented (and therefore set aside) by *sadness*; finally, it evolves towards *serenity* or *ecstasy*, respectively as a result of a decrease and increase in intensity.

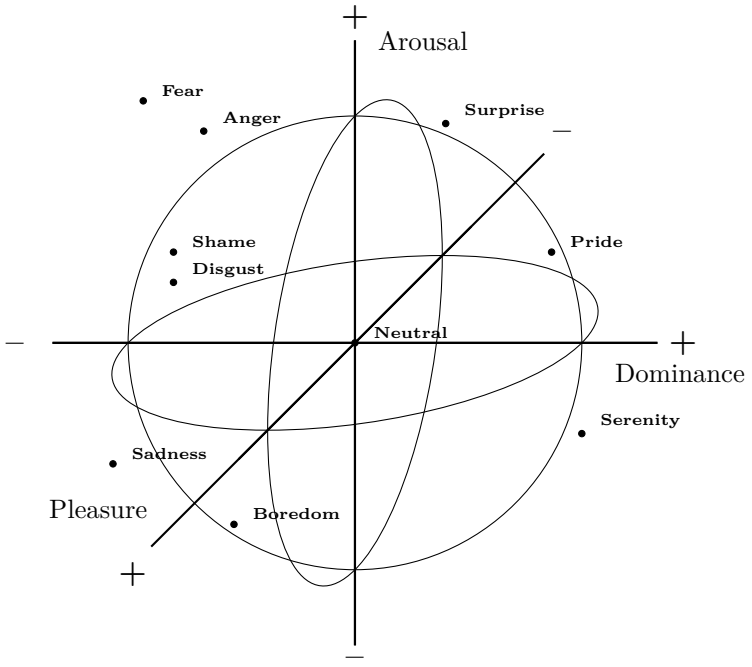


Figure 3.3: Russell and Mehrabian's *PAD model of emotions* — adapted from [Russell and Mehrabian 1974a, 22ff.] The model is based on a series of experimental studies in which participants were asked to evaluate a variety of environmental stimuli —such as photographs or verbal descriptions of physical spaces— using semantic differential scales. Each scale consisted of bipolar adjective pairs (e.g., pleased–annoyed, stimulated–relaxed, controlling–controlled), capturing affective reactions across a range of perceptual and emotional dimensions. Factor analyses of the resulting ratings consistently revealed three orthogonal dimensions: *Pleasure*, *Arousal*, and *Dominance*. These dimensions provided the basis for a continuous affective space, within which thee different emotions could be located as coordinate points.

Albeit introducing a continuous and more dynamic conception of affective states, Plutchik nonetheless remained anchored to the notion of *basic emotions* as biologically rooted, evolutionarily shaped response patterns. Conversely, a more substantial departure from this essentialist framework has been advanced by the American psychologists James Russell and Albert Mehrabian [1974; 1974; 1977] with the *PAD model of emotions*. In PAD, emotions are organized along three quantitative dimensions, respectively *Pleasure*, *Arousal* and *Dominance* [Figure 3.3]. Despite being a *dimensional* model like Plutchik's circumplex one, it significantly differs from this latter in that it rejects the discrete emotional taxonomy behind BET-based approaches. Instead, it embraces a non-categorical view of affect in which emotions are understood as configurations, regions or gradients of psychological states in a continuous affective space[50].

In particular, emotions are considered not to be organismic and genetically predetermined. Instead, they are treated as *psychological constructions* that are always mediated by perceptual, linguistic, cognitive and cultural factors, beyond biology. In other words, no ontologically independent objects correspond to the terms we use in English for categorizing the different emotional states: Joy, fear, anger, etc., in fact, only denote culturally shaped categorizations of specific configurations of valence and arousal, which vary in a context-dependent way. If so, basic emotions survive just as a functional concept (i.e., with no ontological implications) associated to the most «primitive, universal and simple» [Russell 2003, 148] level of affective experience —namely, as *core affects*, i.e., continuous fluctuations of valence (pleasure–displeasure) and arousal (activation–deactivation) levels felt prior to any contingent and linguistic categorization into discrete, determined emotional states (see also [Russell and Barrett 1999a; Russell and Barrett 1999b; Barrett 2006]).

[50] Russell [1980] later revised the PAD model into a two-dimensional structure, excluding *Dominance* as a core axis. He demonstrated through multidimensional scaling of emotional terms that most variance in affective meaning could be captured using just two orthogonal dimensions: *Valence* (Pleasure–Displeasure) and *Arousal* (High–Low activation). The *Dominance* dimension, while statistically present, was considered less explanatory and thus omitted in favor of a more parsimonious and empirically robust model.

During the last decades of the 20th century, the constructivist criticisms by Russell (and later Lisa F. Barrett) to BETs were also accompanied by a radical reevaluation of emotions as a scientific object of inquiry outside psychology, and primarily in the humanities and in the social sciences. Here, emotions came as well to be understood no longer as mere internal states or biological responses, but as shaped also by social norms, cultural scripts, and even historical contexts. This reconceptualization positioned them as central to social interaction, communicative behavior, moral order, and collective life: Since emotions depend on the interaction with the environment —as, among others, especially the emerging EC paradigm was highlighting [Section 3.2.7]— and this environment is made up not only of physical stimuli but also of other (human) subjects, cultural objects, moral norms, etc., it follows that emotions themselves have an interrelational and cultural dimension.

Developing this idea was at the core of what came to be known as the *Emotional Turn* in the social and human sciences (see e.g., [Lemmings and Brooks 2014]). Recalling the *Linguistic Turn* of the 1970s —which pointed out that language does not merely describe the world but participates in constituting reality (see e.g., [Greco and Stenner 2008; Plamper 2010])— it argued that emotions play a similar role in shaping the interaction with the world as a meaningful experience.

In sociology, for example, many scholars began to foreground the embeddedness of emotions in social frameworks [Kemper 1978; Hochschild 1979; Thoits 1989; Averill 1990; Frijda and Mesquita 1993; Parkinson 1996]: Emotions are elicited by others and directed toward others; they have tangible effects on the feelings, reactions, and behaviors of the people around us; they operate as communicative acts that articulate evaluations, claims, and positions within a shared social space. Their form, intensity, and appropriateness are shaped by institutional settings, conventions and norms, depending also on the nature of the social relationship between the subjects involved —what one feels, and how one feels it, for example, differs if one is alone, among intimates, or in the presence of strangers (see e.g., [Buck et al. 1992]). Emotions are also intersubjectively contagious, spreading through bodily attunement, mimicry, and reciprocal feedback. Moreover, the emotional expressions of others play

a constitutive role in how we appraise and interpret situations, serving as affective cues that guide our own responses within social environments.

Drawing on Allport's [1935; 1937] holistic psychology, as well as on more recent attempts to apply Wiener's [1948] cybernetics [Section 1.1.1] and control systems theory to the modeling of psychological processes (see e.g., [Powers 1973]), even formal theories appeared seeking to translate such sociological insights into mathematical accounts of emotional-social dynamics. Among the most notable ones, there was Heise's [1977; 1979; 1987] *Affect Control Theory* (ACT). According to ACT, emotions emerge from social interactions. In a social context, individuals *define* their *situation* by assigning social identities to themselves and to other individuals and objects in the same context —e.g., "student" and "teacher" at school. Towards each individual, object, event or behavior in the context, the individual develops an affective response that can be assessed along three dimensions: *Evaluation* (E) —e.g., good *vs* bad, *Potency* (P) —e.g., big *vs* little, and *Activity* (A) —e.g., fast *vs* slow. Affective responses involve two components: *fundamental sentiments* (FSs), depending on the assigned identity and remaining stable over time —e.g., teachers are typically felt by the students as good (E), authoritative (P), and calm (A); and *transient feelings* (TFs), occurring in *deflations*, i.e., when something in the context happens to deviate from FSs —e.g., when a teacher behaves badly (E) and/or irresponsibly (P) and/or and harshly (A) toward the students. Importantly, affective responses are mediated by cultural and social schemes —FSs and TFs towards teachers, in fact, may have different EPA characterizations depending on culture and social context. On the contrary, EPA themselves, as dimensions of affective response, are *cross-cultural*. Now, FSs and TFs function as a control system: In a continuous feedback loop, the former act like governors that cause the individual's behavior to adjust in order to reduce discrepancies with the latter and reacquire homeostasis. In general, individuals tend to confirm the FSs as evoked by the initial definition of the situation, holding events that may cause a deflating TFs to be highly improbable (and vice versa). Does TFs instead deflate from FSs, then individuals reinterpret the situation and, with it, the assigned roles, their EPA profiles, viz. FSs themselves towards the elements of the context: «New identities are assigned to those on the scene in order to provide new

confirmable sentiments that make sense of recent and ongoing events» [Heise 1977, 165] —e.g. the students reassess the teacher as actually bad (E) and/or irresponsible (P) and/or and harsh (A). The way in which individuals interpret situations, respond emotionally to events, interact with the context, change FSs, etc. can be even predicted through a set of empirically derived mathematical equations that Heise provided also as a basis of a computational implementation of the model, INTERACT, loosely reminiscent of HOMUNCULUS [Section 2.2.2] and, in its conceptual foundations, at the core of many efforts to build CMEs in the 1990s (see e.g., [Vyzas and Picard 1999] and [Section 4.1]).

Sociological theories about emotions and ACT, hence, represented not only a move beyond the standard psychological view of emotions as fundamentally private and internal psychophysical states, but also beyond the biological reductionism about affectivity purported by BET.

This stance, in fact, began to be increasingly challenged by a growing body of ethno-anthropological and cultural research. This specifically tried to highlight the heavy dependency of emotions on frameworks, practices and norms that significantly vary across different cultures: What counts as a salient emotion in one culture may be entirely absent or differently construed in another. In this respect, for example, Michelle R. Rosaldo's [1980] ethnographic work in the Philippines, showed that, for the *Ilongot* people, emotions like *liget* (a form of socially valorized anger or energy) are tightly bound to cultural values and ritual practices, opposing the notion of discrete, universal emotional states. Similarly, Catherine A. Lutz' [1988] studies among the *Ifaluk* people in Micronesia showed that emotions such as *fago* (a blend of compassion, love, and sadness) are experienced and interpreted in ways that resist any direct mapping onto Western emotional taxonomies.

Later, the core ideas of this *emotional constructivism* were extended to encompass even a *historical perspective*, which began to emphasize that emotions are not only socially and culturally variable, but also *historically* contingent: What and how people emote in the present may have been different in the past, and will probably change in the future. Thus, from the mid-1980s onwards, existing efforts to understand emotions as historical objects made in history of psychology, historical sociology, cultural history, historical anthropology and other disciplines

(see e.g., [Febvre 1941; Elias 1976; Gay 1984-1998], etc.) eventually converged in a new, autonomous and systematic research field, the *history of emotions*, trying to systematically investigate how emotional norms, vocabularies, expressions, and experiences are shaped by —and in turn shape— specific historical contexts. Within this emerging field, scholars began to reconstruct the emotional repertoires of past societies, identifying how certain feelings were encouraged, discouraged, or rendered intelligible within distinct temporal frameworks. This approach entailed not only tracing changes in emotional expressions, but also analyzing the social technologies —rituals, pedagogies, moral codes, medical discourses, etc.— through which emotions were regulated and codified over time. Authors such as Stearns and Stearns [1985], Reddy [1997, 2001] and Rosenwein [1998, 2002] were among the first to argue that emotions are shaped by historically specific norms, institutions, and discourses, and to develop methods for tracing how emotional expressions, classifications, and expectations are organized differently across times and settings. Their work emphasized that feelings are not simply experienced, but are taught, regulated, and contested within concrete social frameworks, leaving identifiable traces in sources such as texts, rituals, and practices (see also [Plamper 2010; Frevert 2011, 2024; Stearns and Matt 2014; Dixon 2023], etc.).

Summary

This chapter has traced the main technical and philosophical transformations that, between the 1980s and 1990s, prepared the emergence of *Affective Computing*. It has examined how the rise of subsymbolic AI, the paradigm of embodied cognition, and the sociocultural reconceptualization of affect collectively reshaped the scientific understanding of emotion. Through these converging trajectories, emotion ceased to be conceived merely as an auxiliary control variable of cognition and came to be viewed as a constitutive, embodied, and socially distributed component of intelligent behavior. From neural computation to enactive and sociocultural theories of mind, affect thus became the bridge linking the biological, the cognitive, and the social dimensions of intelligence.

In a broader sense, these developments marked the passage from an abstract and disembodied conception of artificial emotion to one grounded in interaction, adaptation, and meaning. They established the epistemic and technical conditions for the birth of AC, in which emotion would no longer be a marginal add-on to cognition but a principle of design for intelligent systems themselves.

In the following [Chapter 4], this *paradigm shift* will be explored in depth by analyzing the conceptual foundations of AC and the way it operationalized in new working computational models of affect years of reflections about non-representationalism, embodiment, situatedness and affectivity. This will enable us, in [Chapter 5], to consider then one of the most direct concretizations of AC's research program, that is, social and affective robotics.

Chapter 4

Affective Computing and the Rethinking of Computational Emotions

In the last [Chapter 3], I examined the (implicit) historical and epistemological roots of *Affective Computing* (AC), focusing firstly on the transition from the symbolic to the subsymbolic paradigm in AI. Central to this shift was the adoption of more flexible and adaptive forms of computation based on neural networks and connectionist approaches. At the same time, a profound reconceptualization of emotion took place in philosophy, psychology and social science: No longer viewed as irrational residues opposed to cognition, emotions came to be understood as embodied, dynamic, and context-dependent processes, deeply integrated with cognition, perception, action, social interaction and cultural norms. This dual transformation —both technical and conceptual— laid the necessary groundwork for a new approach to modeling affect computationally.

In this chapter, I build on these considerations to introduce the core tenets of AC. [Section 4.1] examines how AC, especially as a programmatic proposal defined by Rosalind W. Picard in the late 1990s, codified and operationalized the developments traced in the previous chapter within a new situated, embodied and subsymbolic conception of CME. [Section 4.2] then traces how Picard's ideas were further developed in later CME research, highlighting the emergence of subsymbolic

approaches to artificial emotions and the shift toward more dynamic, embodied, and context-sensitive models of artificial affect. Finally, in [Section 4.2.2] and [Section 4.3], I point out that AC, especially because of its emphasis on emotion recognition, introduced a form of *physiological-behavioral reductionism* about emotions in CME practices. Historical reasons are given for this particular development, and some of its consequences up to contemporary CME research are explored.

4.1 Rosalind Picard's Revolution

4.1.1 Enhancing Human-Machine Interaction

> «Today it is easy to find people who spend more time interacting with a computer than with other humans. Every day people enter the online communities of the Internet where they communicate with each other *through* computers. Daily interaction between humans and computers has billions of dollars of economic impact, not to mention psychological impact [...].
>
> It is not unusual for intelligent people to attribute emotion to things that clearly do not have emotion [...]. Emotion plays an essential role in communication [...]. Affect recognition and expression are necessary for communication of understanding, one of the greatest psychological needs of people [...]: [E]ven a puppy can tell when you are angry with it [Negroponte 1995].
>
> However, unlike the puppy, computers today cannot even tell if you are pleased or displeased. They will scroll screenfuls of information past you regardless of whether you are sitting forward eagerly in your seat, or have begun to emit loud snoring sounds. Computer-based communication is affect-blind, affect-deaf, and generally speaking, affect-impaired. A quantum leap in communication will occur when computers become able to at least recognize and express affect».

So spoke Rosalind W. Picard [1997, 14-15], American computer scientist at MIT, on the eve of the new millennium. Toward the end of the 1990s, computers were rapidly spreading into homes, schools, and workplaces, becoming part of the daily routine for many people. Email, word processing, and early web browsing turned the computer into a common domestic appliance, used for both work and leisure. As dial-up connections brought people online, even ordinary users began to interact with computers for communication, information, and entertainment on a regular basis.

The growing presence of computers in society was slowly transforming them from mere passive instruments to active social agents, i.e., subjects onto which social meanings and roles —conversational partners, co-workers and everyday companions— could be projected [Rheingold 1993; Turkle 1995]. People were beginning to relate to computers in the same way they would to fellow humans —a phenomenon that was even attracting the interest of empirical psychology [Turkle 1984; Reeves and Nass 1996], as well as of artists and (cyberpunk) novelists [Coupland 1995; Ephron 1998].

This transformation was so profound that it already sparked new cultural and philosophical reflections about human identity and the ethical and social concerns deriving from the entanglement between people and machines [Postman 1992; Terry and Calvert 1997; Katz et al. 2001]. But most importantly, it galvanized the establishment of new and autonomous fields of interdisciplinary scientific research: Situated at the crossroads of psychology, computer science, design, philosophy and many other disciplines, HMI (*Human-Machine Interaction*), HCI (*Human-Computer Interaction*), HRI (*Human-Robot Interaction*), etc. aimed at investigating the modalities, dynamics, and implications of interactive relationships between humans and computational systems (see e.g., [Preece et al. 1994]), trying to develop interfaces, systems and machines capable of engaging users in intuitive, efficient, and interactionally effective ways.

And exactly from the need to build more *responsive* computers developed Picard's [1997] proposal of a framework for *affective computing*, i.e., «*computing that relates to, arises from, or deliberately influences emotions*» [3, my italics]:

> «Most people use PCs, which [...] are embarrassingly unreliable [...]. Complaints abound regarding this or that piece of software [...]. However, the fact is that computers receive the brunt of people's complaints, only they do not yet have the ability to recognize these complaints. *People express their frustration to the computer, but it cannot see it or do anything about it.* Giving computers the ability to pinpoint what they do that most frustrates users, and to see just how much it really frustrates them, would be a step toward helping designers of computer tools make improvements [...].
>
> Most people know very little about how computers work and find computers and software increasingly hard to understand [...]. [U]sers have been caught in a spiral of increasing *complexity*, where figuring out how

> to get the machine to do something they want can be so frustrating that it is easier to change what one wants than to change the machine to do what one wants [...]. [C]omputers are either rigid and inflexible, or so loaded with options that the user has to invest huge amounts of time trying to determine how to set them all. [*T*]*he computer simply does not care what its user thinks or feels; it does not speed up when he is bored, slow down when he is confused, or try to do things differently when he is frustrated with its current modus operandi* [...]. Today's computers are far from being human-centered systems; they cannot even see if they have upset their most valuable customer» [248, my italics].

For this reasons —Picard [1997, 247, my italics] pointed out— the computer should be able «to function with intelligence and sensitivity toward humans [...], to recognize its user's affective expressions, and to respond *intelligently*, especially if [he] indicates frustration, fear, or dislike of something it can change». In a word, it «should have the skills of *emotional intelligence*, including an ability to manage its own emotional mechanisms, and to use them for improving its cognitive and rational functioning».

Early CMEs (1960s-1980s) emerged as rather scientifically driven attempts to build either accurate theoretical models of human cognition, in which emotions worked as operational variables in catalyzing and assisting computational processes (AAIMCO), or computer programs able to simulate human emotions (AIME) [Chapter 2]. Conversely, as made clear from the long quotations above, Picard's [1997] AC proposal seems to have stemmed rather from a very practical, engineering need, in an ever more technologically entangled society: giving affective capabilities to machines in order to improve their efficiency, usefulness, adaptability and interactions with humans.

Namely, in her book, Picard [1997, Ch. 3] explored a wide range of potential applications of this idea across different domains (see also [Section 8.3.4]). For instance, affective computers —beyond, say, mere emoticons— may assist in conveying inflections, prosody and therefore communicative intents in written texts through appropriate text-to-speech conversion techniques. This may be useful in the context of assistive technologies, where emotionally aware systems could support individuals with disabilities or mental health conditions (e.g., seeing-impaired or autistic people) by interpreting and responding sensitively to their emotional states. In customer service, moreover, computers equipped

with emotional capabilities might recognize dissatisfaction or confusion in users and adapt their responses in real time to improve feedback interaction. Or still, educational software capable of detecting student frustration and adjusting their behavior accordingly may help in developing effective learning strategies, improving engagement and reaching learning outcomes. Even in entertainment, games could become more compelling and immersive by modulating their behavior based on user affect.

Many of these AC applications have been since then realized in tangible and widely adopted technologies. Yet, they have often remained partial and devoid of the depth envisioned in Picard's original proposal. In this sense, almost thirty years later her book has been published, her critique of everyday technology as fundamentally *unaffective* seem to remain still true. Far from being only the result of technical obstacles, this situation also reflects deeper tensions concerning the role of emotions in computing —tensions already Picard [1997, Ch. 4] warned about as possible sources of ethical, psychological, and social risk. Artificial emotions, for example, may be used to manipulate or deceive, paradoxically inverting the original aim of the proposal. Too emotional machines can conflict with expectations of "cold" objectivity, efficiency and competence, making them be perceived as unreliable and infantilizing. Furthermore, systems capable of accessing or inferring users' emotional states raise serious privacy concerns, especially when such data are collected, shared and sold without consent.

4.1.2 Reconceptualizing Affect in Machines: Theoretical and Technical Aspects

But let us now more closely concentrate on Picard's [1997] original proposal. Besides responding to the mentioned concrete design challenges, Picard's [1997] AC also framed, implicitly and explicitly, a profound theoretical account of artificial emotions that reflects both the achievements reached by previous efforts to model emotions computationally, and the new understanding of emotions developed in contemporary medical, psychological, philosophical and cultural debates about human mind and affectivity [Chapter 3].

4.1.2.1 Artificial Emotions as a Form of *Intelligent* Sociability

In this respect, an essential aspect of her argument has been the emphasis on the *social* and *interrelational* dimension of emotions —understandable as directly aligned with AC's aim itself to foster more effective interactions between humans and machines, by then increasingly conceived as social agents. This orientation had already been evident in some early CMEs such as HOMUNCULUS, PARRY, or FEELER [Section 2.4.3]. However, in such models, the sociality of emotions often remained implicit or functionally subordinated, as an operational variable, to other (cognitive) computational mechanisms —in HOMUNCULUS, for example, as an instrument for driving rational decision-making in simulated social environments; or, in FEELER, as a possible trigger of emotions generation processes (interrupts in execution plans caused by real or simulated individuals). In early CMEs, in other words, affective states were not expressions of genuinely relational dynamics.

In contrast, what AC proposed was exactly to place the relational aspects of emotion at the core of computational system design, attributing to human-machine interaction the same interrelational characteristics of human-human interaction, as something that occurs in a communicative space shaped by emotional cues, expectations, and feedback. Emotions were now considered to be useful in machines not only for optimizing cognitive tasks, but also for gaining contextual awareness, adapting responses and, generally, contributing to an integrated and dynamic affective ecology of social subjects with (hopefully) symmetrical interactional capabilities.

Thus, rather than just expanding upon early CMEs' intuitions about emotions as social objects, AC introduced a true *paradigm shift* in this respect, by casting *even* computational emotions as mediators of social interaction, knowledge and communication, not completely reducible to mere functional states. Of course, such reconceptualization echoed the new sociological insights about emotions developed between the 1980s and the 1990s [Section 3.3]: AC, in fact, (implicitly[51]) presented itself as

[51] Although Picard [1997] does not discuss explicitly and in depth sociological theories of emotions, references to authors such as Averill [1990], along with recurrent acknowledgments of the sociocultural conditioning of emotions and their communicative function, make it difficult to deny that the broader debate on emotions as

an extension to the realm of *artificial* emotions of the characterization sociologists and anthropologists were giving of human *affects* as socially shaped phenomena, regulated by norms, roles, expectations, and deeply embedded in the structural dynamics of social life —thereby suggesting that these insights could be directly taken as the target of a computational formalization.

In this sense, giving emotions to machines —and not merely in the form of internal ends-in-themselves computational states, but of relevant means of interaction— amounted, in AC, to transforming them from functionally isolated tools into socially *situated* agents. Emotions help navigating and *enacting* an environment populated also by other social subjects, human or artificial, enabling the sense-making of their decisions, actions, intents and affects themselves [Section 3.2.8]. In other words, emotions —and artificial emotions too— are integral to, and not separable from, *cognition*, whereby cognition is enacting an (intrinsically social) environment [Section 3.2.7].

As pointed out in [Section 1.2], from a historical point of view, the recognition of the deep intertwining of emotion and cognition was not a novel achievement of late-1990s science. Conversely, it was rather a view that psychology and philosophy had been already discussing from at least the 1930s-1940s, being even anticipated by key figures of cognitive science (see e.g., [Simon 1967; Minsky 1985], etc.). Even within AI, from the early 1960s, the idea that emotions play a crucial role in cognitive processes was operationalized into the creation of the first CMEs. In the 1980s and 1990s, neuroscience and embodied approaches to the mind further strengthened this view, showing that emotions are constitutive of perception, reasoning, decision-making, problem-solving and other core "cognitive" functions of the human mind (see e.g., [LeDoux 1984; Damasio 1994], etc. and [Section 3.2.8]).

social objects in the 1980s-1990s [Section 3.3] contributed to the general theoretical background in which AC emerged. Also, in other technical works expressly aligned to AC's theoretical guidelines (see e.g., [Vyzas and Picard 1999]), social theories of emotions like ACT [Section 3.3] are taken as explicit operational frameworks in the artificial modeling of emotions.

Indeed, Picard [1997, x, 1-13, 48-49, 218–219, 247-248] demonstrated a clear awareness of this body of literature and debate[52], and explicitly acknowledged the essential relationship between emotions and intelligence as a foundation for the effective implementation of AC technologies and their social fruitfulness. The aim was no longer that of just exploring the functional relation between emotion and cognition as in early CMEs (see e.g., [Section 2.2]), but that of making computers emotionally, and thus socially, *intelligent*, whereby this intelligence cannot be but *inherently* emotional from the outset:

> «Today we have evidence that emotions are an active part of intelligence, especially perception, rational thinking, decision making, planning, creativity, and so on. They are also critical in social interactions, to the point where psychologists and educators have re-defined intelligence to include emotional and social skills. It is no longer accurate to think of emotion as some kind of evolutionary relic or as some kind of luxury, unnecessary in either case for intelligence. Instead, emotion has a critical role in cognition and in human-computer interaction. Computers do not need affective abilities for the fanciful goal of becoming humanoids; they need them for a meeker and more practical goal: to function with intelligence and sensitivity toward humans» [247].

4.1.2.2 Recognition, Multimodality and Embodiment

Especially in light of the social connotation of artificial emotional intelligence in AC, the divergence from early CMEs becomes evident also in the fact that, whereas early CMEs —AIME in particular [Section 2.3]— focused prevalently on *emotion generation*, AC emphasized a complementary and increasingly central dimension of the computational modeling of affect, that is, *emotion recognition*. For enhancing the interaction with the users, in fact, socially intelligent affective computers not only need to express emotions appropriately, but also —and most importantly— to *recognize* them. Picard's [1997] AC contributed to catalyzing the development of new techniques for emotion recognition, providing a systematic and programmatic consolidation of early, yet scattered and mostly rudimental, attempts to build emotionally responsive computers. In AC,

[52] Yet, Picard [1997] does not show particular depth and accuracy in treating early CMEs and their historical underpinnings in psychology, philosophy and cognitive science. This is understandable, given the book's technical and design-oriented focus, even though it sometimes leads her to overestimating the conceptual novelty of some aspects of her proposal.

these efforts were integrated within a unified and socially-oriented research framework, within which the influence of the embodied approach and the subsymbolic paradigm became increasingly apparent.

As observed above, the embodied approach in psychology and philosophy of mind —building upon the ecological views of Gibson [1979] and Hatfield [1988]— posited that the mind cannot be conceived as a computational system separated from the body and from the environment. Instead, the body, and its sense organs, used for interacting with the environment, are an *integral* and *indiscernible* part of the mind [Section 3.2.7]. Now, in this regard, Picard's [1997] AC —especially through its implications on social and affective robotics [Chapter 5]— represented a notable attempt to systematize and operationalize this idea in computation[53]. Artificial emotions, as not detached from artificial intelligence, are not just functional internal states of a machine, but part of its broader capacity to interact with other subjects. If so, the machine necessarily requires a form of embodiment to enact its environment and make it affectively and thus socially meaningful. In other words, following Picard [1997], to be emotionally intelligent the computer must be equipped with suitable *multimodal hardware* (i.e., capable of retrieving and integrating stimuli from distinct sensory channels —visual, auditory, physiological, etc.) that allows it not only to "perceive" the external world, but also to gather emotionally relevant information from it.

Here, reinforcing the long-acknowledged relationship between emotions and the body [Section 3.2.1], she outlines a large spectrum of devices and sensors that can possibly constitute the "sensing body" of the computer, and serve for collecting emotionally relevant physiological and behavioral *signals*. These range from basic apparatus like cameras and

[53]Interestingly, this anticipates some relatively recent intuitions formulated in direct applications of EC to affective phenomena (see [Section 3.2.8] and [Footnote 44]). However, as in the case of sociological theories of emotions [Footnote 51], Picard [1997] does not explicitly reference foundational works in ecological psychology or the embodied mind paradigm. Nevertheless, she does cite authors —most notably (and indeed not surprisingly) roboticists like Rodney Brooks and Rolf Pfeifer [Chapter 5]— who, during the 1990s, were actively exploring the computational implications of EC. Thus, while the influence remains partly indirect, EC's emphasis on the integration of sensing, cognition, and emotion —visible in AC's focus on emotion perception through sensors and adaptive, embodied interaction— is undeniably reflected in the foundations of AC, to the extent that EC implicitly constitutes a substantial part of its theoretical background.

microphones for capturing behavioral signals like facial expressions, pupil dilation, vocal tone, gestures, gait, posture, etc. to sensors for detecting physiological parameters such as heart rate, galvanic skin conductance, body temperature, respiration rate, blood pressure and oxygen levels, hormone and neurotransmitter levels, muscle activity, brain activity, etc. Notably, among the latter there are also *affective wearables*, devices designed to remain in close and continuous contact with the user's body, allowing for the real-time collection of emotionally salient bodily signals that are normally inaccessible to humans [Picard 1997, Ch. 8] —this should theoretically allow in computers even better performances in emotion recognition tasks compared to humans.

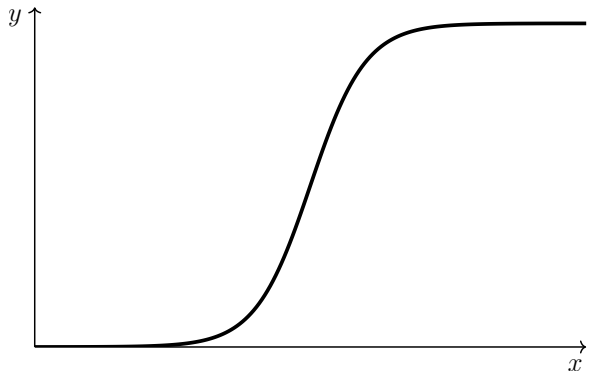
Once these heterogeneous physiological and behavioral signals have been collected by the machine's "sensing" hardware, they are processed in such a way that, from them, the user's emotional state can be eventually inferred (i.e., recognized and attributed to her). To this aim, given their analog and continuous form, they firstly have to be converted into a form that is computationally tractable. Thus, they are sampled, normalized and digitized, which produces a representation of the signals as a discrete, time-independent series of binary numbers that the computer can use for further elaboration [Picard 1997, 163]. For instance, a respiration signal sensed through the stretch of an elastic chest band is sampled (at a suitable frequency) and stored as a sequence of binary values indicating expansion over time. After digitization, signals undergo a process of feature extraction in which numerical values are computed from the raw data to summarize aspects that are relevant to emotional dynamics. For example, from the digitized respiration signal, the system might extract the maximum or minimum amplitude of chest expansion, the rate of change in breathing, or the number of breath cycles per minute.

At this point, the system determines whether and to what extent the input so extracted can possibly trigger an emotional response. In general, emotional activation follows a *linear* trend: Higher input values lead to more intense emotional responses [Picard 1997, 150]. In humans, for instance, higher breathing rate correspond to, say, higher excitement. Emotional response, however, is determined by an uncountable number of uncontrollable factors: The same input may cause a higher or lower emotional activation depending on an individual's temperament, mood,

cognitive expectations, situational context, biochemical conditions like hormone and neurotransmitter levels, etc. For example, an increased breathing rate might elicit a strong emotional response in a highly anxious person, while producing little to no activation in someone who is naturally calm, physically trained to tolerate sudden physiological changes, or is simply doing sport. This means that linear systems can only approximate the behavior of an emotional system within specific boundaries: Inputs under a specific *threshold* will produce no significant emotional response, whereas extremely strong inputs will *saturate* activation, ceasing to generate proportionally greater and differentiable responses. If so, emotional activation can be only modeled as a *nonlinear* process. To this aim, relying on linear systems theory and digital signal processing theory, Picard [1997, 151ff.] proposes a nonlinear sigmoid function for emotional activation:

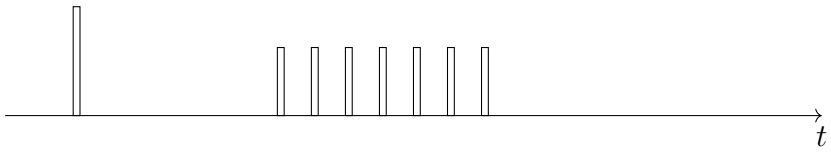
$$y_{\text{act}} = \frac{g}{1 + e^{-(x-x_0)/s}} + y_0 \tag{4.1}$$

where x is the input —i.e., any possible stimulus that may trigger an emotional event (e.g., the frequency or intensity of breathing); y_{act} is the output —i.e., the value of emotional activation; s controls the steepness of the curve —how abruptly the system moves from no response to full activation— and is associated with personality, temperament or reactivity (e.g., a highly reactive person might respond strongly to a slight increase in breathing rate); x_0 sets the activation threshold, shifting the curve horizontally and modeling mood influences (e.g., a person in a negative mood may require a smaller increase in breathing to trigger a response); g determines the amplitude of the response and is linked to arousal (e.g., someone under physiological stress or excitement will show a stronger activation for the same respiratory input); and y_0 shifts the curve vertically, being correlated to cognitive expectations (e.g., if one expects something stressful to happen, even normal breathing might produce an anticipatory emotional activation). As x approaches and surpasses x_0, the system enters an almost linear region, where the response increases proportionally. Beyond this range, when x becomes very large,



the function flattens, and y approaches its maximum value $g + y_0$ reflecting saturation.

Now, of course, emotions are not instantaneous events. Rather, they exhibit temporal dynamics: They begin with rising intensity, and then naturally decade within minutes, unless reactivated by new similar stimuli, that prolong and amplify it. For instance, a sudden increase in breathing rate due to stress may trigger an emotional response that gradually fades, but if new stressors continue to appear —such as repeated negative thoughts or external pressure— the emotional state is sustained or even intensified. In order to clarify this, Picard [1997, 145ff.] refers to an example: the sound produced by a bell struck by a hammer. When the bell is struck once, the resulting sound —influenced by the physical characteristics of the bell (size, material, etc.), paralleling the role of mood, temperament, etc. in emotions— reaches a peak and then gradually fades away. However, if it is struck repeatedly, the sound persists and even grows louder, as the successive strokes build upon the residual resonance of the previous ones [Figure 4.1].



(a)

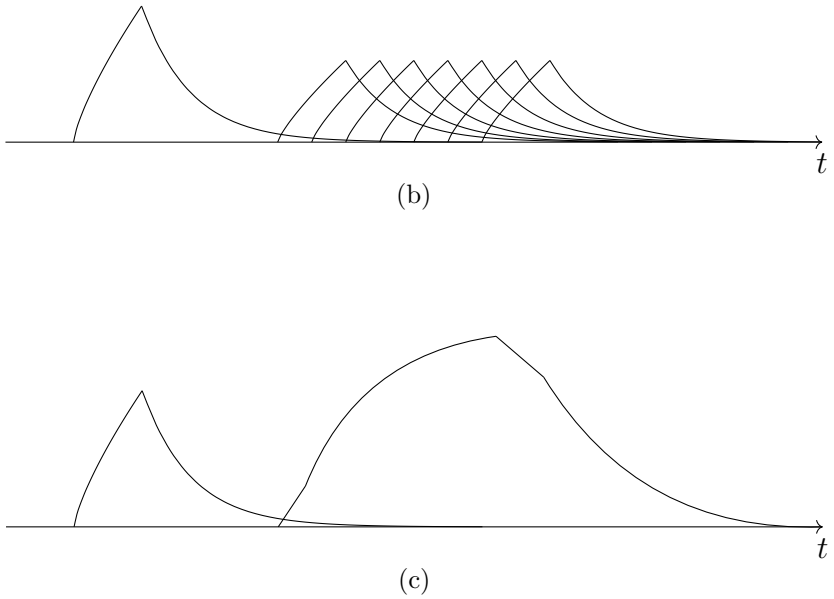


Figure 4.1: Illustration of emotional activation modeled through the bell metaphor —adapted from [Picard 1997, 146]. (a): Sequence of stimuli over time (hammer strokes with different intensities), represented as discrete events; (b): Commensurate responses generated by each stimulus —a single stimulus produces a sharp peak followed by exponential decay, whereas repeated stimuli prolong and amplify the response; (c): Overall intensity of the composite signal resulting from the sum of all successive responses.

Thus, if the computer is to represent emotional signals adequately, it is not sufficient for it to produce, given an input x, a static activation value. Instead, it must be additionally able to model an emotional response in the time[54]. A dynamics such as that of the bell can be mathematically described using standard exponential solutions to first-order linear differential equations for decay processes, for example:

$$y(t) = a \cdot e^{-bt} \tag{4.2}$$

where $y(t)$ is the output intensity of the sound produced by the bell at time t; a the initial intensity of the sound immediately after the stroke; and b a constant determining the rate of the sound decay. Now, a similar approach can be used in computers to mathematically model the internal evolution of an emotional response over time: The activation value

[54]The acknowledgement of the temporal dimension of emotions was however still present in some AIME (e.g., FEELER [Section 2.3]), albeit sometimes not treated with the same theoretical and technical depth of AC.

$y_{\text{act}} = a$ output by the sigmoid [Equation 4.1] represents the initial intensity of the emotional response, which then evolves in time and decays following a predictable curve $y(t)$ [Equation 4.2].

4.1.2.3 Emotional Pattern Recognition and Learning

At this stage, the system possesses a *low-level* internal representation of emotionally relevant *signals*. Yet, this do not yet amount to the computer having recognized a specific emotion: In fact, all what it has is just an implicit information indicating that an emotionally significant event is occurring, but still lacks the capacity to categorize it as anger, joy, sadness, etc. For this, the signals need to go through at least two other representational subprocesses: a *medium-level* one for *pattern recognition*, and a *high-level* one for *emotional category attribution* —processes in which the influence of 1980s–1990s developments in connectionist and subsymbolic computational approaches [Section 3.1] becomes particularly evident, as also reflected in Picard's [1997].

Emotions —whether expressed through physiology, facial movements, voice, or behavior— are rarely conveyed as simple and discrete units of information. Instead, they are usually expressed as more or less organized *patterns* of multimodal expressive cues. For example, fear may involve an increased breathing rate, facial muscle contraction, and a tightening of the voice. Taken individually, none of these cues is sufficient for indicating fear, whereas this is generally not the case when they are considered collectively. Now, in order to effectively recognize an emotion, the computer must identify in the signals the patterns of information that are typically characteristic of a certain emotion, and suitably represent them in a computationally tractable form. In other words, the process of artificial emotion recognition can be partly casted as a problem of *pattern recognition* [Picard 1997, 165].

Once the patters have been identified, the computer can eventually attribute correspondent emotional labels (joy, fear, etc.) to them. This result fulfills the recognition task[55], being output in explicit form to the

[55]It is important to notice that the fact that a computer can recognize affective states through such procedures does not amount, for Picard [1997], to saying that it can recognize emotions. What the computer recognizes, in fact, is only the expression of these emotions. Some aspects of this ontological claim will be broadly discussed

user or being internally used by the machine for further adapting its behavior towards her in interaction with other modules of its cognitive architecture (e.g., memory, attention, etc.). In this respect, emotional pattern recognition is obviously simplified if one either assumes a small set of discrete emotional categories (such as in BETs [Section 3.3]), or a small number of dimensions along which affective states can be distributed (like in dimensional theories of emotion [Section 3.3]). Under these theoretical assumptions, in fact, the variety of the possible outputs is constrained, which lowers the overall complexity of the recognition process [Picard 1997, 170].

By the 1990s, a number of computational approaches to pattern recognition and modeling had appeared —ranging from deterministic, rule-based techniques to more flexible data-driven methods (see e.g., [Therrien 1989; Schalkoff 1992; Bishop 1995]). However, few such techniques specifically dealt with computational tasks involving emotions. Essa and Pentland [1997] (see also [Yacoob and Davis 1996]), for example, proposed a more nuanced version of FACS (see [Section 3.3] and [Footnote 47]) that —paralleling Picard [1997]— also accounted for the *temporal decay dynamics* in emotions (application, release and relaxation) as expressed by facial movements —an aspect initially disregarded by Ekman and Friesen [1978], limited instead to just static, time-independent emotional expressions. The system receives as input continuous video streams of a human face in motion. After applying appropriate algorithms for normalization and noise reduction, low-level face motion signals are captured across successive frames using optical flow techniques, specifically in the form of changing parameters of a geometric mesh modeled onto the face of the subject [Figure 4.2]. The evolution of such parameters reflects an estimate of the deformation of facial muscles over time. From this, characteristic sets of values (features vectors) quantifying the intensity and distribution of muscle activation are extracted, i.e., expressive *patterns* potentially correlated to emotional states. Thus, at this point, the system initiates emotion recognition by calculating the

in [Part II]. However, as already stated in the [Introduction], this work takes this distinction for granted. In other words, it is here not under discussion that machines cannot "*really*" recognize or ontologically have emotions —a question that is, indeed, neither posed nor answered here in the case of human affectivity.



Figure 4.2: Geometric mesh applied on a face (single video frame) — adapted from [Essa and Pentland 1997, 2]. The changing position of the mesh nodes across video frames is modeled as a matrix of displacement vectors, capturing the spatio-temporal dynamics of facial motion.

distance of features vectors from prototypical expressive-emotional templates, previously derived from recorded sequences of expressions. The template corresponding to the highest similarity score is then selected as the recognized emotional category, thus completing the classification process. Of course, by appropriately define deformation parameters in accordance with a specific emotional template, the process can be reversed, i.e., adapted to *emotion generation*. In both cases, even in absence of neural networks and machine learning, the model operates sub-symbolically following a connectionist logic: It does not rely on discrete emotional categories or semantic labels, but processes facial expressions as distributed patterns of activation encoded in continuous-valued vectors. These are classified not through rule-based inference, but through direct geometric calculations on low-level, temporally structured representations.

Other models conceptually similar to Essa and Pentland's [1997] existed, that applied emotional pattern recognition mostly to *monomodal* expressive inputs —Cahn [1990] and Roy and Pentland [1996], for example, designed some primitive computational methods for recognizing emotions from speech. With Picard's [1997] AC *revolution*, a coherent and comprehensive theoretical foundation was eventually provided to such technical efforts trying to make computers able to recognize emotions. In particular, Picard [1997, Ch. 6] suggested new technical approaches to artificial emotion recognition that integrated *multimodal* expressive signal and advanced *machine learning* techniques beside traditional statical

pattern recognition —direct reflections, as said, of the influence of EC [Section 3.2.8] and subsymbolic AI [Section 3.1] on AC's new conception of artificial emotion.

Picard [1997, 185ff.] and Vyzas and Picard [1998; 1999], for example, devised a new ML-based method for teaching a computer to recognize emotions:

> «The idea is to have the computer observe multiple signals gathered while a person is experiencing an emotion [...] and learn which patterns of physiological signals are most indicative of which affective state. Later, when the system is given only raw signals from a person, then it can use what it has learned previously to try to recognize which affective state most likely gave rise to the signals» Picard [1997, 185].

In the training phase —that in which the computer builds a model of emotion— emotionally significant information is extracted from a heterogeneous set of four physiological signals: EMG, BVP, GSR and Respiration. These are monitored through sensors worn by an actress deliberately expressing emotions (in particular, no emotion, anger, hate, grief, platonic love, joy, romantic love, and reverence) over a period of twenty days. The physiological signals are gathered, sampled, normalized and digitized. From the resulting data, a set of six statistical features is extracted for each signal. Features are numerical summaries that represent different aspects of intensity, variability, and change *over time* in bodily responses associated to specific emotional occurrences[56]. This yields a total of 24 features for each emotional episode: A single emotional expression on one day is represented by a point in a 24-dimensional space, and all emotional occurrences on all days are represented by a dataset of

[56]Let S_n^i be the value of the ith signal at the nth sample point, for $n = 1, \ldots, N$ and

$$i = \begin{cases} 1 & \text{EMG} \\ 2 & \text{BVP} \\ 3 & \text{GSR} \\ 4 & \text{Respiration,} \end{cases}$$

whereas

$$\tilde{S}_n^i = \frac{S_n^i - \mu^i}{\sigma^i} \quad i = 1, \ldots, 4$$

the normalized signal. The six features are: (1) *Mean of the raw signal*:

$$\mu^i = \frac{1}{N} \sum_{n=1}^{N} S_n^i \quad i = 1, \ldots, 4$$

160 points distributed in the feature space. Each such point captures the unique configuration of physiological activity associated with a particular emotional occurrence, enabling comparisons across different emotions and different days.

So-represented physiological signals are then submitted to ML pattern recognition algorithms, that identify consistent physiological regularities associated with specific emotional states —in particular, k-NN classifiers, *Maximum a Posteriori* (MAP) estimators, *Sequential Floating Forward Search* (SFFS) and *Fisher Projection* for dimensional reduction (see [Vyzas and Picard 1998] and [Section 3.1]). Since each point in the space corresponds to a labeled instance of a deliberately expressed emotion, the problem can be framed as one of *supervised* classification, for

—measuring the overall level of physiological activation during an emotional episode, such as sustained muscle tension or elevated skin conductance; (2) *Standard deviation of the raw signal*:

$$\sigma^i = \left(\frac{1}{N-1} \sum_{n=1}^{N} \left(S_n^i - \mu^i \right)^2 \right)^{1/2}$$

—representing the degree of fluctuation around that level, which may indicate instability, agitation, or physiological lability; (3) *Mean of the absolute first difference of the raw signal*:

$$\delta_1^i = \frac{1}{N-1} \sum_{n=1}^{N-1} \left| S_{n+1}^i - S_n^i \right|$$

—capturing the speed at which changes occur in the signal, reflecting how rapidly the body responds or reacts; (4) *Mean of the absolute second difference of the raw signal*:

$$\tilde{\delta}_1^i = \frac{1}{N-1} \sum_{n=1}^{N-1} \left| \tilde{S}_{n+1}^i - \tilde{S}_n^i \right| = \frac{\delta_1^i}{\sigma^i}$$

—indicating the abruptness or irregularity of those changes (i.e., how turbulent the response is); (5) *Mean of the absolute first difference of the normalized signal*:

$$\delta_2^i = \frac{1}{N-2} \sum_{n=1}^{N-2} \left| S_{n+2}^i - S_n^i \right|$$

—an absolute measure of signal speed, allowing comparisons across different days; and (6) *Mean of the absolute second difference of the normalized signal*:

$$\tilde{\delta}_2^i = \frac{1}{N-2} \sum_{n=1}^{N-2} \left| \tilde{S}_{n+2}^i - \tilde{S}_n^i \right| = \frac{\delta_2^i}{\sigma^i}$$

—highlighting sudden and irregular shifts in the signal, independently of its overall strength or amplitude [Vyzas and Picard 1998, 175-176].

the system learns to associate regions of the feature space with particular emotional categories given in advance and not emerging directly from the data. By this procedure, the system will thus know what the typical physiological patterns are for each emotion, being later able to classify new raw inputs outside the original dataset. Of course, as evident, classification does not depend here on pre-programmed, explicit symbolic rules. Similarly, the ML algorithms in the system operate on continuous numeric representations of physiological states and not by manipulating discrete symbolic knowledge related to emotional categories. In other words, this system exactly exemplifies the kind of subsymbolic methodology that began to gain prominence in CME during the 1990s[57] [Section 3.1].

4.2 New AC Methodologies in CME: An Overview

As seen in the previous [Section 4.1], Picard [1997] marked a revolutionary moment in the history of artificial emotions. Her AC proposal systematized previous CME efforts, often isolated or just exploratory, into a new, organic and coherent scientific paradigm and programmatic research framework, integrating it with novel psychological and philosophical insights from cognitive sciences and with advanced techniques in computation. But most importantly, she introduced a bold new conception of "machine intelligence": Designing cognitively efficient machines requires forms of *computing* which are intrinsically and from the outset *affective*, whereby artificial emotions too need to be embodied. Only this kind of *affective artificial intelligence*, in fact, can enable machines to properly engage in human social environments, whereby exactly emotions are what can mediate interaction, adaptability and understanding.

[57] Other examples of computational techniques for emotional pattern recognition concern probabilistic models such as *Hidden Markov Models* (HMMs) —which describe temporal processes through *Markov chains*, where the current state depends only on the immediately preceding one. In such models, the underlying (*hidden*) emotional state cannot be directly observed but can be inferred from observable physiological or behavioral signals. This approach allows the system to capture the sequential and dynamic nature of emotions, and was among the first probabilistic methods adopted in AC [Picard 1997, Ch. 6].

In the wake of this proposal, the field of computational modelling of emotion underwent a profound transformation. Of course, already some late pre-AC CMEs (e.g., FEELER) featured elements linking emotion to cognition, temporality, embodiedness, sociality etc. —no conceptual change, after all, originates completely *ex nihilo* (see e.g., [Chang 2007]). It was however only from the mid-1990s onwards, that a wide range of CMEs began to emerge which, despite the functional heterogeneity, aligned with AC and explicitly shared its core technical and epistemological premises. A (partial) list of AC-influenced CMEs developed between the 1990s and the 2000s is given in [Table 4.1] (see also [Marsella et al. 2010; Lin et al. 2012; Kowalczuk and Czubenko 2016]).

Name	**Year**	**Task Domain**	**Typology**
AR (Affective Reasoner) [Elliott 1992]	1992	*Simulation of emotional reasoning and interaction in agents.*	S
OZ [Bates 1992]	1992	*Real-time emotionally expressive agents in interactive fiction, using behavior sequencing, motivational models, and expressive channels.*	H (S, C)
NML1 [Beaudoin 1994]	1994	*Goal management in an autonomous agent designed for a "nursemaid" scenario, using procedural reasoning and complex scheduling.*	H (S, C)
WILL [Moffat and Frijda 1994]	1994	*Emotion-based decision-making architecture.*	H (S, C)
CATHEXIS [Velásquez 1996]	1996	*Emotion generation and modulation of behavior in autonomous agents using affect programs and motivational systems.*	AC/H (S, SS:ANN*, C)
Armony [Armony et al. 1997]	1997	*Model of fear conditioning, simulating thalamo-amygdala and cortical pathways based on empirical data.*	H (SS:ANN-ML, C)

SILAS [Blumberg 1997]	1997	*Virtual dog agent displaying lifelike behavior through ethologically inspired motivation, including adaptation, intentionality, and learning in real-time interactive environments.*	H (SS:ML*, C)
VP [Vyzas and Picard 1998] (see also [Healey and Picard 1998; Vyzas and Picard 1999; Picard et al. 2001; Klein et al. 2002; Scheirer et al. 2002], etc.)	1997	*Training computers on multimodal emotion recognition from physiological signals collected through wearables.*	AC/H (SS:ML*, C)
TABASCO [Staller and Petta 2001]	1998	*Tractable architecture for situated agents with emotion-based reactive and deliberative behavior using layered appraisal processes.*	H (SS, C)
EMILE [Gratch 2000]	2000	*Emotion-driven reasoning and planning for affective agents in educational applications, using plan-based appraisal.*	AC/H (S, C)
FLAME [El-Nasr and Seif 2000]	2000	*Adaptive model for emotion generation, learning, and behavior selection in artificial pets and affective agents based on fuzzy logic.*	AC/H (SS:ML, C)
GL [Gmytrasiewicz and Lisetti 2000]	2000	*Emotion-based modulation of decision-theoretic reasoning in intelligent agents using expected utility.*	AC/S
ACT-R ext. [Belavkin 2001]	2001	*Simulation of emotional influence on problem-solving using ACT-R with stochastic optimization and motivational control.*	AC/H (S, SS:ML, C)
H-CogAff [Sloman 2001]	2001	*General architecture of the human mind supporting primary, secondary, and tertiary emotions via layered cognitive processing.*	AC/H (S, SS, C)

ParleE [Bui 2002]	2002	*Emotional speech recognition and generation in conversational agents with Bayesian belief modeling.*	H (S, SS:ML, C)
ALEC [Gadanho and Custódio 2003]	2003	*Emotion- and cognition-driven learning in robots using both reinforcement learning and rule extraction to improve adaptation in real-world tasks.*	H (S, SS:ANN-ML, C)
CBI [Marsella 2003]	2003	*Emotion-sensitive pedagogical agents used in interactive drama to teach coping skills, specifically targeting health-related interventions like pediatric cancer.*	S
ACT-R ext. [Fum and Stocco 2004]	2004	*ACT-R extension integrating emotion to modulate cognitive functions like memory recall and reasoning under pressure.*	H (S, SS:ML*)
EMA [Gratch and Marsella 2004]	2004	*Real-time emotion generation in conversational agents using dynamic appraisal, coping, mood, and reappraisal in a domain-independent setting.*	S
ALMA [Gebhard 2005]	2005	*Affective model for virtual agents incorporating transient emotions, mood, and stable personality traits using a three-layer structure.*	H (S, C)
EM-ONE [Singh 2005]	2005	*Common-sense reasoning and affective behavior in agents using a six-layered cognitive architecture inspired by* [Minsky 2006].	H (S, C)
FearNot! [Aylett et al. 2005]	2005	"Fun with Empathic Agents Reaching Novel Outcomes in Teaching": *Emergent narrative and empathic interaction in educational virtual drama for anti-bullying applications with affective agents*	H (S, C)

MAMID [Hudlicka 2005]	2005	*Simulation of cognitive-affective processes in decision-making agents for crisis scenarios, with modules for emotion, appraisal, and behavior generation.*	H (S, C)
ACT-R ext. [Cochran et al. 2006]	2006	*Integrates emotion with ACT-R to model user frustration and adaptive interaction in human-computer systems.*	H (S, SS)
EM [Reilly 2006]	2006	*Real-time emotion modeling in believable characters using OCC-based appraisal to drive expressive and adaptive behaviors.*	AC/H (S, C)
MS [Malfaz and Salichs 2006]	2006	*Architecture for robot decision-making based on internal drives, motivational states, and basic emotions.*	AC/H (SS:ML-RL, C)
Meyer [Meyer 2006]	2006	*Emotion-based reasoning model integrating beliefs, desires, and intentions in agent-oriented programming.*	H (S, SS)
THESPIAN [Mei et al. 2006]	2006	*Interactive drama framework with autonomous agents employing belief modeling and POMDP-based planning for emotionally coherent character interaction.*	H (SS:ML*, C)
AGC [Cãnamero and Avila-García 2007]	2007	*Emotion modeling in autonomous agents via hormone-like homeostatic regulation affecting adaptive behavior.*	AC/H (SS, C)
Soar-EMOTE [Marinier and Laird 2008]	2008	*Extension of Soar architecture integrating emotion-influenced decision-making using PEACTIDM model in time-pressured cognitive tasks.*	AC/H (S, SS:ML)
WASABI [Becker-Asano 2008]	2008	*Affective architecture for autonomous agents using a dimensional PAD model to generate real-time emotional states and facial expressions.*	AC/H (S, SS, C)

Soar-PEACTDM [Marinier et al. 2009]	2009	*Integration of emotion mechanisms into Soar cognitive architecture to influence attention and memory in a simulated pilot.*	AC/H (S, SS, C)
ActAffAct [Rank 2009]	2009	*Emotion-based action selection in synthetic agents for interactive narrative, integrating appraisal and behavioral influence mechanisms.*	AC/H (SS, C)
Ahn [Ahn 2010]	2010	*Modeling affect-influenced human decision-making under uncertainty using an affective-cognitive framework combining Prospect Theory and Reinforcement Learning.*	AC/H (SS:ML-RL)

Table 4.1: Partial list of computational models of emotions developed in the 1990s-2000s. Year, summary of task domains, and typology are indicated. AC: The model explicitly builds upon the programmatic ideas and technical guidelines of Picard's [1997] *Affective Computing*; S: Symbolic model; H: Hybrid model (involving symbolic and/or cybernetic [Section 2.1] and/or subsymbolic components); SS: Subsymbolic model; ANN: The model involves connectionist neural networks; ANN*: The model implements ANN-like mechanisms without using actual neural networks; ML: The model involves ANN-based machine learning techniques; ML*: The model involves other, non-ANN-based machine learning techniques; RL: The model involves reinforcement learning. The vast majority of the models has been implemented as a computer software or as a formal cognitive architecture for robotic systems. They are both concerned with emotion recognition and generation tasks. Clearly, all models are AIME in the sense of [Chapter 2].

Far from being isolated technical solutions, these models, taken together, build up a historically and conceptually coherent configuration that illustrates how the establishment of the AC paradigm contributed to redefining both the function and architecture of artificial emotions.

Differently from early AAIMCOs [Chapter 2], emotions in post-AC CMEs are not functional states or regulatory mechanisms that just *augment* the models as to their performance in purely cognitive tasks. Instead, similarly to late AIMEs, in them emotions amount to a primary computational task, being this related to *emotion recognition* (e.g., VP — see [Section 4.1.2.3]) and/or *emotion synthesis* or *generation* (see also [Picard 1997, Ch. 7]). In this respect, affective states are designed to support socially meaningful behavior, acting as mediators of interaction, ne-

gotiation, and alignment between agents and human users (e.g., FearNot!, CBI). For the same reason, they are modeled as online, temporal phenomena, unfolding across time, shaped by mood dynamics (e.g., ALMA, EMA, etc.), memory (e.g., ACT-R ext., Soar-PEACTIDM, etc.) and adaptive learning processes (e.g., ALEC, MS, etc.).

Now, the need to implement emotions as integrated, dynamic mechanisms within complex architectures, rather than as separate modules influencing cognition or behavior, is also reflected in the substantial diversification of computational strategies adopted by these models. The vast majority of them, in fact, rely on *hybrid* computational frameworks, ranging from symbolic and rule-based architectures —e.g., for representing goals, plans, evaluations, decisions, etc.— to cybernetic approaches —using control-based feedback loops and homeostatic regulation to manage reactive and adaptive responses through the continuous adjustment of numerical variables representing levels of physiological arousal, intensity, activations, thresholds, etc. [Chapter 2]. As already in AIMEs, these two computational layers are often functionally distributed, as modules or subroutines, within the same model: While symbolic components are employed for high-level interpretative and planning processes, cybernetic mechanisms handle the adaptation of the system to new environmental inputs.

4.2.1 Early Connectionism in CME: The Case of CATHEXIS

As a reflection of 1980s advancements in AI, however, most AC and post-AC CME also involve subsymbolic modules, or even complete subsymbolic architectures, that do not thus rely on explicit rules and representations for recognizing or generating emotions. Some of them, for example, use systems based on fuzzy logic to generate graded affective responses based on imprecise inputs (e.g., FLAME), or implement emotion through hormone-like diffusion mechanisms and internal drives (e.g., MS), enabling continuous modulation without relying on discrete symbolic categories. In others, especially toward the end of the decade, some early, albeit constrained, experiments appear of ANNs [Chapter 3] ap-

plied to CME, offering a flexible and biologically inspired methodology for modeling artificially affects as highly dynamic and adaptive processes.

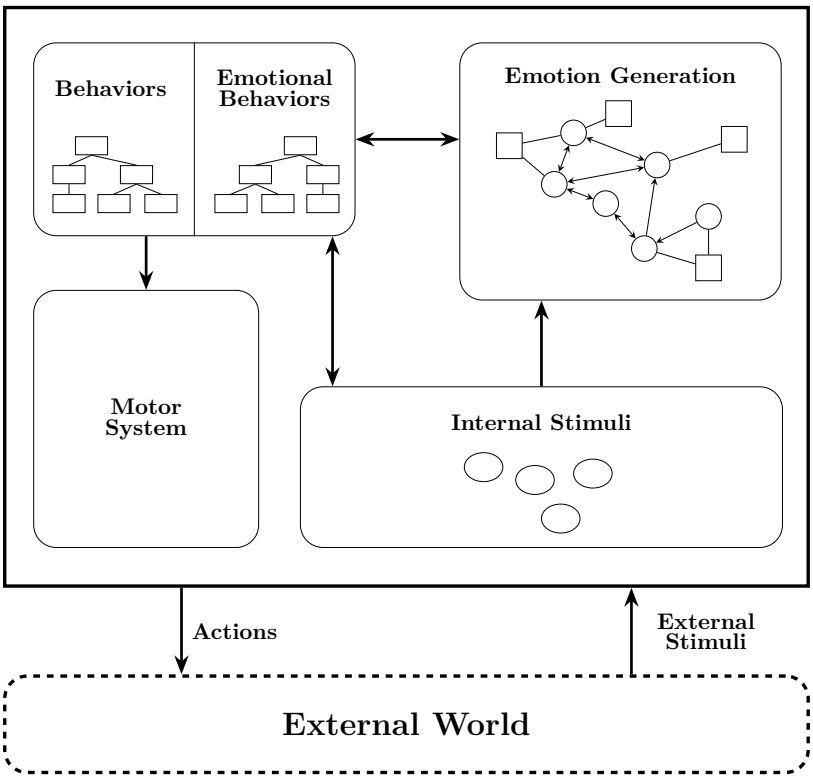


Figure 4.3: High-level view of CATHEXIS' architecture —adapted from [Velásquez 1996, 38].

One of the earliest examples of CME beginning to incorporate primitive ANN-like features into affective processing is Velásquez's [1996] CATHEXIS[58] (see also [Velásquez and Maes 1997; Velásquez 1997] and [Picard 1997, Ch. 7]). CATHEXIS (from the Greek κάθεξις, *cāthexis*, meaning "retention", see [Footnote 30]) is a psychologically, ethologically, and neurobiologically inspired model for emotion generation, which also accounts for the influence of affect on behavior (and vice versa). This is directly reflected in its architecture [Figure 4.3], which is primarily composed of an *Emotion Generation System* (EGS) and an *Emotional Behavior System* (EBS), tightly interdependent between each other.

[58]From a historical point of view, it is interesting to note that Picard, Vyzas, Velasquez, Healey, Maes and other AC pioneers were or still are affiliated scientists of MIT Media Lab. This center has thus been an important institutional booster of the AC revolution during the early 2000s.

As a theoretical basis for modeling emotions computationally, it explicitly takes basic emotion theory in the formulation of Johnson-Laird and Oatley [1989; 1998], Izard [1991], Ekman [1992], Plutchik [1994], etc. [Section 3.3][59], integrating it with Minsky's [1985] notion of *mental proto-specialists* —see [Figure 4.4].

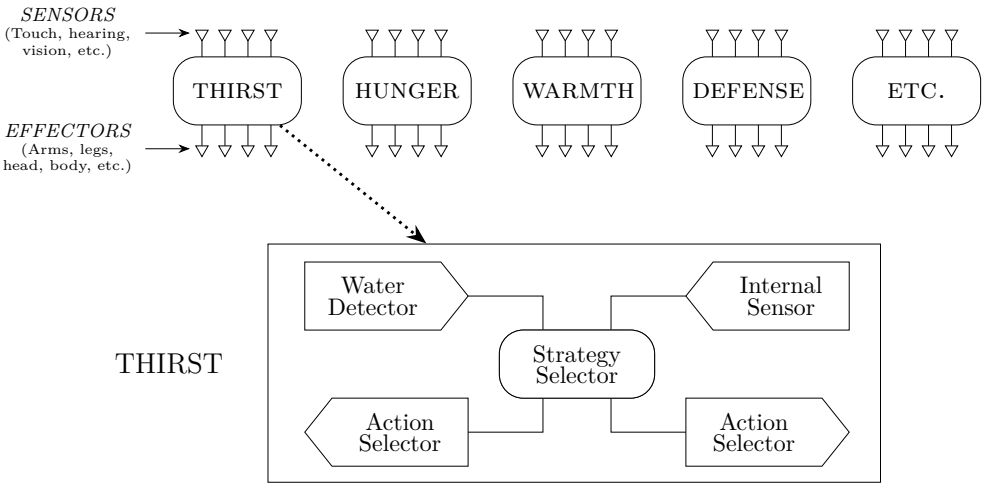


Figure 4.4: Examples of proto-specialists —adapted from [Minsky 1985, 165].

Proto-specialists are primitive and simple mental mechanism dedicated to handling very specific basic needs (e.g., thirst) by connecting certain sensory (e.g., sight of water) and internal (e.g., dry mouth) inputs to specific actions (e.g., reaching the water and drinking). These *agents* act like *little autonomic minds* that are focused only on their own task. They are hardwired in the brain and share the same resources as to bodily sensors and effectors. Initially independent at birth, they begin to interact as the development progresses, forming the basis for ever more complex cognitive and behavioral schemes. This means, however, that in the adult they must compete or cooperate to control the system.

Directly from this, Velásquez [1996, 48ff.] derives a conception of affective life (emotions, moods, temperaments, etc.) as a complex *network* of different, semi-autonomous and interacting modules —*Emotion*

[59] In particular, Velásquez [1996] refers to later formulations of the theory that treat basic emotions as *emotions families* rather than single, well-defined affective states, whereby members of the same family share similar characteristics as to eliciting conditions, expressive and behavioral consequences, patterns of physiological activation, etc. —see also [Footnote 45].

Proto-Specialists (EPSs)— each representing exactly one of six basic emotions —*Anger*, *Fear*, *Distress/Sadness*, *Enjoyment/Happiness*, *Disgust*, and *Surprise*. Depending on wether specific eliciting conditions are met, they "fire", generating the correspondent emotion and producing an appropriate behavioral outputs.

In EGS, such conditions are organized into four distinct categories of both cognitive and non-cognitive *Elicitors* EL_i [Roseman et al. 1990; Izard 1993]: *Neural* ($i = 1$) —e.g., neurotransmitters, hormones, drugs, electrical and chemical brain stimuli, etc.; *Sensorimotor* ($i = 2$) —e.g., nervous stimuli, muscle activity, body posture, facial expressions, etc.; *Motivational* ($i = 3$) —e.g., drives like thirst or hunger, pain, other emotions themselves, etc.; and *Cognitive* ($i = 4$) —e.g., evaluations, beliefs, memory, etc.. In the model, correspondingly, four kinds of sensors monitor «both external (e.g., events in the environment) and internal sensory stimuli (e.g., drive levels, feedback from sensorimotor processes) for the existence of the appropriate conditions that would elicit the emotion represented by that particular [emotion] proto-specialist» [Velásquez 1996, 48-49].

Elicitors EL_i have themselves *Eliciting Conditions* EC_k that determine and contribute to their activation. EC_k are specific to *Elicitor* i —e.g., probability or controllability of an event for the cognitive elicitor; the particular motivation for the motivational, etc. In general, the value of *Elicitor* i at time t is given by

$$EL_i^t = \sum_k EC_i^k \tag{4.3}$$

Of course not all EC_k have the same *weight*, which varies depending on both k itself and the *Elicitor* they are associated with. In general, the value of *Eliciting Condition* i at time t is given by the product of its value v_i and its weight w_i:

$$EC_i^t = v_1 \cdot w_i \tag{4.4}$$

Elicitors cause an EPS to fire, producing an emotion whose intensity is proportional to the value of its elicitors. The *Intensity* of an *Emotion* i at time t (EI_i^t), however, also depends on regulatory mecha-

nisms that determine the *activation* and *saturation* of the ESP, as well as the *duration* of the emotion produced —aspects mathematically modeled in CATHEXIS under the direct influence of [Picard 1997] (see [Section 4.1.2.2]). The activation threshold α_i for *Emotion* i defines the lower bound above which inputs from the elicitors contribute to the intensity of the generated affective state, successively inhibiting or exciting other ESPs and eventually released to EBS for behavior generation [Figure 4.5]. In general, the output O_i^t of an EPS for emotion i at time t is

$$O_{it} = \begin{cases} EI_i^t, & \text{if} \quad (EI_i^t > \alpha_i) \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

On the other hand, the *saturation threshold* ω_i specifies for each EPS the maximum intensity beyond which additional input has no further effect on the intensity of the *Emotion* i.

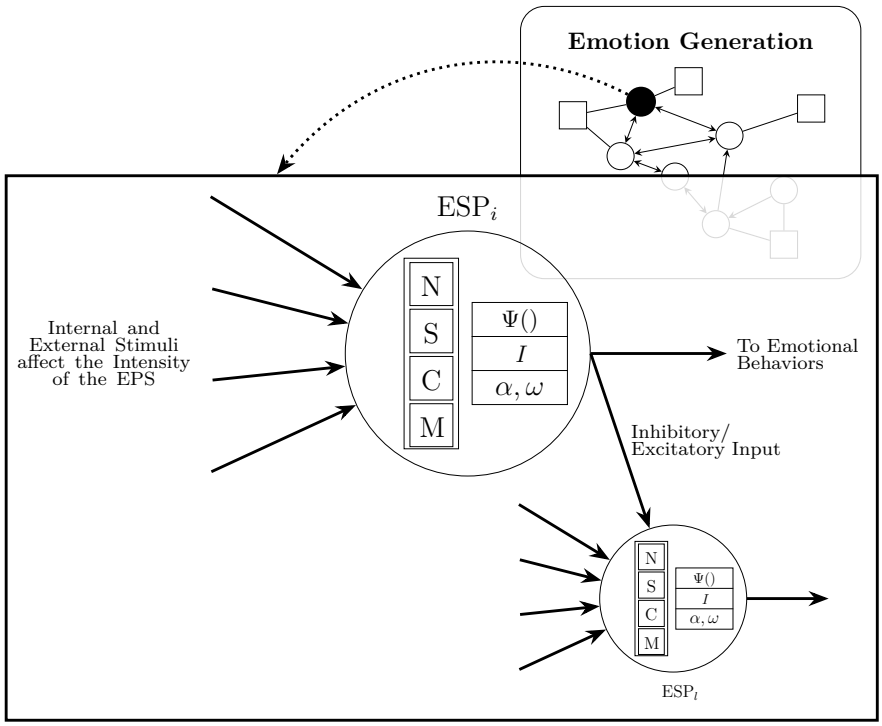


Figure 4.5: CATHEXIS' *Emotion Generation System* and EPS — adapted from [Velásquez 1996, 50].

These thresholds can be adjusted to model individual differences in mood and temperament, with lower α values increasing emotional sensitivity,

and higher ω values allowing for more intense responses. Finally, a decay function $\Psi()$ represents the progressive reduction of the emotion's intensity over time in the absence of new sustaining input, modeling the natural fading of affective states. At this point, *Emotion Intensity* can be formally modeled as

$$EI_i^t = \chi\left(\Psi(EI_i^{(t-1)}) + \sum_k EL_i^k + \sum_l (EX_i^l \cdot EI_l^t) - \sum_m (IN_i^m \cdot EI_m^t), 0, \omega_i\right) \quad (4.6)$$

where EI_i^t is the Intensity of the *Emotion* i at time t; $\Psi()$ the decay of *Emotion* i; EL_i^k the contribution from *Elicitor* k to the activation of *Emotion* i; EX_i^l the *Excitatory Input* from *Emotion* l to *Emotion* i; EI_l^t the *Intensity* of *Emotion* l at time t; IN_i^m the *Inhibitory Input* from *Emotion* m to *Emotion* i; EI_m^t the *Intensity* of *Emotion* m at time t; and $\chi()$ a function used to clamp the intensity of *Emotion* i within $[0, \omega]$.

Now, once produced, EPS output O_i^t with intensity EI_i^t [Equation 4.4] is fed in into EBS for the generation of behaviors. In particular, it functions as a *Behavior-Releaser* (BR), capturing the agent's emotional state at a specific time, which the EBS then uses to select the most fitting *Emotional Behavior* (EB) for the agent (e.g., "withdraw from conflict", "lower gaze to avoid tension", etc.). EBs have two components: an *Expressive* one (responsible for producing observable affective outputs such as facial expressions, body posture, and vocalizations) and an *Experiential* one (in charge of updating the agent's internal state by modifying motivational variables or emotional conditions over time) [Figure 4.6].

There are several EBs which compete for controlling the behavior of the agent. EBS must thus choose among them, relying on the value of each EB, recalculated on every EGS-EBS cycle: The EB with maximum value controls the agent. In general, EB values depend on BRs —including, besides emotions (EPS outputs), also motivations (moods, drives, pain, etc.) and multimodal stimuli from the external world [Figure 4.3]:

$$EB_i^t = \sum_k BR_i^k \quad (4.7)$$

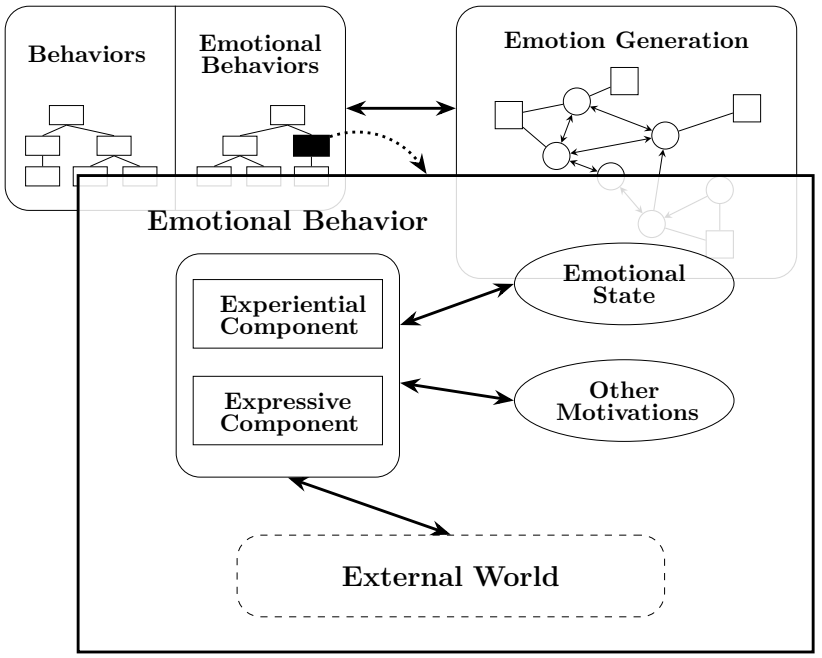


Figure 4.6: CATHEXIS' *Emotional Behaviors System* —adapted from [Velásquez 1996, 57].

where EB_i^t is the value of *Emotional Behavior* i at time t and BR_i^k the value of *Behavior Releasers* k relevant to *Emotional Behavior* i. As for EC, the value of BRs depends on its specific nature and to the EB they are associated with (for example, if the releaser corresponds to an internal state —such as a drive like hunger or an active emotional condition— its value is typically a function of the underlying motivational intensity; Conversely, if the releaser is linked to an external stimulus —e.g., a loud sound or a bright light—its value will be proportional to the intensity of the corresponding sensory input).

CATHEXIS has also been implemented as an OOP-based computer program written in C++ and Java [Velásquez 1996, Ch. 4]. Each EPS is coded as an autonomous module, maintaining its own internal variables to handle intensity, activation and saturation thresholds, and temporal decay dynamics in the way described above. Excitatory and inhibitory connections between EPSs are implemented as weight matrices, enabling the emergence of complex emotional patterns through simple local interactions. EGS module influence EBS, which controls the output behavior of a simulated agent, "Simón the toddler" [Figure 4.7], who is able to

display coherent emotional reactions through continuous real-time computation of its basic emotion activations.

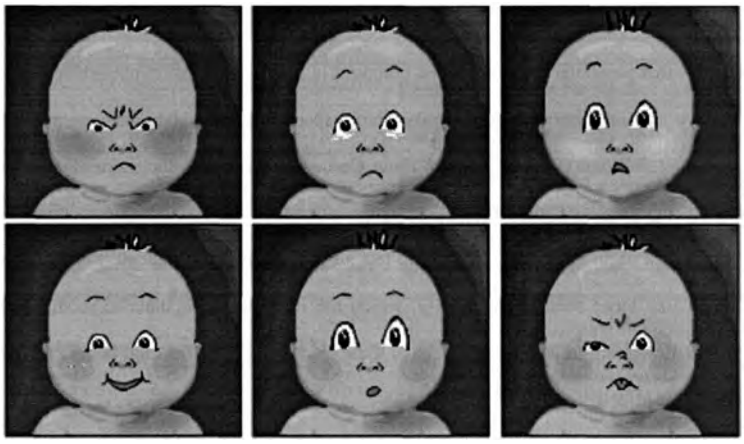


Figure 4.7: Facial expressions for different Emotional Behaviors in "Simón the toddler" CATHEXIS' virtual agent —adapted from [Velásquez 1996, 81].

The model's modules operate *in parallel*, each processing relevant stimuli and contributing to behavioral responses independently. This parallelism marks a crucial departure from earlier CME systems like FEELER [Section 2.3], where emotional knowledge was encoded uniformly in IF-THEN production rules, without differentiation among distinct affective processes. In CATHEXIS, by contrast, there is no centralized emotion engine, no supervisory affective controller. The system's behavior instead emerges from the dynamic and *distributed* interplay of its modular components: Multiple EPSs may be co-activated, giving rise to a rich web of mutual influences —excitatory or inhibitory, reinforcing or suppressive— whose temporal interactions can yield complex affective phenomena —including blends (e.g., *Grief*, *Jealousy*) and mixed emotions (e.g., *Schadenfreude*)— and complex behaviors, in turn influencing emotion generation. This computational integration between emotion and behavior —whereby emotion is not treated as a mere discrete functional state like in previous CMEs but as a dynamic process— directly aligns CATHEXIS with AC's research program, as to the particular emphasis given to the fundamentally interactive and situated nature of affect.

In the model, in fact, emergent affective states are not the result of rule-based deductions, but arise from the system's distributed architecture itself —an approach that had in those years also a counterpart

in robotics [Section 5.2]. This clearly echoes relevant connectionist intuitions about subsymbolic and pattern-sensitive computation [Picard 1997, 209] that computer scientists were beginning to apply also to the artificial modeling of emotions at the end of the 1990s. In this respect, CATHEXIS probably represents the first example of a CME applying the core ideas of ANNs-based computation [Section 3.1] to emotions —albeit not yet in the form of an explicit ANN model built on backpropagation, learning, etc. which will be done later (see e.g., [Armony et al. 1997]): Each EPS namely functions analogously to a neuron, receiving inputs from multiple elicitors or other EPSs, computing its activation based on weighted influences and thresholds, and propagating outputs that excite or inhibit other EPSs modules. Its excitatory and inhibitory connections, encoded as weight matrices, instantiate a networked structure of local processing units whose collective dynamics yields emergent global states and complex (non-representative and continuous) knowledge about emotions and emotional behaviors.

4.2.2 Epistemic Gaps: Cognition and Physiology

As seen in the previous section, computer scientists began to concretize the proposal of AC soon after the publication of Picard's [1997] seminal book. Several models were developed that integrated emotion, adaptation, and learning within complex and distributed subsymbolic architectures, considering affectivity as an essential component of a new conception of AI: embodied, socially grounded, and interactionally meaningful —a fact which is particularly evident in the case of AC-inspired social robotics, as we will see in [Chapter 5].

Following early cognitive science and early CME (see [Chapter 1] and [Chapter 2]), these models also gave a concrete computational form to an idea that, after long debate, had by then become widely accepted across the scientific community: That emotion is fundamentally intertwined with cognition, and that each depends on the other to operate effectively. Of course, as seen above, this is a point that Picard [1997, 194, my italics] herself had repeatedly highlighted, with a special emphasis on its social consequences:

> «One agent may act like an office assistant to help you process mail [...]. The assistant that cannot read your emotional expression, *reason*

> about what your emotions might be, and *learn* what is important to you —when not to interrupt, for example— will act *unintelligently*. If the agent cannot have a mechanism for the equivalent of "feeling bad" for causing you distress, then it is likely to repeat this behavior. An ability to "feel good or bad" does not merely effect the agent's ability to learn, but helps it prioritize and *choose* among all its actions —*learning*, *planning*, *decision-making*, and more.

Correspondingly, she considered cognition to be one of the *five fundamental criteria* for building an *emotion system* —criteria that, if (partially or fully[60]) implemented, would allow one to legitimately say that a computer, a program, or a robot[61] *have emotions*[62] [Picard 1997, 70-71 and 193]:

1. *Emotional behavior*
   System has behavior that appears to arise from emotions;
2. *Fast primary emotions*
   System has fast "primary" emotional responses to certain inputs;
3. *Cognitively generated emotions*
   System can cognitively generate emotions, by reasoning about situations, especially as they concern its goals, standards, preferences, and expectations;
4. *Emotional experience*
   System can have an emotional experience, specifically:
   - Cognitive awareness,
   - Physiological awareness,

[60] It is important to note that such criteria are not necessary conditions for a system to be affective. AC namely operates on a *graded* and *context-dependent* logic: Some machines may require emotional capacities to function appropriately, others may not; In some cases, emotions are relevant only in certain contexts or to varying degrees. This flexible and situated approach reflects the pragmatic orientation of the field, which prioritizes functionality over strict ontological commitments [Picard 1997, 68, 193, 249, etc.]

[61] Throughout her work, Picard [1997] uses the term "computer" in a broad and inclusive sense, referring not only to traditional computing machines, but also to a wide range of artificial systems, including software programs, robotic agents, animated characters, and other synthetic entities capable of perceiving, processing, or displaying affective states. This extended usage reflects the interdisciplinary and application-oriented nature of AC, which spans from purely virtual agents to embodied robotics —as it will be described in [Chapter 5].

[62] Although Picard does raise the question of what it might mean for a machine to "have emotions", she does not engage with it in sustained philosophical depth. Rather than offering a rigorous conceptual analysis, she proposes five operational criteria that —if implemented— would justify such a claim. However, these criteria are not always immune to classical objections from the philosophy of mind and emotion, nor do they show substantial awareness of the ongoing debates around the nature of affective experience, intentionality, or the distinction between simulating and instantiating emotional states.

- Subjective feelings;

5. *Body-mind interactions*
   The system's emotions interact with other processes that imitate human cognitive and physical functions, for example:
   - Memory,
   - Perception,
   - Decision making,
   - Learning,
   - Concerns,
   - Goals,
   - Motivations,
   - Attention,
   - Interest,
   - Prioritizing,
   - Planning,
   - Sentic modulation,
   - Immune system functions,
   - Regulatory mechanisms.

At the same time, however, Picard [1997, 194] also complained that, in practice, the general tendency in the development of CMEs continued to be one prioritizing only the cognitive criterion (3) at the expense of the others:

> «The emotions implemented [in agents] today are primarily cognitively generated, the third component only [...]. The agents have some simple cognitive emotions, and they can usually express these emotions, but they do not have the ability to recognize the emotions of people, to experience or show empathy, or to benefit internally from the functions that emotions can provide».

Indeed, if one looks at the history of CMEs both before [Chapter 2] and after [Section 4.2] AC, one cannot but actually observe this persistent imbalance. In AAIMCO [Section 2.2], affective components were primarily conceived as modulators of cognition. In AIME [Section 2.3], early attempts to simulate emotional states through reactive or rule-based modules began to incorporate minimal embodiment or contextual responsiveness, yet the focus remained largely on the cognitive dimension of emotion. Similarly, in AC and post-AC CMEs, despite a growing attention to the social and bodily factors influencing emotions —as well as a clear recognition of new interrelational motivations behind the development of CMEs— the core architecture of most models was still derived

from computational renderings of classical cognitivist, appraisal and rational theories of emotions ([Schachter and Singer 1962], [Lazarus 1984], [Scherer 1984; 1999], [Frijda 1986; 1988], [Ortony et al. 1987], etc.) —see [Figure 4.8].

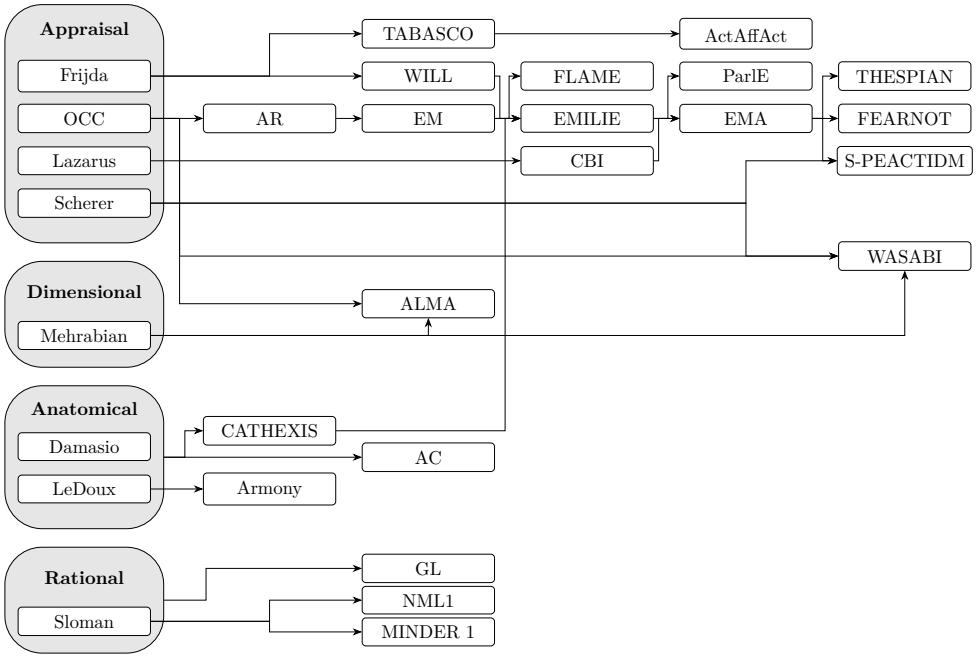


Figure 4.8: Theoretical derivation (appraisal, dimensional, anatomical and rational theories of emotion [Chapter 3]) of most AC and post-AC CMEs [Table 4.1] —adapted from [Marsella et al. 2010, 24][63]. As it can be easily noticed, most models developed from cognitivists approaches to emotions.

In such theories, emotions were predominantly conceived in inference-driven terms: Affective states emerge from [Lazarus 1984; Scherer 1999]

[63]The diagram is partial: It does not contain information about all models of emotions in psychology, philosophy, medicine, etc. or about all CMEs developed in the 1990s-2000s; nor it accounts for cross-derivations of some models —e.g., CATHEXIS [Section 4.2.1], based on multiple theoretical approaches to emotions, cognitivist, and non-cognitivist). Interestingly enough, however, Marsella et al. [2010, 24] title it «A history of computational models of emotion». Of course, as pointed out in [Chapter 1] —and as emerging from the entire [Part I] of the present work— doing a history of CME requires broader conceptual categories, reference to a larger time span, and methodological tools from the (Philosophy and) History of Science that the authors do not seem to take into account. Yet, Marsella et al. [2010] is one of the extremely few works trying to approach AC and post-AC CMEs from a historical point of view.

—or even consist in [Ortony et al. 1987; Frijda 1988]— *evaluative judgments*, i.e., *appraisal processes* whereby agents interpret the significance of events by relating them to their beliefs about the world, their current goals, and the confirmation or violation of their expectations. This "logic of emotions" obviously found a direct and technically straightforward implementation in CMEs: Adequate (and typically symbolic) computational mechanisms were designed to artificially reproduce the process of appraisal, by mapping the agent's internal representations of beliefs, goals, and expectations onto emotional categories. Of course, in this respect, the central modelling task became that of *emotion generation* [Section 4.1]: enabling agents to produce affective states appropriate to their appraised situation, and to express them in ways that aligned with their reasoning structures. Other dimensions of affectivity —such as bodily dynamics, perceptual receptivity, emotional attunement, situated interaction— tended to remain underdeveloped, often deferred to peripheral modules or treated as implementation details.

On the contrary, although the cognitive dimension of (artificial) emotions remained central to her proposal, Picard [1997] stressed the importance of complementing it with other components of affectivity, especially those contributing to a possible augmentation of the empathetic and interactional abilities of machines. In particular, as discussed in [Section 4.1], the traditional focus of CMEs on emotion generation had to be balanced by a corresponding attention to *affect recognition*, as well as —aligning with EC— to the simulation of *body–mind interactions* involved emotional processes: Sentic modulations, hormonal signals, metabolic states, and even immune responses were all considered relevant factors to implement in systems for an optimal modeling of artificial recognition and expression of emotions [Picard 1997, 66–68].

Indeed, this new approach to CME, systematized in the AC proposal, found a first concrete realization especially in the models developed by Picard herself and her collaborators at MIT Media Lab in the late 1990s (see [Section 4.1.1], [Section 4.1.2.3], [Table 4.1] and [Footnote 58]). Velásquez's [1996], Vyzas's [1998] or Healey's [1998] CMEs, for example, were based on a complex conception of emotion, involving both cognitive and non-cognitive aspects. A multimodal approach was commonly employed in both affect generation and, more prominently, affect

recognition, reflecting the view that emotions are expressed and perceived through coordinated physiological and behavioral signals. Particular emphasis was thus placed in such models on the *physiological* dimensions of affective processes, treated as primary sources for inferring or displaying emotional states.

Now, at closer look, the focus on the bodily components of affectivity, while broadening the general scope, accuracy, and efficiency of cognitivist CMEs, implicitly led to a conception of emotion within AI and computer science that began to tend toward a form of *physiologic-behavioral reductionism*. Notwithstanding the rich theoretical frameworks from which they developed, and despite the ambitious premises they had —integration of affect and cognition, social embeddedness of artificial emotions, etc.— models like CATHEXIS or VP, with their emphasis on internal drives, autonomic responses, and low-level expressive signals, subtly ended up reinforcing the idea that, at least at a functional level, emotions can be computationally encoded and decoded primarily through structured patterns of physiological activation and observable behavior.

As pointed out above, however, AC's stress on physiology and behavior —even with some tendencies towards a latent functional reductionism— remained relatively marginal within the mainstream trajectory of CME, particularly during the early phases of consolidation of AC as a research field. In fact, until at least the early 2010s (see [Section 4.3] and [Section 6.1]), affective cognitivism continued to serve as the dominant conceptual framework for most CMEs. Far from being accidental, this *epistemic gap* between cognition and physiology in 1990s-2000s CMEs emerged from a composite set of historical, epistemological and disciplinary reasons.

Firstly, as mentioned before, the "logic of emotions" advanced by classical appraisal theories had the advantage of being conceptually rigorous, and therefore relatively easy to implement within symbolic or hybrid rule-based systems. These theories offered clear and directly operationalizable affective ontologies —beliefs, goals, expectations, plans, emotions, etc.— as well as formally defined sets of inference rules enabling artificial agents to simulate emotional states without requiring complex integrations of perceptual or bodily processes. In this sense, the dom-

inance of cognitivism in CME was basically a consequence of practical and technical-computational constraints. Secondly, still at the earliest stages of development of the subsymbolic paradigm, AI remained heavily shaped by representationalist frameworks [Section 3.1], which continued to provide both the educational scaffolding for researchers in the field [Planck 1948], and the theoretical basis for the most prominent emotion models accepted within CME community (e.g., OCC). Also, standard ontologies, software libraries, and computational frameworks rooted in those theories were repeatedly employed over time —and this not merely for reasons of convenience, but to ensure a shared scientific grounding, criteria of empirical testability, reproducibility and interpretability, pedagogical continuity, and institutional-academic legitimacy. Furthermore, until the early 2010s, a large number CMEs found real-world applications predominantly in entertainment (e.g., video games, storytelling, etc.) or educational contexts, where the symbolic management of emotions was often sufficient, and no high degree of embodiment or real-time interaction was required.

By contrast, more embodied and socially-oriented subsymbolic CMEs relied on technologies and methodologies that were still marginal within the field, thus limiting reproducibility, interpretability, and acceptance. Computational tools and hardware infrastructures available throughout the 1990s and early 2000s were not yet mature enough to support robust, real-time multimodal sensing and processing. Physiological data like galvanic skin response, heart rate, facial expressions or voice modulation were not only difficult to acquire, but also highly susceptible to noise and hard to interpret consistently. Consequently, the integration of such signals into functional affective architectures turned generally out to be technically onerous and methodologically unstable, especially when compared to the more mature tools of knowledge representation and symbolic reasoning in AI. Despite the improvement of the algorithms, ML and ANNs still lacked the large, annotated datasets and computational resources needed for an efficient elaboration and classification of the noisy, high-dimensional input typical of physiological and affective data. Also, the Internet had only just begun enabling large-scale data collection and interconnectivity, while most (wearable) hardware was still inadequate in terms of processing speed and parallelism for de-

ploying perceptual models in real time [Section 3.1]. Even in robotics —and especially in *social robotics*, where the embodied and socially situated aims of AC found a more concrete and coherent implementation than in mainstream computer science starting in the early 2000s [Chapter 5]— hardware constraints remained decisive: Robots were often slow, rigid, and poorly adapted to the nuanced expressive demands of social interaction. Under such conditions, physiologically grounded approaches —however promising in theory— remained a initially marginal line of inquiry in the broader landscape of CME, while cognitive models continued to offer a more feasible and, crucially, *functionally* and *operationally* attractive pathway for emotion simulation in artificial agents.

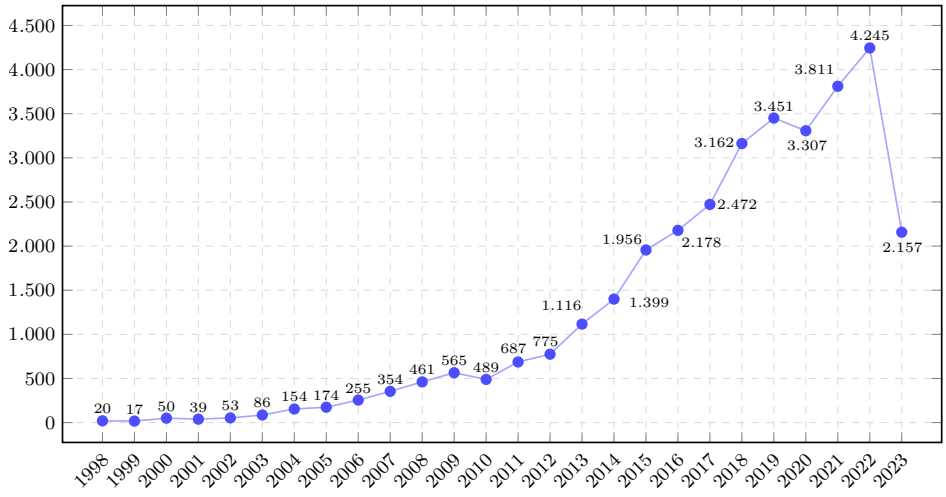


Figure 4.9: Number of papers published on "Affective Computing" from 1997 (January 1) to 2023 (September 25) —from [Pei et al. 2024, 8][64]. The sharp surge in the number of publications from 2010 onwards can be clearly noted.

4.3 Physiological-Behavioral Reductionism: Some Notes Contemporary AC

By the beginning of the 2010s, AC had become an established research field in science and engineering, with the number of scientists working and

[64] Attempts to repeat the search of [Pei et al. 2024] —aimed at completing the graph for years 2023–2025— did not yield their same results. Although the exact numbers are therefore taken here as approximate, the overall upward trend remains confirmed despite minor variations in the paper counts.

publishing in the area steadily increasing over the following fifteen years up to the present day —see [Pei et al. 2024] and [Figure 4.9]. Although it has remained a relatively niche domain compared to the broader directions of AI and computer science, even specialized subfields within the discipline have nonetheless been able to crystallize during this time: first and foremost, *emotion generation* —building on earlier CME tradition— and *emotion recognition* —more closely tied to AC's original proposal, with journals, conferences and research groups individually dedicated to each of them (see e.g., [Alexandre et al. 2020; Zhang et al. 2020a; Wang et al. 2022] and [Afzal et al. 2024]).

Emotion recognition research, in particular, while still relatively exploratory compared to emotion generation efforts in the early 2000s, has by the late 2010s emerged as the leading direction within AC, with a growing number of papers published about it in the course of the whole decade (see [Figure 4.10] and [Ma and Yarosh 2023], as well as [Kowalczuk and Czubenko 2016; Ojha et al. 2021; Wang et al. 2022; Pei et al. 2024], etc.). The improvements in hardware capabilities, more reliable multimodal sensors, and increasingly sophisticated learning algorithms have significantly fostered the implementation of real-time, data-driven models of affect. What had once been limited by infrastructural and methodological constraints has thus become a central and increasingly autonomous research direction, bringing affect recognition closer to practical deployment across diverse interactive systems in different domains of everyday life.

---

[65] Ma and Yarosh [2023, 1664] distinguish between *generation*, *expression* and *regulation* —expressed emotion may be, for example, (purposely) different from generated emotions or emotions produced within regulatory mechanisms. In general, I make no claim here for rejecting this distinction —with some aspects of it to be occasionally discussed in [Part II]. Rather (and still consistently with the distinction), I group together these three components based on the "direction" of the relationship between the emotional system and the environment:

(a) CME $\longleftarrow$ *Environment*

(b) CME $\longrightarrow$ *Environment*

whereby (a) —*inward*: Detection/Recognition (the system *receives* emotionally meaningful signals from the environment —which may also include its body, cognitive states, etc.); and (b) —*outward*: Generation, Expression, Regulation (the system *produces* or *emits* emotionally meaningful signals to the environment —which may also include its body, cognitive states, etc.).

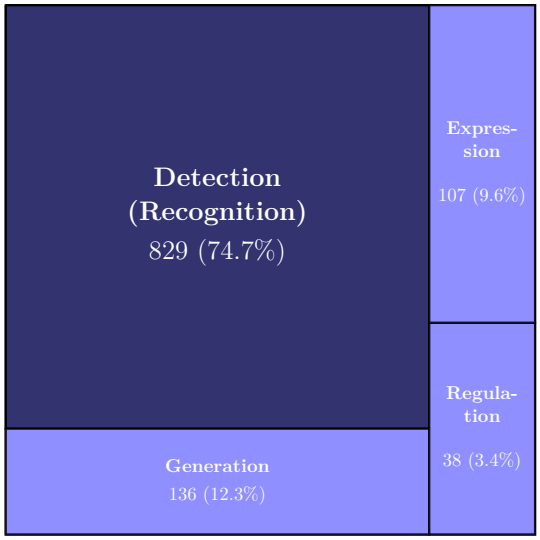


Figure 4.10: Adapted from the systematic review of AC research papers by Ma and Yarosh [2023, 1659]. The authors review 1.110 AC publications from the period 2000-2017 with the aim of understanding tendencies in their *a priori* theoretical underpinnings. Papers are classified along *Function* —addressing *why* affect is computed: *detection* (inferring emotion from input signals), *expression* (displaying affect to others), *regulation* (managing emotional states), and *generation* (producing or modeling emotional responses); *Component* —referring to *how* affect is computed: *feeling* (subjective emotional experience, typically captured via self-report), *cognition* (appraisals, beliefs, and other emotion-related thought processes), *motivation* (action tendencies or goal-directed drives), *physiology* (autonomic or central nervous system activity), *behavior* (observable expressions such as facial or vocal signals), *social* (interpersonal or cultural dynamics shaping affect), and *multiple* (integration of two or more of the above); and *representation* — involving *what* affect is computed: *biomarker* (low-level physiological or behavioral features used without labeling emotions), *discrete* (categorical emotions such as anger or joy), *dimensional* (affective states mapped in continuous spaces, e.g., valence–arousal), and *hybrid* (combinations or mappings between categorical and dimensional models). The figure shows the distribution of the papers over function (i.e., computational task): Between 2000 and 2017 the vast majority of CME research produced within AC (74,4%) has regarded affect detection, whereas only 25,3% of it affect generation[65].

Despite the popularity of affect recognition as a research trend, however, the epistemic gap between emotional cognitivism and physiological-behavioral reductionism discussed before [Section 4.2.2] has continued to exist, being particularly reflected in AC's disciplinary articulation even after 2010. On the one hand, in fact, during the new decade [Kowalczuk and Czubenko 2016] and beyond [Ma and Yarosh 2023], cognitivist theories of emotions have been still functioning as a key theoretical back-

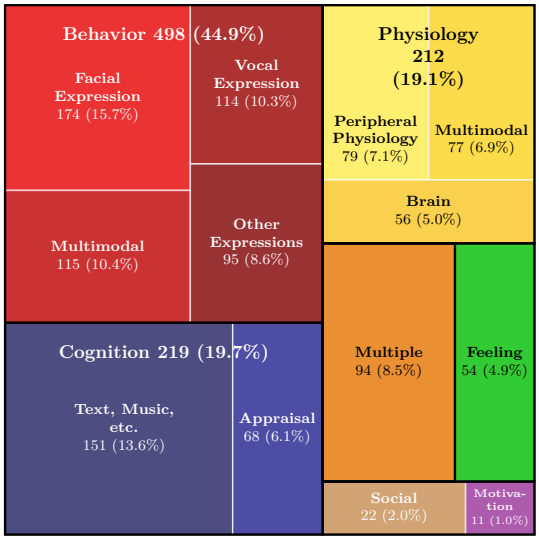


Figure 4.11: Distribution of 2000–2017 AC papers across modalities and components (see caption of [Figure 4.10]) —adapted from Ma and Yarosh [2023, 1661]. Physiological and behavioral components jointly underpin 64% of the theoretical approaches adopted in contemporary CMEs, while cognitive ones account for only 19,7%. The distribution highlights a dominant orientation toward physically and behaviorally grounded models of affect, often to the detriment of cognitively mediated structures.

ground for many CMEs, which have often relied on the operationalization of appraisal-related constructs to model affectivity. In line with the tradition, a considerable number of cognitivist CMEs in the 2010s have concerned themselves with affect generation —see [Table 4.2]. Following AC's new research trends, however, more and more of such models have begun to deal with affect recognition as well. Yet, cognitivist research efforts in this direction have largely remained confined to modally bounded tasks such as sentiment classification in textual or musical data [Figure 4.11], which eschews any modelling of emotion as a temporally unfolding, socially situated, or interactionally emergent phenomenon in the form envisaged in the program of AC.

On the other hand, it was precisely the growing centrality of affect recognition as a general research orientation within AC that naturally led to a consolidation of the physiological-behavioral perspective in the artificial modeling of emotion [Table 4.2]. Grounded in the theoretical premises of Picard's [1997] original proposal and further articulated in her early work (see [Section 4.1.2]), this perspective has progressively



Table 4.2: Distribution of 2000–2017 AC papers across functions and components (see caption of [Figure 4.11]) —adapted from [Ma and Yarosh 2023, 1664]. Over half of the models for emotion recognition (53,3%) rely on behavioral and physiological signals, while cognitive components are still well represented in affect generation (expression, regulation, generation) —see [Footnote 65]— alongside behavioral-physiological ones (7,3% and 10,6%, respectively).

strengthened throughout the 2010s, often in an increasingly —albeit not always explicitly acknowledged— reductionist sense. During this period, in fact, the vast majority of CMEs have come to equate emotions with their observable and measurable manifestations [Figure 4.11], privileging what can be externally tracked over more representational or cognitively mediated accounts.

The *dualism* between cognitivist and physiological-behavioral approaches to emotions has continued to characterize AC research until today. For both theoretical and practical reasons, researchers often —and not always consciously— fall into forms of emotional reductionism that identify emotions (and thus artificial emotions) almost *exclusively* with their external and measurable correlates. Even in contexts where machines are designed to operate as socially competent agents and expected to engage in affectively meaningful interactions —the case of social robotics will be indeed examined in all details [Section 6.1], cognitively grounded, internally mediated, and subjective aspects of emotions —long highlighted by research traditions in philosophy and psychology as central to affective phenomena— are frequently sidelined in favor of more readily detectable outputs and inputs such as facial expressions or phys-

iological signals —which are often also perceived as simpler and more straightforward to implement.

Summary

This chapter has analyzed the emergence of *Affective Computing* as both a technical paradigm and a philosophical redefinition of computational emotions. Building on the subsymbolic and embodied approaches discussed in [Chapter 3], AC reformulated the problem of emotion in machines as one of recognition, expression, and interaction. Initiated by Picard's proposal, it defined affect as a measurable and computationally tractable dimension of intelligent behavior, integrating physiological, behavioral, and contextual data to enable machines to detect, interpret, and simulate emotional states. Subsequent developments expanded these principles into multimodal architectures capable of integrating vision, speech, gesture, and physiological sensing, together with adaptive models based on learning and feedback. In this way, emotions became not only computable entities but communicative processes essential to human-machine exchange, grounding the idea that affective capacities are central to sociability, empathy, and intelligent interaction even in computers.

The chapter has furthermore shown that, owing largely to its focus on emotion recognition, the development of AC has progressively privileged the physiological and behavioral dimensions of artificial affect over its cognitive and evaluative aspects. This tendency has given rise to a form of *physiological–behavioral reductionism* within the field, one that by and large persists to the present day.

More on this reductionism and its philosophical implications, particularly in contemporary robotics and HRI, will be said in [Part II]. Before turning to that, however, it is important to consider how similar reductionist tendencies have emerged within robotics —a domain that, in many respects, constitutes a direct operational extension of AC's core principles. In [Chapter 5], the evolution of affective and social robotics will thus be examined, from the early history of emotional automata to modern humanoid platforms, showing how the theoretical and computational models of AC have been progressively embodied in machines capable of expressing, perceiving, and regulating emotions within social interactions conducted in the physical world.

Chapter 5

From the Models to the Real World: The Rise of Social Robotics

In the last [Chapter 4], I discussed the core principles of AC, focusing not only on its technical aspects but also on its epistemological foundations and theoretical shortcomings —particularly the issue of the physiological-behavioral reductionist notion of emotion.

As argued there, AC's main contribution lies not merely in its technical architectures but, more profoundly, in its theoretical reframing of artificial emotions as simultaneously *social*, *embodied*, and *cognitive*. A truly intelligent agent is such insofar as it can *multimodally* communicate with humans and their *physical* environment, elaborating emotionally meaningful inputs and outputs that enhance its sociability and interrelational abilities.

Building on these premises, the following chapter examines how these principles —initially confined to software implementations— have progressively been translated into the design of *physical* agents. In particular, it traces how the AC conception of emotions as social, embodied, and cognitive found a direct and natural application within the field of *robotics*, where an artificial agent's ability to interact with humans crucially depends on its bodily presence and on its capacity to display, recognize, and respond to affective signals.

In [Section 5.1], I outline the main historical developments of robotics, with particular attention to its often-neglected emotional dimension —from the 18th-century musical automata to the analog machines of cybernetics and the early robots developed within GOFAI. Then, in [Section 5.2], I explore how, during the second half of the 20th century, new technological advances and theoretical frameworks —especially those emphasizing embodiment, motivation, and situated intelligence— opened the way for applying Picard's principles to actual embodied agents. Finally, in [Section 5.3], I show how AC directly influenced the emergence of social robotics as a distinctive research field, where artificial emotions became not merely a technical feature but a fundamental condition for robots to operate as credible social partners —with Cynthia L. Breazeal's *Kismet* standing as the paradigmatic case in which Picard's vision of multimodal, socially embedded affectivity has been concretely realized in a physical machine.

5.1 Notes on the History of Affective Robotics

Despite the centrality of robots in science, technology and popular imagination, a general history of robotics aimed at systematically and critically reconstructing its conceptual, technical, and cultural developments is still largely lacking. In technical handbooks or journal issues on robotics and *Human-Robot Interaction* (HRI), it is common to find brief introductory paragraphs on the early developments of robotics. Such contributions, however, typically contain just fragmented and anecdotal information about technical milestones, patents, notable prototypes and traditional topoi of the history of technology —the Golem in Hebrew mythology, the myth of Pygmalion, Ctesibius's hydraulically powered statues, von Kempelen's *Mechanical Turk*, the origin of the word "robot" in Čapek's writings [Footnote 17], etc. (see e.g., [Stone 2005; Siciliano and Khatib 2016; Goswami and Vadakkepat 2018], etc.). By contrast, some more extensive historiographical efforts available so far have primarily focused on specialized areas of robotics —most notably industrial [Wallén 2008;

Gasparetto and Scalera 2019] and especially medical [Challacombe et al. 2006; Kalan et al. 2010; Shah et al. 2014; Lane 2018]) robotics.

Yet, mostly written by engineers or practitioners rather than by professional historians of science and technology, these rather descriptive accounts do not engage with the broader conceptual evolution and epistemological foundations of robotics. They tend to neglect how theoretical assumptions, cognitive models, and shifting notions of intelligence and embodiment have historically shaped —and been shaped by— robotics research. They fail to address how the development of robotics is embedded in the broader evolution of AI (and AC), and how ideas and developments in the history of science, technology, psychology, cognitive science, neuroscience and philosophy (e.g., [Chapter 1]) have influenced —and been influenced by— the emergence and transformation of robotics *as a scientific research field*, from early mechanical automata and cybernetic machines to symbolic AI robots and today's social-affective robots. Of course, this holds even more visibly when it comes to social and affective robotics, to which no historical-epistemological analysis has been *specifically* dedicated to date.

The aim of this section is therefore to offer such a historical-epistemological reconstruction, focusing in particular on the development of affective robotics. Rather than attempting a complete historical overview —unfeasible and unnecessary here— the goal is to highlight the key conceptual, technological, and epistemological moments in the broader histories of automation, AI and CME that led, around 2000, to the idea of building not only artificial systems capable of processing or simulating emotion, but also *physically embodied* machines designed to sustain advanced social interaction precisely through their emotional capacities.

5.1.1 18th-Century Musical Automata

In comparison to the history of, say, computers or cars, doing a history of robots is, to some extent, a challenging task. In fact, while computers and cars are well-defined and clearly identifiable objects —as many other technological artifacts or scientific concepts, robots do not seem to be quite so. From a conceptual perspective, to the object we call "robot" a number of different properties can be attributed: physicality,

anthropomorphism, mobility, interactivity, receptivity, activity, autonomy, sociality, intelligence, specificity and multimodality among the most characteristic ones[66]. None of these properties, however, is sufficient and necessary to provide a univocal definition of this particular object: A vending machine is autonomous and specific, but we would not call it a robot; a Roomba is neither anthropomorphic nor social, yet, in this case, we define it as such.

The concept of robot is thus a rather elusive one, whose definitional instability is often compensated by a naïve, pre-theoretical and intuitive use both laypeople and scientists make of it —a use which is clearly shaped also by the cultural, popular and collective imaginaries surrounding it[67] [Section 1.2.3]. It is therefore not surprising that, within the history of robotics, references often recur to the broader history of automation technologies —from the steam engine and the Jacquard loom to early control systems— where robotics is implicitly treated as a late expression of more general mechanical principles that, it is assumed, were already applied in the past.

In what follows, I do not attempt to reconstruct a linear or comprehensive sequence of technical developments, nor to retroactively identify "robots" where none yet existed. Rather, my aim is to trace a set of historically situated artifacts, ideas, and design strategies that —while not robotics in any strict or modern sense— nonetheless contributed to shaping the conceptual space from which affective robotics could later emerge. From this perspective, a particularly instructive case is represented by 18th-century musical automata, some of the earliest attempts to incorporate affectivity into artificial devices showing at least some of the properties inherent to the concept of "robot" (primarily, physicality, anthropomorphism and autonomy).

[66] Of course, the nature and definition of these properties are themselves subjects of ongoing debate. As discussed more in details in [Section 6.2.5], a notable example is anthropomorphism, which philosophers, psychologists, and HRI specialists currently discuss as to whether it is limited to the physical and external characteristics of an object or also extends to behavior, cognition, and other non-physical aspects (see e.g. [Cheetham et al. 2011; Fink 2012; Von Zitzewitz et al. 2013; Torricelli et al. 2019; Jastrzab et al. 2024]).

[67] Albeit without particular analytical and philosophical depth, some few attempts have been made to articulate necessary and sufficient conditions for a general, formal definition of "robot" —see e.g., [Mataric 2007; Guizzo 2020].

As widely known, the 18th century was a period of great technological innovation, that eventually led to the First Industrial Revolution. Significant advancements were especially made in precision mechanics and clockmaking, where improved gear and escapement mechanisms could enhance the accuracy and reliability of increasingly complex automatic systems (see e.g., [Musson and Robinson 1989]). Interestingly, these advancements were also applied to a branch of mechanics with deep historical roots —most likely tracing back even to Classical Antiquity: that of *musical automation*[68] [Section 8.2.1]. Champions were, in this respects, especially French, German and Swiss clockmakers, who began to insert in highly sophisticated clocks rotating cylinders, small pipe organs or moving flute-playing figures able to perform elaborated melodies —in some cases, even composed by renowned composers of the time such as Händel, Mozart, Hayden, Beethoven, etc. [Buchner 1960, 70-72].

Yet, the most impressive musical automation experiments of the century were certainly the *musical androids* [Chapuis and Droz 1949; Margócsy 2015; Westermann 2016]. Unlike musical clocks, the internal mechanism of these anthropomorphic mechanical figures —an intricate system of levers, weights and rotating cylinders— *could be* distinguished from the means of music production itself. These androids were namely able to perform music on real, life-size musical instruments otherwise playable by humans.

[68]The history of musical automation is very old (see e.g., [Fowler 1967; Koetsier 2001; Chen et al. 2018], etc.). The Greek Ctesibius and Herons of Alexandria are known to be among the first inventors to build automated musical instruments and mechanical singing birds [Woodcroft 1851; Papadopoulos 2007]. In the 9th century, the brothers Mūsā in Baghdad described the first programmable musical machine, an automatic flute player powered by air [Mūsā 1979; al–Ḥassan 1986, 60; al–Jazarī 1989; Paz et al. 2010, 60-92; Krzyzaniak 2012, 2ff.]. During late Middle Ages and early Renaissance, mechanical carillons were built in church and city hall towers all over Europe. The flourishing of musical engineering occurred however in the 16th-17th century, with Athanasius Kircher [1650] describing the design of innovative automated organs in his musical treatises, and barrel organs being build in the major European courts —e.g., the 1596ca. (still functioning) barrel organ of *Fontana dell'Organo* built by Luca Blasi at Palazzo del Quirinale in Rome, Italy [Fronzuto 2008]. As shown here, it is noteworthy that music has historically acted as both a driver and a repository of ideas in the field of automation in general, and affective robotics in particular (see [Zanzarella 2024] and [Chapter 8]).

Le Flûteur, crafted by the eminent French inventor Jacques de Vaucanson (1709–1782), stands as a famous example of 18th-century musical automaton —see [Doyon and Liaigre 1966, 49–108] and [Figure 5.1].

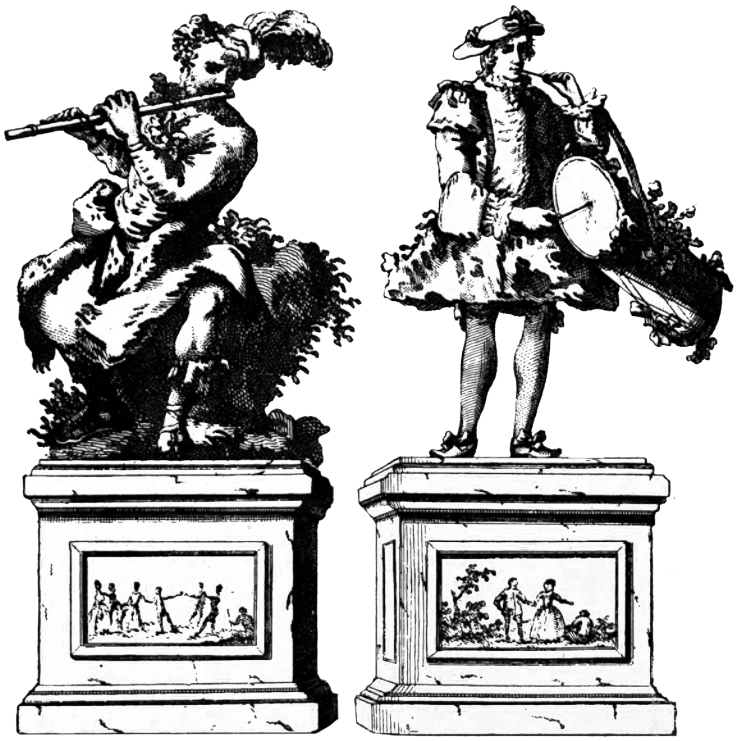


Figure 5.1: Vaucanson's *Le Flûteur* and *Le Joueur de tambourin* (1738) –detail from the poster for the presentation of Vaucanson's automata in Strasbourg (1746), engraving in [d'Allemagne 1902, 222].

Inspired by *Le Berger flûteur* —a marble sculpture by Antoine Coysevox (Jardin des Tuileries, 1716)— it was built between 1737 and 1738 and made capable of playing twelve different melodies. The first exhibitions of the automaton to the public took place in 1738 in Paris, where it was visited also by the members of the *Académie royale des sciences*, who gave their approbation to the work. From 1741, together with other two automata by the same author, *Le Joueur de tambourin* [Figure 5.1] and *Le Canard digérateur*, the flutist began a long travel across French, Italy, England, the Netherlands, Germany and Austria, where, after having had several owners, it was allegedly exhibited for the last time in mid-19th century before completely disappearing without trace[69].

[69]The veridicality of this information, however, was denied already in 1882: «The automatic flutist is not in Vienna, as generally believed» [Charton 1882, 122, my tr.].

In a *mémoire* presented to the *Académie*, Vaucanson described his automaton and explained all the physical and mechanical principles behind it [Vaucanson 1738, 9ff.]. The automaton was composed of a life-size human figure of about $5,5$ foot (=∼1,70 m) tall, placed on a pedestal of $4,5$ foot (=∼1,40 m) hiding large part of the mechanism that made the figure move and play. The figure held in the hands a real Baroque transverse flute, which could be replaced with other flutes of the same kind without consequences on automation and performance. Sound was produced by the automation itself: A crank mechanism activated bellows in the pedestal, which pumped air up to the mouth of the figure. Air was then blown into the instrument, producing sound. In the mouth, suitable movements of the tongue controlled quantity and pressure of the air to be emitted in accordance with the frequency of the notes to be played[70]. The different notes were finally produced by proper positioning of the fingers on the holes of the flute. The movement of the fingers, covered with leather to imitate the softness of human skin on the flute holes, was controlled by a system of levers activated by the rotation of a pegged cylinder. As in previous clocks and Kircherian water organs [Footnote 68], the cylinder "stored" all musical information and could be "programmed", by opportune arrangement of the pins, with several different melodies.

Inspired by Vaucanson's pioneering work, also other musical automata were build during the 18th century. Among the most known ones there are for example *La musicienne* (or *The Harpsichord Player*) —see [Figure 5.2], built by the Jaquet-Droz family in La Chaux-de-Fonds (Switzerland) between 1767 and 1774[71] [Perregaux 1906; Perregaux and Perrot 1916; Voskuhl 2019], and Peter Kinzing's and David Roentgen's *Joueuse de tympanon* ("Dulcimer player") —see [Figure 5.3], build in Neuwied (Germany) between 1783 and 1785 and gifted to Queen Marie

[70] The regulation of air emission, air pressure, tone frequency, etc. required Vaucanson to undertake a prior analysis of all the physical-acoustical, anatomical and music-technical aspects of flute playing, which he did in the first part of the *mémoire* [Vaucanson 1738, 4-10] —his considerations about acoustics and flute technique, however, were not entirely accepted as accurate by the experts of the time (see e.g., [Quantz 1752, 4]).

[71] With other famous automata such as *The Writer* and *The Draughtsman*, *La Musicienne* belongs to a group of about ten automata build by the Jaquet-Droz family between 1730 and 1810 [Voskuhl 2019, 4-5].

Antoinette of France[72]. Both automata worked on nearly the same mechanical principles applied by Vaucanson: They used cylinders, levers, cranks, etc. for storing musical sequences and making the figures play music on real instruments, thereby enabling the execution of short performances with a high degree of accuracy.



Figure 5.2: *La musicienne* or *The Harpsichord Player* (La Chaux-de-Fonds, Switzerland: 1772–1774) by Pierre and Henri-Louis Jaquet-Droz —from Musée d'art et d'histoire, Neuchâtel (Switzerland).



Figure 5.3: *Joueuse de tympanon* (Neuwied, Germany: 1783-1785) by Peter Kinzing (clockwork) and David Roentgen (cabinet) —from Musée des Arts et Métiers, Paris (France).

For that, all these automata displayed a high degree of motor complexity. Beyond the articulation of the fingers on keyboards or flutes, they could even move their chest and perform realistic movements of the head, eyes, and arms, simulating breathing, facial expressions and other complex biological functions. These bodily gestures were not random, but synchronized with the character of the musical pieces they were playing, serving thereby an important expressive and communicative function: that of *conveying emotions* and *eliciting affective responses* in the audience [Voskuhl and Riskin 2010].

This expressive purpose must be understood in the context of 18th-century musical aesthetics. During this period, the so-called *Affektenlehre* (from the German, "theory of affects") still framed implicitly the

[72]Differently from Vaucanson's automata, though, they have survived until present times, and can be seen and heard respectively at the Museum of Watches in Le-Chaux-de-Fonds and at the Musée des arts et métiers in Paris.

experience of musical performance and listening [Section 8.1]. Broadly speaking, according to this theory, composers and performers could encode specific emotional meanings into a piece by properly arranging musical structures —forms, dynamics, tonalities, harmonic progressions, etc.— univocally associated with them. Listeners, in turn, were expected to decode these patterns and emotionally resonating to them [Glüxam 2020]. Within such a framework, expressive bodily movement was not merely a secondary aspect of performance, but a necessary one: It was namely an integral component of a persuasive and emotionally communicative performance (see [Bach 1753, 115–133], and [Voskuhl 2019; Voskuhl and Riskin 2010]).

In this sense, *Le Flûteur*, *La Musicienne*, and the *Joueuse de tympanon* were not simply technical wonders or entertainment devices. They constituted early, though limited, attempts to artificially simulate emotions by a mechanized reproduction of their external expressive correlates. In this sense, they may be considered among the first historical examples of what would later be conceptualized as an *affective artificial agent*.

This consideration also fits with how historians have come to interpret these automata —namely, as *scientific models* designed to simulate life and biological processes within a scientific and philosophical paradigm in which such processes were still largely conceived in mechanical terms [Voskuhl 2019; Riskin 2021]. For instance —given also Vaucanson's exchanges with the *Académie*[73] and his technical treatise— it is likely that his *Le Flûteur* was as a part of a broader effort to empirically substantiate his acoustical theories [Riskin 2021, 119ff.]. Within this framework, then, the expressive and coordinated movements of these automata —though limited and predetermined—may be seen as attempts to reproduce, or at least approximate, emotional and embodied phenomena within a mechanistic scientific paradigm. While they were not intended to replicate subjective emotional experience, their functioning nonetheless relied on observable patterns of affective expression that, for the first time, could be operationalized, mechanized and embedded in a machine.

[73] A similar case is that of the *Joueuse de tympanon*, in turn gifted by Queen Marie Antoinette also to the *Académie royale des sciences*.

5.1.2 Early Experiments in Robotics: Cybernetics Creatures

The mechanical innovations introduced during the century of the Enlightenment underwent continuous refinement in the following decades. The 19th century, in fact, saw a dramatic surge in automation technologies, especially within industrial settings, with the invention of programmable industrial machinery, the adoption of assembly-line methods and the rise of electricity as a new power source for automation. Furthermore, the 19th century was also a period of extraordinary scientific achievements of great technological significance, with groundbreaking studies on thermodynamics and statistical mechanics, the establishment of electromagnetism theory and the first experiments with computing [Footnote 12].

The growing complexity in machinery eventually led to the need of more effective *human-machine interfaces* and *control systems*, able to coordinate multiple components, and ensure both work and power efficiency in increasingly automated environments. It is precisely in this context that the first formalized *control theories* emerged, aimed at mathematically describing the ways of *controlling* a dynamic system, i.e., the ways of optimally *regulating* it so to allow the performance of a desired behavior. In this respect, one of the first scientists to propose a mathematical theory of control and optimal regulation was the Scottish physicist James Clark Maxwell [1867], who formally described the behavior of the *governor*, a self-regulating mechanism already applied in machines like Watt's steam engine for adjusting speed automatically and proportionally to varying loads. Maxwell modeled the governor as a dynamic system using differential equations, and analyzed the conditions under which its *feedback-based* regulation remains stable or becomes unstable.

Maxwell's [1867] early control theory would later serve as a foundation upon which, during the 20th century, a new science, *cybernetics* [Chapter 1], began to take shape, especially through the work of Norbert Wiener [1948] —who in turn referred to some of Maxwell's original insights. In particular, Wiener's [1948] cybernetics developed from the attempt to ascertain the heuristic value of control theory's mathematical principles for better understanding not only machines, but also *biologi-*

cal systems, from simple cells to complex organisms, behaviors and even ecosystems.

Now, the hypothesis about the functional equivalence of self-regulatory mechanisms across machines, organism-environment dynamics and biological systems —formalized in the principles of cybernetics— led many scientists in biology, ethology, neurophysiology and psychology to the idea of building machines to use as *epistemic* (*physical*) *models*[74] for empirically investigating and testing it. It is precisely in this context that the first robotic experiments were conceived, as a means to study biological systems and emergent behaviors relying on the structural and functional analogies between organisms and machines.

Among the earliest and most notable *cybernetic robots* there are William Grey Walter's [1950; 1950; 1951; 1953] "Tortoises" [Holland 2003]. American neurophysiologist at the Burden Neurological Institute in Bristol and expert in brain research, Walter built, between the 1940s and the 1950s, various animal-like electromechanical devices with the aim of studying the possible emergence of complex behaviors in structurally simple organisms —the most known ones are the *Machina speculatrix* (from the Latin "thinking machine") viz. ELMER (*Electro Mechanical Robot*); and the *Machina docilis* (from the Latin "obedient machine") viz. ELSIE (*Electro Light Sensitive with Internal and External Stability*) [Figure 5.4].

The tortoises —name given because of their shape, and after the turtle of *Alice in Wonderland* by Lewis Carroll— were extremely minimalist in their design, with three motorized wheels for locomotion and a rechargeable battery. They could sense their spatial environment through a "perceptual" system consisting of a photocell (a sensor for detecting light levels) and a bump sensor (for detecting contact with objects). External inputs from it were then processed by an analog "brain", consisting of a simple electronic circuit based on two vacuum tubes that

[74] As clear from [Section 5.1.1], the epistemic use of robots as physical models for testing scientific theories was not a completely original idea.

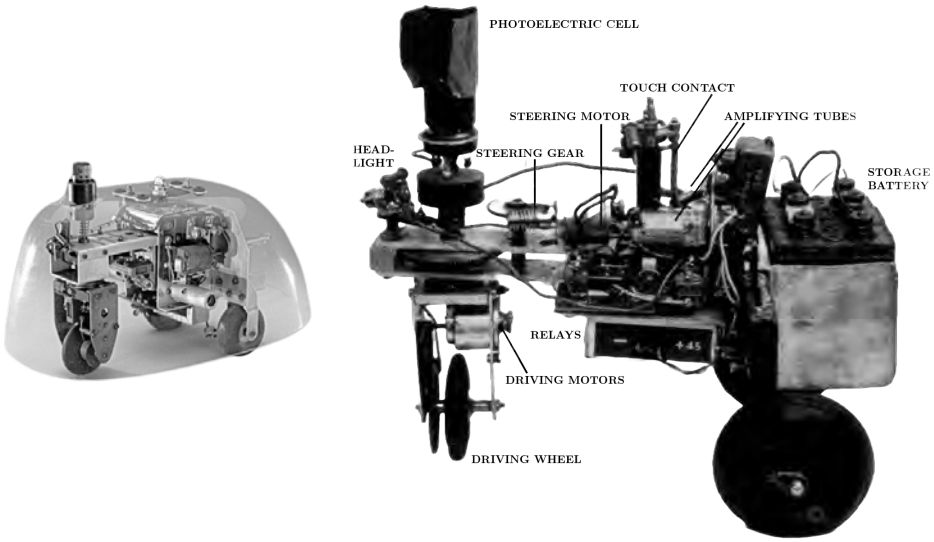


Figure 5.4: Left: Example of Walter's robotic tortoises —from the catalog of the National Museum of American History, Washington (USA); Right: Components of Walter's robotic tortoises (external shell removed) —adapted from [Holland 2003, 354].

continuously linked sensory information directly to motor output[75] [Figure 5.5].

Despite their mechanical simplicity, the tortoises were capable of exhibiting relatively complex behaviors, most notably phototaxis, i.e., the capacity to actively move toward a light source —a capacity also employed by several biological organisms as a means of orientation in the environment. Through the photocells, the robots could detect variations in light intensity. When a light was sensed, the sensors triggered the motors, making the robots orient themselves toward it and start moving

[75] *Analog circuits* process incoming signals in continuous form —for example, light intensity (which varies continuously) is converted into an electric current whose intensity can take any value within a continuous range. In contrast, *digital circuits* sample continuous signals at specific time intervals and convert them into a sequence of discrete 0s and 1s, which can then be used for further processing. The first digital circuits appeared in the 1930s and were still based on vacuum tubes —which functioned, in this case, not by directly linking continuous inputs to outputs like analog circuits, but as binary, on/off switches representing information about continuous signals in a digital form. Such switching circuits, as demonstrated by Shannon [1937], could be modeled and optimized through Boolean algebra —a conclusion that provided a rigorous theoretical foundation for digital logic design. The invention of the transistor at Bell Labs in 1947 —a semiconductor component that amplifies or switches electric current in the circuit— eventually allowed the construction of smaller, faster, and more reliable devices, paving the way to modern computers and robots.

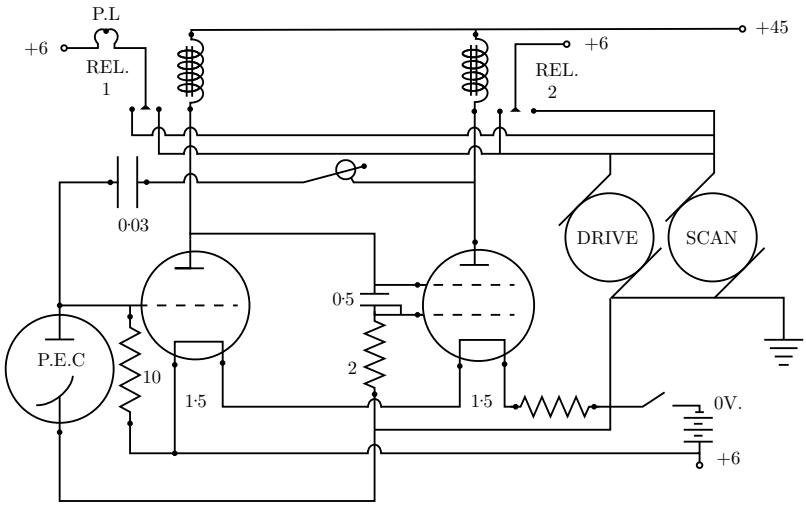


Figure 5.5: Diagram of Elmer's analog circuit —from [Walter 1953, 242].

in that direction. Changing the position of the light source caused the robots to reorient themselves accordingly, producing a behavior that resembled active and *intentional* pursuit or tracking. Obstacles on their path caused the bump sensor to temporarily override the motor outputs generating reactive turning movements until the robots found new clear trajectories toward the light.

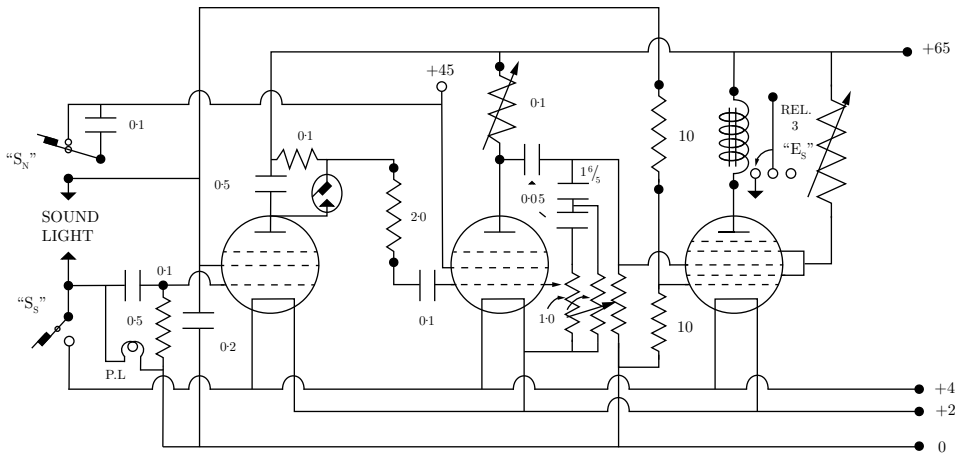


Figure 5.6: Diagram of the circuit of CORA analog module for associative learning —from [Walter 1953, 247].

Crucially, these behaviors were *emergent*, and, surprisingly, just from relatively trivial electromechanical arrangements. They did not result

from direct, top-down programming: There was no central control unit processing inputs and issuing instructions according to stored sequence of formally defined actions. Instead —anticipating some core claims of EC [Section 3.2.7]— the robots operated entirely through real-time *coupling* between perception and action in the environment, governed by continuous feedback loops. Their movements —approaching a light source, avoiding obstacles, searching for charging stations, etc.— emerged directly from the dynamic interaction between the sensory inputs and analog motor responses, without requiring any form of internal representation or symbolic processing.

These robotic experiments showed that, in general, complex behavior can emerge from simple structures. But Walter [1951; 1953] went even further, exploring *associative learning* in artificial systems. Already studied by Pavlov [1927], psychophysicists and behaviorists [Section 3.2.1.1], associative learning involves a subject developing new behavioral responses by forming connections between previously unrelated stimuli. To assess whether his robots could display this kind of adaptability, he thus devised an analog module for associative learning, CORA (*Conditioned Reflex Analogue*), to be incorporated into his robots [Figure 5.6]. CORA enabled the robot to form simple associations through continuous interaction with its environment. For example, if a neutral sound (e.g., a whistle) was paired with an aversive event (e.g., collision with an obstacle), CORA caused the robot to begin to turn away at the sound alone, as if anticipating the negative consequence. Here as well, *cybernetically*, such complex learning behaviors *emerged* from the robots' simple sensorimotor structure interacting with the environment in a continuous feedback-loop, without explicit programming and symbolic representations.

A similar experiment was conducted by Claude E. Shannon [1951], who, in addition to his foundational work in logic and the mathematical theory of communication [Footnote 10], also had a strong interest in electronics and cybernetics (see also [Shannon 1953a,b]). As in Walter's case, Shannon's experiment revolved around a form of *bioinspiration* and

biomimesis[76]: Namely building a simple electromechanical robot-mouse able to explore, memorize and re-navigate a maze through a continuous process of feedback-driven learning and trial-and-error adaptive behavior.

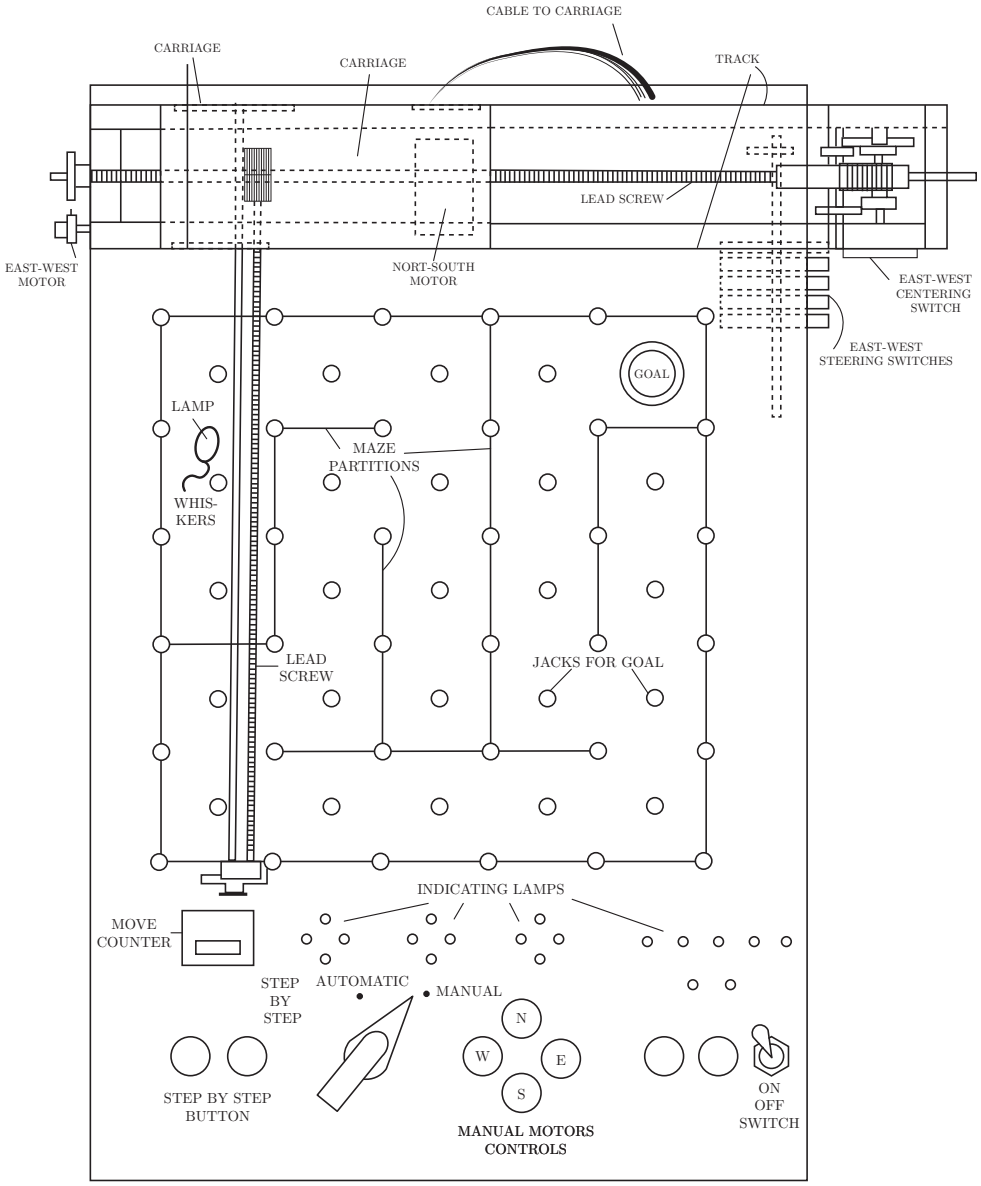


Figure 5.7: Outline of Shannon's maze-solving machine —from [Shannon 1951, 174].

[76]Indeed, Walter and Shannon may be regarded as the true initiators of *biorobotics*, a scientific discipline aimed at replicating aspects of living organisms through robots (see also, [Tamborini and Datteri 2022; Tamborini 2024a,b].).

Shannon built several prototypes of his system[77] [Figure 5.7]. Broadly speaking, it consisted of a rectangular wooden box with a metal surface mounted on the top, divided into twenty-five squares by metal walls placed between them, removable to manually define the maze layout. Underneath, two electric motors controlled a magnetic platform capable of moving horizontally and vertically, one cell at a time. A small wooden mouse —opportunely named *Theseus*— was placed on the surface. A bar magnet lodged in it made it follow faithfully the movements of the platform below. The mouse had no internal components, but carried a metal whisker arm on its front. When the whisker bumped into a wall, it completed an electrical circuit through the conductive surfaces, signaling a collision to the electronics below. This comprised about seventy-five[78] electromagnetic relays, grouped into pairs assigned to each square, representing the mouse's "memory", and the remaining ones controlling motor movement, direction sequencing, and other functions of the platform.

Each time the mouse entered a new square, the system attempted movement in the direction currently set by the state of the relay pair associated with that square. If the path was clear, the movement succeeded, and the relays were left in that state, storing the successful exit direction. If instead the movement led to a collision —detected by the mouse's whisker touching a wall and closing the circuit— the motor reversed, bringing the platform back to the center of the square. At the same time, a mechanism repeatedly attempted successive directions, rotating the mouse's orientation by 90 degrees each time, until a valid path was found.

The successful attempt to move led to a change in the state of the relays associated with the corresponding square, which eventually made

[77] Despite their importance in the development of robotics, Shannon's electromechanical mice are relatively neglected in the history and philosophy of science —some few (and mild) exceptions are e.g., [Somenzi and Cordeschi 1994; Cordeschi 2002; Boden 2008; Dupuy 2009; Kline 2015]. This is likely due also to the scarcity of historical sources, as Shannon himself never published comprehensive and systematic articles or books about them, with most available documentation appearing scattered across personal notes, handwritings, unpublished texts, demonstrating interviews, and non-academic magazines (see e.g., [Shannon 1952a,b]).

[78] Sources vary on the exact number of relays used. Shannon's original descriptions mention approximately 110 relays, including memory and control logic. However, other technical accounts often refer to totals ranging from 65 to 75, depending on which components are included in the count.

the mouse reach an electric terminal at the end of the maze as an ideal piece of cheese. After that, the system had a "memory" of the correct path to the cheese, represented by the overall state of the relays. In fact, when the mouse was placed again at the start, with the same maze layout, it could use this memory to navigate directly to the goal, without further trial or correction.

Albeit still based on analog circuitry and feedback-loops characteristic of the cybernetic approach, Shannon's *Theseus* and its mechanisms for memory and learning[79] anticipated interesting aspects of the (GOF)AI robotics that will be developing from the 1960s onwards [Section 5.1.3]. Relay states retained information about correct directions, effectively encoding a distributed trace of the mouse's interaction history with its environment. This trace was not symbolic in the digital sense, nor constituted an internal map in the full-blown cognitive sense. And yet, it represented a form of knowledge the system could use as a set of "rules" for effectively behaving in the environment. In this sense, in contrast to Walter's tortoises —whose behavior was entirely determined by the *immediate coupling* between sensory inputs and motor responses— Shannon's mouse already instantiated a subtle operational distinction between perception (viz. physical interaction with the environment) and stored internal knowledge. It is precisely this architectural principle —that autonomous behavior can be governed by internal representations of the environment— that would later become central in (GOF)AI robotics.

With Shannon's experiment, thus, the cybernetic principle that complex (artificial) behavior can *emerge* from simple (artificial) mechanisms received once again convincing support. To an external observer, the behavior of the robotic mouse seeking the "cheese", much like that of the robotic tortoises seeking the light, appeared —«*frightfully*»[80] [Wal-

[79]Some critics (see e.g., [Gabor 1956]), however, pointed out that Shannon's maze-solving mouse did not genuinely simulated biological learning —a criticism indeed directed also to Walter's tortoises [Somenzi and Cordeschi 1994; Cordeschi 2002, 163ff.] Its adaptive behavior, in fact, was based on distributed changes in the external environment (state of the relays), and not on some forms of real, internal and active, memory.

[80]The description of these behaviors as *frightfully* intentional does not stem from academic sources, but rather from popular media —magazines and newspapers of the time (e.g., [Walter 1952])— which captured the astonishment and occasional unease of lay audiences encountering such machines for the first time. These accounts

ter 1952, 11]— goal-directed, resembling *intentional* action, as if driven by an internal *motivation* toward the goal[81].

The same elementary mechanisms that govern task execution and regulation of limited resources in the interaction with the environment can likewise give rise to forms of *emergent affective behavior*. An explicit extension of such cybernetic insights to the realm of affectivity was proposed by the Japanese psychologist and roboticist Masanao Toda [1962; 1982a,b]. As an earlier representative of the great Japanese robotics tradition[82] (see [Section 5.2.1] and [Section 8.2.1]), and at the beginning of the theoretical reflection on artificial models of emotion [Section 2.1], Toda criticized structuralist and behaviorist academic psychology [Chapter 1] for fragmenting the study of human mind into isolated faculties like perception, memory, language, emotion, etc., examined through narrowly defined tasks, mostly under artificial and unrealistic laboratory conditions. Rather, he argued for a shift towards the study of complete, self-sufficient systems that could operate autonomously in their environment and reveal the integrated nature of behavior. This perspective led him to formulate a "science fiction"[83] *gedankenexperiment* to illustrate this thesis in action, that is, the famous *Solitary Fungus Eater* (SFE) —an interactive scenario in which the reader assumes the role of

offer valuable insight into the *naive* perception of artificial agents by non-specialist observers unfamiliar with cybernetics and early robots. Interestingly, this popular response anticipates, by nearly two decades, Mashairo Mori's [1970] formalization of the concept of *uncanny* in 1970, which will become one of the most discussed in HRI until today —see [Section 6.2.5].

[81] These impressions already foreshadowed important epistemological questions about the nature of robotic agency, still today at the core of the debate in HRI and philosophy of robotics. A notable example is the problem of attribution of mental states to robots [Thellman et al. 2022], which will be discussed in [Section 7.1].

[82] Japan has indeed cultivated a remarkable tradition in the development of robotic technologies. While most existing historical narratives of robotics —including the present one— remain grounded in a Western epistemological framework, it would be particularly valuable to examine the cultural, religious, philosophical, and technical foundations that shaped early experiments in robotics in Japan, in order to assess the reciprocal exchanges and cross-influences with the Western tradition rooted in cybernetics, AI, cognitive science, and AC. Such an inquiry constitutes one of lines along which the present research will be developed in future, in order to contribute to a more inclusive and genuinely cross-cultural history of robotics.

[83] Again, it should be noted that science-fiction had a primary role in boosting the development of research ideas and technologies in AI and robotics, especially between 1950s and 1960s [Section 1.2.3].

a remote operator of a minimal autonomous agent in a simplified, yet unpredictable environment:

> «You are a remote control operator of the robot miner nicknamed "Fungus-Eater", sent to a planet called Taros to collect uranium ore, which uses wild fungi growing on the surface of the planet as the main energy source for its biochemical engine. The uranium ore and fungi are distributed over the land of Taros, which is covered mainly with black and white pebbles, and little is known about the mode of their distribution. As the operator you can control every activity of the Fungus-Eater, including the sensitivity of the fungus- and uranium detection devices. All the sensory information the robot obtains will be transmitted here and displayed on this console so that you will feel as if you are the Fungus-Eater itself.
>
> Note that your mission is to collect as much uranium ore as possible, and your reward will be determined in proportion to the amount of uranium you collect. Note also that the amount of fungi you collect and consume during your mission is irrelevant to the reward. Remember, however, that every activity of the Fungus-Eater, including the brain-computer operations, consumes some specified amount of fungus-storage. Never forget that the Fungus-Eater cannot move to collect further uranium ore or fungi once it runs out of its fungus-storage, and your mission would be over then and there. Good luck!» [Toda 1982b, 95].

In its very design, the SFE is a cybernetic robot. Like Walter's tortoises or Shannon's mice, its behavior is governed by a closed dynamic system based on continuous feedback loops. The need to constantly balance uranium collection with fungal recharge for avoiding "starvation" results in a process of real-time action selection driven by the robot's ongoing interaction with its environment. In so doing, the SFE does not follow a predefined set of formal rules by which inputs are processed and outputs emitted. Its decision-making arises entirely from the dynamic coupling with the environment. Its behavior is adaptive, rather than deterministic: SFE's environment is partially uncertain (uranium and fungi randomly distributed) and therefore not symbolically mappable. Thus, environmental stimuli do not trigger fixed responses, but interact with the system's changing internal state for producing action. In this sense, perception, decision-making, and action unfold as continuous and mutually interdependent dimensions of the system —which makes Toda an important anticipator of several core ideas that will be later developed by ecological psychology, embodied cognition and affective neuroscience [Section 3.2.7].

For these reasons, from the SFE a form of intelligent (i.e., intentional, motivated, goal-oriented) behavior emerges. And it is precisely in this respect that Toda conceptualizes it also in affective terms. Emotional behavior arises whenever the system must reorganize its priorities or adjust its behavioral tendencies in response to changes in internal state or environmental conditions —whether in situations of tension, scarcity, or relative abundance. These episodes are not triggered by any single and discrete stimulus, but emerge from the global configuration of the system's internal dynamics and its interaction with the environment. Crucially, such patterns may also be interpreted from the outside as emotional in character: Urgency, hesitation, apparent frustration, or even efficiency and opportunistic drive become visible not because they are encoded as such, but because they emerge as coherent, adaptive responses to the evolving demands of the environment and the shifting internal conditions of the agent. Emotion, for Toda, is therefore not a separate function but a phase in the ongoing regulation of behavior —inseparable from cognition, perception, action, and motivation, and indeed co-constitutive of them.

5.1.3 Early Experiments in Robotics: (GOF)AI Robotics

Toda's SFE —and its later version EFE (*Emotional Fungus Eater*) [Section 2.1], more explicitly focused on emotions[84]— has been extremely influential. The situated, non-representational, embodied and motivational conception of emotions as emergent behavior articulated in this cybernetic proposal has inspired the design of many models in CME (see e.g., [Pfeifer 1988; Pfeifer 1994]), AC (see e.g., [Picard 1995, 61]) and social and affective robotics [Section 5.3].

Cybernetics and cybernetic robotics continued to develop as scientific research programs all through the 1960s and 1970s. Other robots similar to Walter's and Shannon's artificial animals were built. For instance, "The Johns Hopkins Beast" [Figure 5.8] was engineered at the

[84]SFE's emotional conceptualization is namely more recent (see [Toda 1982a,b]) than SFE itself. However, already in the paper in which SFE is presented for the first time, Toda [1962, 182] anticipates some idea on its emergent affectivity.

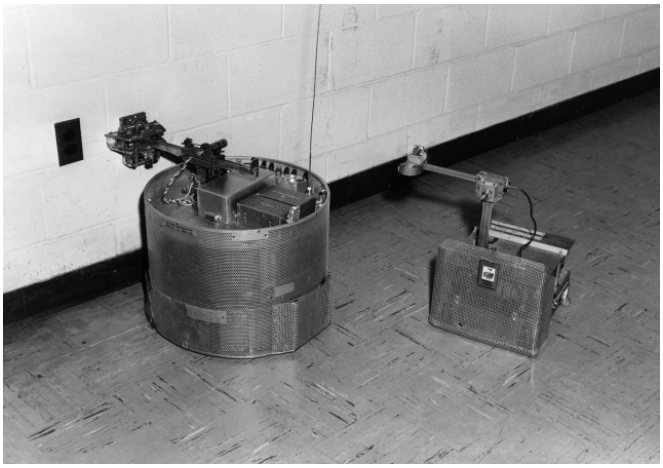


Figure 5.8: "The Johns Hopkins Beast" robot —from [Moravec 1999, 19].

Johns Hopkins University Applied Physics Laboratory in 1964, equipped with onboard analogue electronics, sonars, light sensors and tactile arms to autonomously roam white hallways, locate dark power sockets, and recharge itself when needed [Moravec 1988; 1999].

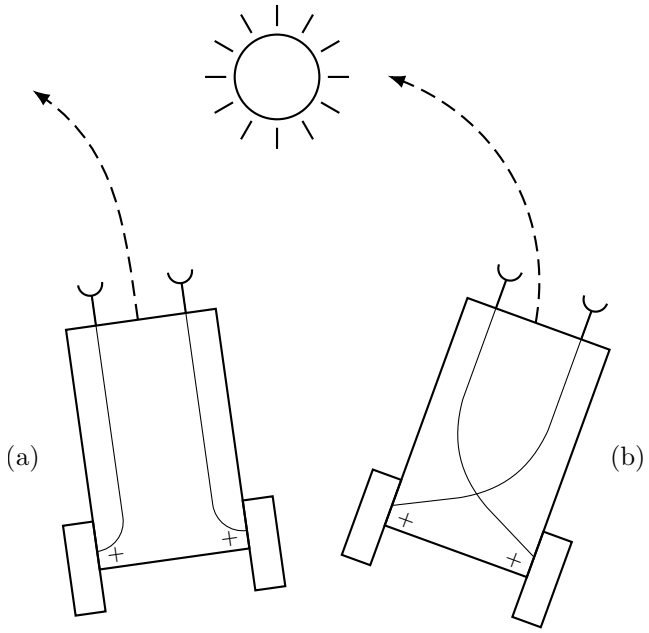


Figure 5.9: Braitenberg's [1984] vehicles (two motors and two sensors with different internal electronic configuration): Vehicle (b) is approaching the light source (photophilia, attraction); Vehicle (a) is avoiding the light source (photophobia, fear) —from [Braitenberg 1984, 8].

Or, still later in the 1980s, similar thought experiments to Toda's SFE were proposed —for instance, Braitenberg's [1984] "vehicles" [Figure 5.9], a series of minimalist robotic agents governed by extremely simple sensorimotor couplings, yet capable of producing rich (emotional) behaviors (e.g., fear, curiosity, aggression) depending on the polarity and placement of their light sensors relative to the motors, causing them to approach (*photophilic*) or avoid (*photophobic*) light sources[85].

While cybernetics' scientific endeavors extended over several decades, they did not always maintain the same momentum. Between the 1960s and the 1970s, in fact, the field began to stagnate, and cybernetics itself gradually transformed from a stable and autonomous discipline into a set of research orientations occasionally embedded within other fields such as psychology, engineering, and computer science [Nilsson 2009; Kline 2015]. This decline was largely due to the intrinsic limitations of its methodological core: Analog models and feedback-based mechanisms were effective in explaining simple behaviors like homeostasis or stimulus-response loops, but proved inadequate when applied to more complex cognitive functions such as language, memory, abstract reasoning, or planning. At the same time, the very interdisciplinary and abstract nature that had initially made cybernetics appealing soon became a source of institutional fragility, as it struggled to consolidate itself within the increasingly specialized and departmentalized structures of postwar scientific research. Its generalist epistemology —rooted in metaphor and analogy rather than formal rigor—gradually fell out of favor, particularly in a political climate where measurable outcomes and disciplinary boundaries were becoming key criteria for funding and scientific legitimacy.

For these reasons, other approaches concerned with the same questions could begin to gain ground. And indeed, from the 1950s, there was another young discipline that aimed to investigate complex biological functions through the construction of artificial systems, that is, AI [Section 1.1]. In sharp contrast with the bottom-up, feedback-driven logic of cybernetic machines, this approach emerged from a top-down model

[85] Strikingly, despite the close conceptual parallels, there is no direct mention of Toda's work in Braitenberg's writings, nor evidence that he was aware of it —even though Toda had proposed nearly identical ideas almost twenty years earlier, at a time when cybernetics was still an active and promising field of research.

of cognition and behavior, treating intelligence as a rule-based manipulation of abstract and discrete symbols. Rather than engaging with the continuous dynamics of sensorimotor interaction, it focused on formal reasoning, problem-solving and the internal architecture of thought.

Compared to cybernetics, which from the outset had been engaging with embodied artificial systems interacting with a physical environment, GOFAI tended to frame cognition as a *solipsistic* process. And still, functional connections to the external world remained, particularly in the field of computer vision [Section 3.2.7], where artificial agents were developed that not only had to process abstract mathematical theorems or diagnostic data, but also to deal with concrete stimuli and situated actions, operating, however minimally, beyond the realm of pure abstraction and formal manipulation. Symbolic architectures were forced to interface with the physical world: Data were collected from the environment through sensors, transformed into symbolic representations, and processed by a digital machine operating on a set of discrete, rule-based procedures stored in memory. These internal computations were then matched against the encoded rules, producing symbolic decisions which, once translated back into motor instructions, resulted in observable actions —such as tracking a moving object or reacting to the presence of physical obstacles.

Both responding to the decline of cybernetics, and following the attempts to design symbolic models capable of elaborating *physical* inputs and outputs, so to truly approximate human intelligence [Nilsson 2009], a new research line thus emerged: *GOFAI-inspired robotics*. This aimed to operationalize high-level cognitive architectures within autonomous, mobile agents interacting with real-world environments. Rather than challenging the disembodied nature of symbolic AI, this line of work sought to demonstrate its sufficiency in the "real" world, i.e., to show that planning, reasoning, and action selection based on internal representations could, in principle, support intelligent behavior outside purely abstract domains.

In this new research line converged researchers from AI, cognitive science and engineering, and its most notable product was *Shakey* the robot [Figure 5.10], the first general-purpose AI robot, built between

1966 and 1972 at the Stanford Research Institute (SRI) under DARPA funding [Nilsson 1984; 2009].



Figure 5.10: Shakey the Robot —from *SRI International*, 1966.

Shakey was a wheeled, mobile robot equipped with a television camera, bump sensors, and a range finder, all mounted on a tall, box-like frame. Its hardware design reflected the need for basic perceptual and navigational capacities, allowing it to detect obstacles, perceive its surroundings, and move within a constrained indoor environment. Shakey's tasks included navigating hallways, opening and closing doors, pushing boxes, and switching lights—simple operations in themselves, but non-trivial when approached from a planning and reasoning perspective.

The core innovation lay in its software architecture, which embodied the full symbolic pipeline of sense-plan-act. Perceptual data from sensors were transformed into symbolic descriptions of the environment and temporarily stored in the memory, which also contained a symbolically structured internal world model and STRIPS (*Stanford Research Institute Problem Solver*), a planning algorithm capable of generating and updating plans through logical inference. Inputs were matched against the current state of the internal world model, allowing the system to determine the relevant preconditions for goal satisfaction. STRIPS then decomposed the target objective into an ordered sequence of subgoals, each linked to a specific operator —i.e., an action schema defined by preconditions and effects. Once a consistent plan was computed, symbolic

instructions were passed to the execution system, which translated them into low-level motor commands. These, in turn, triggered concrete actions in the environment —such as turning, navigating around obstacles, pushing objects, or aligning with landmarks— while new sensory data were constantly fed back into the system to monitor progress and detect discrepancies between expected and actual states.

This modular integration of perception, representation, planning, and execution made Shakey a paradigmatic case of GOFAI robotics: a machine designed not to adapt or learn, but to reason about its environment and act accordingly, under the assumptions of a fully representational and rule-governed model of cognition.

5.2 A New Approach to Robotics

5.2.1 The Limitations of GOFAI-Robotics

Shakey marked the beginning of a sustained effort to construct autonomous machines in which symbolic reasoning and deliberative planning could directly control embodied action —other examples are Cart [Moravec 1976], HILARE [Giralt et al. 1979], Rover [Moravec 1982], CARMEL [Congdon et al. 1993], Flakey the robot (successor of Shakey) [Saffiotti et al. 1993], etc.

GOFAI-inspired robots, however, soon revealed profound limitations. Their ability to cope with external environments, in fact, crucially depended on the availability of complete, accurate and detailed symbolic representations of paths, obstacles, objects, etc. in the space. In order to reason and plan effectively, systems like Shakey, in other words, necessarily required a fully specified world model —a condition that, as it could easily be expected, was rarely met outside the controlled settings of the laboratory.

The difficulty of handling the complexity of real-world environments was further paralleled by issues of reactivity. The computational cost of deliberative planning over symbolic structures proved namely to be extremely high, leading to unnaturally long delays between perception and action, and therefore to *shaky* movements or *flaky* decisions. This lack of real-time responsiveness made these robots fragile, unable to cope with

unpredictable changes or noisy and partial inputs in variable environments. The rigidity of symbolic pipelines, along with the brittleness of their rule-based reasoning, rendered them ill-suited for dynamic physical interactions —which was exactly the task in which autonomous agents were expected to operate though.

From GOFAI's perspective, the challenges for the development of autonomous robots were numerous (see also [Section 1.1.2] and [Section 1.2.4]). Yet, perhaps the most serious hurdle was how to represent, within their cognitive architecture, the logical boundary between those properties of the world affected by the robot's actions and those that remain unchanged —a difficulty that came to be known as the *frame problem* (FP). Originally emerged and formally defined in AI by McCarthy and Hayes [1969], FP quickly extended into broader philosophical discussions, still resonating today in the debate about epistemology of common sense[86] [Dennett 1978; Shanahan 2016]. It is about something that may appear trivial to us, but is in fact deeply non-trivial for machines: You know, for instance, that when you turn on the light in your room, this affects only some aspects of the environment (e.g., illumination), but not others (e.g., the position of the chair, the color of the pen, the temperature of the computer, etc.). Indeed, representing only the actual consequences of an action, while ignoring the potentially infinite list of non-consequences, turned out to be a major challenge for symbolic systems.

More formally, using McCarthy and Hayes's [1969] *situation calculus*, FP emerges when trying to logically represent how the world changes as a result of an agent's actions. Such changes are described as transitions from situation s to situation s' —whereby s and s' are discrete states of the *entire* world at different instants of time— *resulting* from an action α performed in s by an agent p:

$$result(p, \alpha, s) = s' \tag{5.1}$$

That in [Equation 5.1] is more generally called a *fluent*. Fluents are functions that express the state of a given property f of a certain item r in a situation s, formally $f(r, s)$. For example,

[86] In philosophy and epistemology, FP has concerned for example the question about how an agent can know what is relevant in a given situation without considering everything, especially after performing an action that modifies the situation.

$$Illumination(room, s) = low \tag{5.2}$$

expresses that, in situation s, the illumination of the room is low. Beyond this kind of *functional* fluents —that can have the most disparate properties as output values— there are also *predicative* fluents, that only have outputs $(true, false)$: $LightOn(s) = true$, for example, expresses that in situation s the proposition $<$*The light is on*$>$ is true, i.e., that the corresponding fact holds in that given situation[87]. Now, if $LightOn(s) = true$ holds, and p performs in s the action $\alpha = TurnOn$, we have that:

$$result(p, TurnOn, s) = s' \tag{5.3}$$

whereby,

$$Illumination(room, s') = high \tag{5.4}$$

Of course, [Equation 5.2], [Equation 5.3] and [Equation 5.4] may amount to represented knowledge in the memory of the robot, together with, say,

$$\begin{aligned} Color(pen, s) &= blue \\ Position(chair, s) &= x \\ Temperature(computer, s) &= t \end{aligned} \tag{5.5}$$

etc. However, what may be not represented is for example that

$$\begin{aligned} Color(pen, s') &= blue \\ Position(chair, s') &= x \\ Temperature(computer, s') &= t \end{aligned} \tag{5.6}$$

etc. *Logically*, [Equation 5.3] *does not* necessarily entail that the properties in [Equation 5.6] hold, so that, for example

[87] In general, as far as I understood it, the codomain of fluents contains both properties and situations. In absolute terms (i.e., without epistemological constraints deriving from the impossibility of representing *all* properties of the world at a given instant of time [McCarthy and Hayes 1969, 18]), in fact, two situations in which objects have the *same* properties *but one*, are two different situations. For this reason, the fluent *result* in [Equation 5.1] produces a new situation as an output —whereby s', in this case, is a property attributed by "*result*" to (p, α) coherently with the fact that α produces at least one change in s.

$$result(p, TurnOn, s) \rightarrow Color(pen, s') = \neg blue \tag{5.7}$$

may still remain true. Thus, in order to guarantee that α produces exactly the changes in [Equation 5.6], and avoid changes such as that in [Equation 5.7], one has explicitly to program *additional axioms* in the robot for *each* unaffected fluent by *each* possible action:

$$\forall f \ \forall \alpha \ \forall r \ \ (f(r, s) \wedge \neg Affects(\alpha, f) \rightarrow f(result(p, \alpha, s))) \tag{5.8}$$

Assuming that there are n actions and m fluents, one would have to explicitly program at least[88] $m \times n$ such *frame axioms*, which leads to a *combinatorial*, and therefore (for the robot's computer) *computational explosion*, incompatible with the idea of an artificial agent acting smoothly in the environment as a human would intuitively do[89].

5.2.2 Rodney Brooks' *Nouvelle AI*

The practical limitations of GOFAI-inspired robotics —aggravated by deep logical issues such as FP— required a decisive shift in the history

[88] One has namely also to consider combinations of actions and fluents, combinations of combinations, etc. which makes the quantity of frame axioms potentially infinite.

[89] FP has received in logic, AI and epistemology a number of different (attempted) solutions. In logic, as we have seen, the *frame axioms* approach soon proved unworkable. For this, [McCarthy 1986] later proposed *circumscription*, a form of non-monotonic reasoning aimed at minimizing unintended changes by default. This paved the way for the broader adoption of non-monotonic logics in AI, which enabled agents to assume the persistence of facts in the absence of evidence to the contrary —an idea that, while conceptually elegant, often proved brittle in complex or dynamic domains. A major refinement came with Reiter's [2001] *successor state axioms*, which provided a way to define directly, for each property, the precise conditions under which it changes or persists across actions. This allowed to noticeably reduce the number of frame axioms needed and avoiding redundancy. In epistemology [Footnote 86], proposed solutions involved tackling FP pragmatically —accepting incomplete formalization and relying on runtime heuristics rather than exhaustive logical coverage [McDermott 1987]; interpreting FP phenomenologically as a sign that cognition is not purely symbolic but embedded in the dynamic, context-sensitive flow of real-world interaction [Dreyfus 1992; Wheeler 2008]; and applying *Global Workspace Theory of Consciousness* [Baars 1993] to FP —a theory in which consciousness is understood as the broadcasting of information to a central cognitive workspace making it accessible to multiple specialized unconscious systems, whereby thus relevance emerges from parallel processes competing for access to a shared, limited-capacity space of attention [Shanahan and Baars 2005].

of robotics towards more situated, adaptive, and environment-responsive models. Accompanied by a broader, *winterly*, erosion of trust in the cognitivist paradigm, the mind-computer analogy and GOFAI computation [Section 1.1.2]— new approaches to robotics began to take shape towards the late 1980s. Paralleling emerging embodied and situated perspectives in the philosophy of mind [Section 3.2.7], as well as Hatfield's [1988] *noncognitive computationalism* —albeit not yet in a connectionist acceptation, these new approaches remained computational, even though no longer grounded in formal rules or discrete symbolic representations.

One of the most influential figures in the shift away from classical representationalist AI robotics has been Rodney A. Brooks, leading roboticist at MIT and founder of iRobot (1990) —the company behind the development of the famous cleaning robot *Roomba* (2002).

Brooks [1989] (see also [Brooks 1990b; 1991a,b,c; 1999], and [Steels and Brooks 1995; Brooks and Breazeal 1998]), criticized the symbolic paradigm primarily for having decomposed intelligence into a multiplicity of «functional information processing units» [Brooks 1990b, 3]. Each unit corresponds to a distinct cognitive faculty —such as language, vision, or planning— and is designed to operate as an independent module, responsible for solving domain-specific sub-problems. These modules are orchestrated by a central symbolic processor that receives inputs from peripheral systems, elaborates them into symbolic representations, and converts them into actions. The system's overall behavior is provided by the *combination* and *coordination* of these discrete, symbol-manipulating components, that, alone, cannot determine any complex behavior of the system. The semantic content of the symbols is irrelevant to the central processing unit. What counts are only the consistency of the formal relationships between them, and the fact that they have a meaning for an external (human) observer —typically entities in the world, from individual objects to numbers, geometric shapes, nations, people, diseases, etc. (*Symbol System Hypothesis*, SSH).

Against this view, Brooks [1990; 1991a;c, etc.] advocated for a *nouvelle AI*, rejecting explicit symbolic representations as superfluous: «It turns out to be better to use the world as its own model» [Brooks 1991c, 140, my italics], whereby representations, if any, are not abstract, but grounded in the physical world (*Physical Grounding Hypothesis*, PGH).

This means, in other words, that the system has to be directly connected to the physical world through a set of sensors and actuators. Recalling early cybernetic approaches [Section 5.1.2], physical stimuli alone are sufficient to generate behavior, while goals, internal states, and outputs are expressed as physical actions, with no need for symbolic intermediation (see however [Footnote 43]). In the absence of abstract symbols and rules, the system has thus to be constructed from the bottom up.

In this way, such a system would avoid the rigidity of GOFAI models, above all by dynamically and time-responsively modulating their goals depending on changing environmental conditions and priorities, maintaining operational continuity despite disturbances, and exhibiting a form of autonomy that is not pre-programmed but arises from its *embodied* and *situated* design. In general, situatedness and embodiment —also inspired by Varela et al.'s [1991] embodied cognition (see [Steels and Brooks 1995], [Section 3.2.7])— are, along with the application of cybernetic principles, the most characterizing aspects of Brooks' [1991b] *nouvelle AI*.

From a technical and implementational perspective, this theoretical outline has translated into an effort to build *complete intelligent systems*, or *creatures* [Figure 5.11], robots based on new, (at least not in the traditional sense) non-representational, computational architecture: the so-called *subsumption architecture*.

Rather than being decomposed into peripheral input- and output processing units, and a central symbolic controller, intelligence is decomposed along lines of possible behaviors. In this sense, creatures are built as networks of *activity-producing systems* (APSs), which are self-contained *functional* units responsible each for a specific behavioral function (e.g., obstacle avoidance, goal-directed navigation, environmental exploration, etc.). Each APS (also referred to as a layer or behavior module) is directly connected to sensors and actuators, translating perceptual inputs into motor responses without requiring intermediation by a central representational system. These layers operate in parallel, and can modulate one another's activity depending on environmental context or task priority.

This modulation occurs according to a specific mechanism of interaction between modules grounded in the system's internal structure. A

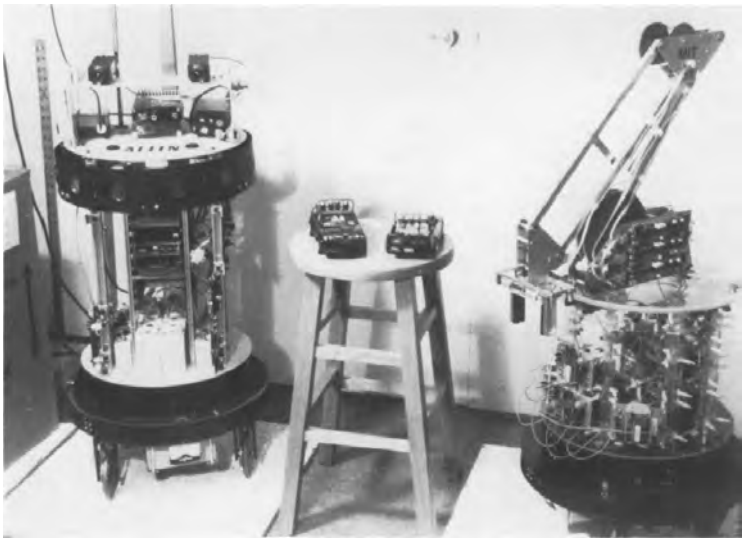


Figure 5.11: Between 1985 and early 1990s, with his collaborators at MIT, Brooks built several *creatures* that concretely exemplified the tenets of his *nouvelle AI*. *Allen* (left), for example, was the earliest one. It used off-board computation and focused on obstacle avoidance and exploration through sonar-based sensing; *Tom & Jerry* (center) demonstrated the minimal hardware requirements for autonomous behavior, running simple layered control entirely on compact programmable logic chips. *Herbert* (right), the most advanced of the group, featured onboard parallel processing and successfully navigated office environments to locate and collect empty soda cans, showcasing task-specific integration of navigation, vision, and manipulation without centralized planning or world models —from [Brooks 1991c, 151].

specific behavior of the system, functionally performed by the correspondent APS, is produced as a result of the coordinated activity of a fixed set of simple *finite state machines* (FSMs) [Figure 5.12]. A FSM is a computational device defined by a limited set of discrete states, exactly one of which is active at any given time. Transitions between states are triggered by sensory inputs, and each transition is associated with a specific output action. In the case of the behavior "`Obstacle Avoidance`", for instance, the APS is realized by a set of FSMs containing states such as "`Move forward`", "`Stop`", "`Turn Left`". The machine remains in the "`Move Forward`" state as long as no obstacle is detected; upon activation of a proximity sensor, it transitions to the "`Stop`" state; then "`Turn Left`" is activated, and once the path is cleared, it returns to "`Move Forward`". No memory or symbolic processing is involved: The complex behavior "`Obstacle Avoidance`" is fully determined by FSMs transition

dynamics between simple states, caused by continuous real-time physical inputs.

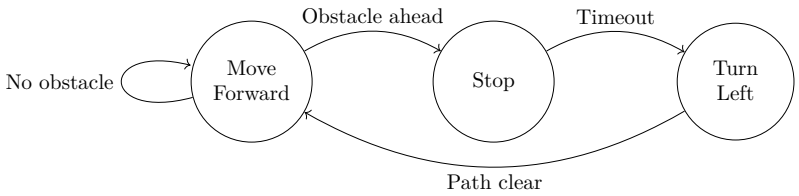


Figure 5.12: Example of a FSM representation for "`Obstacle Avoidance`" behavior in a Brooks-style subsumption-based robot. The behavior is implemented as a sequence of discrete states —"`Move Forward`", `Stop`, and "`Turn Left`"— with transitions triggered directly by sensory input. No symbolic representation or memory is involved: the system responds reactively to the physical environment in real time, illustrating how simple control structures can produce adaptive behavior.

Now, as mentioned, APSs operate independently. Yet, the overall system's behavior derives from the concurrent activity of multiple APSs, which may compete or interfere with one another. The coordination between APSs, however, is not managed by a central controller, but realized through the layered arrangement of the architecture itself. Higher-level behaviors —typically involving more context-sensitive or temporally extended responses like "`Explore Environment`" or "`Follow Target`"— modulate, or, more technically, *subsume* lower-level ones —involving simpler, reactive routines such as "`Locomotion`" or "`Obstacle Avoidance`" [Figure 5.13]. This dynamic of subsumption does not override or disable lower levels, rather it locally regulates their influence on the overall systems' behavior through two structural, complementary mechanisms: *inhibition* —which prevents a lower-level APS from activating; and *suppression* —which intercepts and replaces outputs of a lower-level APS once activated. For example, "`Explore Environment`" may suppress the ongoing activity of "`Obstacle Avoidance`" in order to carry out a deliberate change of direction, or inhibit it entirely while scanning the surroundings. In both cases, however, the lower-level APS will remain active and responsive, continuously monitoring the environment and ready to resume control as soon as higher-level demands are lifted.

In this context, hence, *intelligence* is not a sum of symbolic processes, but something that *emerges* directly from the system's structural coupling with the environment through perception and action. Brooks'

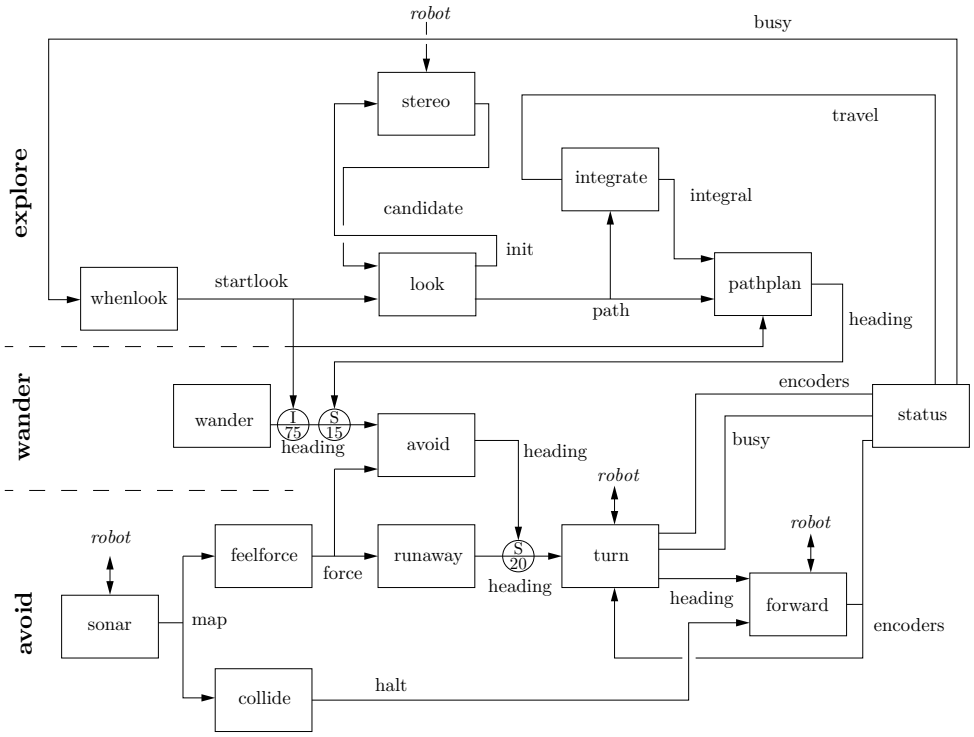


Figure 5.13: Example of a subsumption architecture organized into three behavioral layers: "`Avoid`", "`Wander`", and "`Explore`". Each layer is composed of simple FSMs (e.g., "`Sonar`", "`Wander`", "`Look`", "`Pathplan`") that process sensor data and generate motor commands. Lower-level behaviors (e.g., "`Avoid`") remain active and responsive, while higher-level behaviors (e.g., "`Explore`") can modulate them via suppression (`S`) and inhibition (`I`) signals, without centralized coordination or internal representations —from [Brooks 1991c, 152].

creatures exhibit complex *behavior* not by following pre-specified plans, but through the bottom-up, immediate, dynamics of their APSs. Similarly to cybernetic machines [Section 5.1.2], the attribution of a "mind" to such *behavior-based systems* —which in AI was traditionally tied to the presence of a centralized control unit— is not a property of the system itself, but a projection by the observer in response to its emergent, coherent behavior —a view that, for Brooks [1991c, 140ff.], also carries evolutionary significance, applying to humans themselves[90].

[90] Brooks [1991c], etc. refers however to some limitations of his approach, above all the level of cognitive complexity that robots based on the subsumption architecture can reach —if this is sufficient for generating behaviors such as smooth navigation and environment exploration, for example, it is not for language, reasoning, or social interaction. Other points concern the risk of increasing complexity and instability as more behavioral layers are added, the unclear potential for learning in fixed-topology

5.2.3 Humanoid Robotics: The COG Project

As a natural extension of the *nouvelle AI* principles already embodied in the earlier *creatures*, Brooks and his collaborators at MIT Artificial Intelligence Lab began to give concrete form to an idea that had long captured the imagination of both science fiction and AI research alike (see also [Section 1.2.3] and [Asimov 1950]): the construction of an intelligent *android* (from the Ancient Greek ἀνήρ,-δρός, *anēr,-drós*, meaning "man" and εἰδής, *eidēs*, meaning "form", i.e., "human-shaped"), that is, an intelligent *humanoid robot* (see [Brooks and Stein 1994; Brooks 1997; Brooks et al. 1998]).

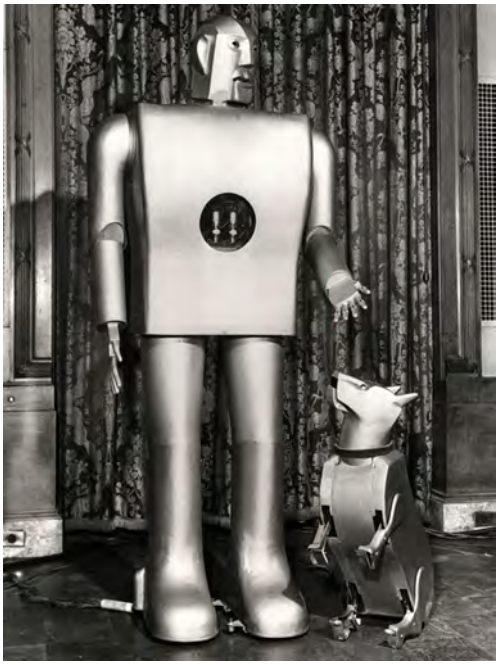


Figure 5.14: The humanoid robot *Elektro* was developed by the American company Westinghouse Electric between 1937 and 1938, and was exhibited at the 1939 New York World's Fair. It was equipped with a voice-recognition system based on audio-tone commands, a limited internal control mechanism, and several actuators; it was able to walk, move its arms, turn its head, light a cigarette, and even "speak" a set of prerecorded phrases —from *The Mansfield Memorial Museum*, Ohio (USA).

Already since the very beginnings of the field, one of the major — and more or less overtly declared— ultimate goals of AI had been that

systems, and the absence of internal representations or centralized control, which raises doubts about the coherence and intentionality of behavior.

of building an android (see e.g., [Turing 1950, 117]), as a machine that really instantiates artificially a form of intelligence comparable to that of a human being, and that is capable of flawlessly interacting with other humans and the real world as humans do. As we have seen, humanoid automation itself has a far-reaching history, tracing back at least to the 18th century[91] [Section 5.1.1]. Apart from few exhibition-oriented curiosities with loose connections to systematic scientific research —see e.g., *Elektro* (1937) [Figure 5.14]— it was only from the 1970s onward that humanoid robots began to be developed as serious objects of engineering and scientific investigation.

Crucial to this shift were major technical advancements and the decreasing cost of electronic engineering components —such as increasingly miniaturized transistors, microchip-based processors, and improvements in systems like the Global Positioning System (GPS) and the Inertial Measurement Unit (IMU). The introduction of more powerful and accurate sensors, cameras, sonars, and radars significantly expanded the perceptual and expressive capabilities of robots, enabling forms of interaction with the environment that had previously been impossible (see, e.g., Sheridan 2016; Shkurti 2023).

Also thanks to these technological breakthroughs, humanoid robotics could thus grow. In particular, it was in Japan that some of the first ambitious and technically sophisticated humanoid platforms began to appear during the 1970s and 1980s (see also [Section 8.2.1]). At Waseda University, for example, researchers started experimenting with life-sized anthropomorphic systems, called WABOTs (WAseda roBOTs), that explored the potential of new robotic technologies that would prove to be very effective and influential for later developments in the discipline (see e.g., [Kato 1974; Sugano and Kato 1987; Hashimoto et al. 1997; Morita et al. 1998; 1999; Setiawan et al. 1999], and also [Hashimoto et al. 2002; Lim and Takanishi 2007; Hashimoto and Takanishi 2018]). Around the same years, similar goals were pursued within the industrial sector —most notably by the robotics laboratories of the Honda Motor Company, which

[91] Earlier examples of humanoid automation are al–Jazarī's [1989] automata —a flute player [Section 5.1.1] and a mechanical waitress among others [Hill 1998]— as Vinci's [2008] mechanical knight (only designed but probability never built in his time) [Rosheim 2006], among others.

launched a long-term project aimed at developing autonomous bipedal machines for real-world environments, in particular, the humanoid robots of the P-series, built between 1993 and 1997 (see e.g., [Hirose et al. 1997; Hirai 1998; Hirai et al. 1998; Hirose and Ogawa 2007]).

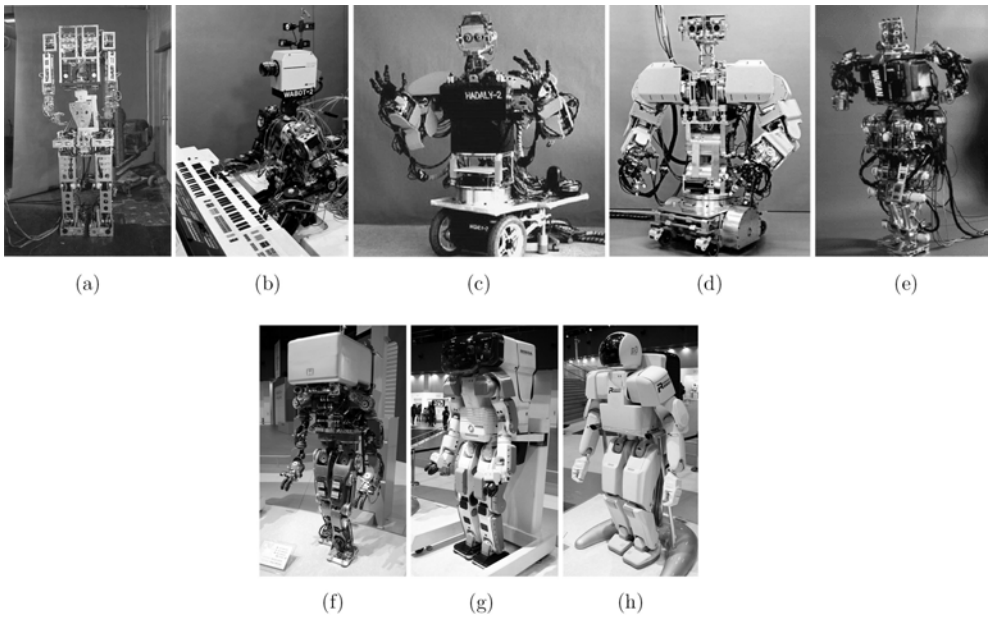


Figure 5.15: Some examples of WABOTs (top row), and HONDA P-series robots (bottom row): WABOT-1, 1973 (a), WABOT-2, 1980 (b), Hadaly-2, 1997 (c), WENDY, 1998 (d), WABIAN-1, 1999 (e), Honda P1, 1993 (f), Honda P2, 1996 (f), Honda P3, 1997 (f) —from Waseda University, 1973-1999 and "Morio" (Wikipedia User), 2008.

All these platforms (see [Figure 5.15, a-h]), whether academic or industrial, were mostly designed to emulate specific aspects of human behavior, such as speech and auditory processing, object manipulation, autonomous movement or even musical performance[92]. While early platforms employed wheeled mobility, these systems increasingly explored bipedal legged locomotion, integrating suitable engineering solutions and control algorithms for dynamic balance and the generation of human-like walking gaits, as well as articulated limb motions in item grasping and

[92] At Waseda University, several prototypes of music-playing robots were developed —notably WABOT-2 (1980), a humanoid robot capable of playing the organ [Figure 5.15, b], and WF-1 (1990), a flute-playing robot that was later refined in subsequent versions (see e.g., [Solis et al. 2006]). To the best of my knowledge, these were the first machines since the 18th century [Section 5.1.1] to combine anthropomorphic form and musical performance within a single system —a telling example of how music and musical intelligence have historically served as powerful catalysts in the evolution of AI and robotics (see also [Section 8.1] and [Section 8.2]).

displacement tasks (see e.g., [Takanishi et al. 1985; Takanishi 1991], and [Aoustin et al. 2018; Schaal 2018] for similar experiments in Europe and USA conducted in the same years).

Yet, despite their impressing technical achievements, these humanoid robots remained mostly anchored to an engineering-driven vision of robotics —concerned primarily with replicating human external appearance and movement dynamics. Their developers were less interested in using them to explore the conditions under which intelligent behavior might arise, or in treating them as scientific tools for testing hypotheses about perception, action, or cognition. Nor were they seriously attempting to «[*build*] *brains for bodies*», to couple «human-like cognitive processes to these systems», so to develop «robots that are both humanoid in form, humanoid in function, and to some extent humanoid in computational organization» [Brooks and Stein 1994, 7ff., my italics].

It was precisely in order to pursue this vision —to build a humanoid robot that was not only shaped like a human[93], but that could develop human-like *cognitive* abilities through embodied and situated *learning*— that Brooks and his team launched the COG project at MIT in the mid-1990s. COG was conceived as a concrete step toward the construction of an android in the full sense: Not merely a machine with a human form, but a platform for investigating how intelligent behavior might emerge from the dynamic coupling of body, environment, and control architecture.

To this aim, Brooks (see, in particular, [Brooks et al. 1998]) started by proposing a new theoretical framework for humanoid robotics based on psychological, ethological, system-theorical, philosophical, linguistic and neuroscientific studies about human intelligence. Rejecting SSH and the assumptions of classical AI and cognitivism [Section 5.2.2], he maintained that humans do not form monolithic and complete representations of the world, but just transient representations of task-relevant portions of it that often remain isolated and cannot fully access each other [Gazzaniga

[93] As pointed out (see [Footnote 66], as well as [Section 6.2.5]), there is no univocal definition of *human-likeness*, as it can be physical, but also behavioral, cognitive, emotional, etc. —see e.g., [Fink 2012; Kühne and Peter 2023]. For example, under certain definitions of "human-likeness", even Walter's turtles [Section 5.1.2] may be perceived as behaviorally human-like —since, like humans, they also *seek* nourishment or reward).

and LeDoux 1978; Weiskrantz 1990; Ballard et al. 1995; Rensink et al. 1997]. Also, human intelligence is not general-purpose —reasoning and decision-making are context-sensitive, domain-specific, and often shaped by emotional, perceptual, and motivational constraints [Stroop 1935; Wason 1966; Damasio 1994]. Furthermore, in humans, control is not handled by a single, central CPU-like structure (usually believed to be the brain) but emerges from multiple, distributed systems acting in parallel across body and brain [Gazzaniga and LeDoux 1978].

Given this essential characterization of human intelligence, thus, a truly *humanoid* robot is not just a symbolic processor encased in a humanoid mechanical shell, but a system whose very organization reflects the structural conditions of human cognition, in particular:

(1) *Developmental Structure.* As in humans, cognition must be built *incrementally*: Lower-level sensorimotor routines support cognitive processes of increasing complexity, serving as the functional substrate for higher competences. Rather than being pre-specified, these must emerge through incremental learning, where each stage builds on the previous one and adapts to the contingencies of real-time interaction (see [Diamond 1990; Johnson 1993; Smith and Thelen 1994; Ferrell and Kemp 1996; Ferrell 1996; Marjanović et al. 1996; Scassellati 1996; 1998]);

(2) *Physical Embodiment.* Human-like intelligence requires human-like engagement with the physical and social environment. By coupling perception and action, the system can operate without relying on internal representations, which ultimately consists in patterns of sensory-motor interaction with the world (PGH). The body also simplifies computational design, as constraints like gravity or social presence need not be explicitly programmed (see [Greene 1982; Brooks 1986; 1991c; Brooks and Stein 1994; Williamson 1998a,b], and especially [Lakoff 1987; Johnson 1987] about EC view of the mind [Section 3.2.7], also appearing in [Varela et al. 1991]);

(3) *Integration of Multiple Sensory and Motor Systems.* In humans, perception and action emerge from the coordination of multiple sensory and motor channels. Likewise, a humanoid robot must integrate vision, audition, proprioception, and motor control in real

time, enabling context-sensitive behavior without central representations and control (see [Wertheimer 1961; Knudsen and Knudsen 1985; Cohen and Massaro 1990; Churchland et al. 1994; Irie 1997; Peskin and Scassellati 1997]);

(4) *Social Interaction.* Human learning unfolds within rich social dynamics —through shared attention, imitation, and guided exploration. These interactions scaffold the acquisition of new skills and regulate developmental progress. For a humanoid robot, then, social capabilities are not only valuable for communicative and assistive purposes, but constitute necessary conditions for adaptation, the learning of complex behaviors, and the development of higher-level cognitive competences (see [Wood et al. 1976; Hobson 1993; Cohen and Volkmar 1997; Breazeal and Scassellati 1998; Breazeal and Velasquez 1998]).

COG was designed as a direct attempt to operationalize this framework [Figure 5.16]. In its physical design, the robot reproduced the upper body of a human adult, with a head, neck, torso, and two arms equipped with articulated fingers. It included stereo cameras for vision, microphones for audition, and a proprioceptive system for tracking joint positions. The entire structure was actuated by a large number of motors (over 20 degrees of freedom, DOF), allowing for coordinated movements of the head, eyes, shoulders, elbows, and hands. Sensory and motor components were distributed across multiple modules, each connected to a dedicated processor, forming a decentralized architecture closely modeled on the human sensorimotor system.

In its software, COG was explicitly designed to mirror the decentralized and layered nature of human cognition. Rather than relying on a single central controller, COG's internal processes were distributed across a network of independently running modules, each responsible for distinct sensorimotor functions. These modules operated in parallel and communicated asynchronously, reflecting the human brain's reliance on multiple specialized, concurrently active subsystems. Building on the principles of the subsumption architecture [Section 5.2.2], the system allowed low-level routines —such as gaze stabilization or reflexive arm withdrawal— to function autonomously, while higher-level behav-



Figure 5.16: The humanoid robot COG —from [Brooks et al. 1998, 59].

iors could emerge from their dynamic coordination. In this framework, cognition was not treated as an abstract computation isolated from the body, but as an emergent property of embedded, real-time interaction between sensory input, motor action, and environmental feedback.

COG was not designed to solve a specific task, but rather to explore how intelligent behavior could emerge from such interaction. Accordingly, its development focused on a set of core sensorimotor and proto-social skills that could serve as building blocks for more complex capacities. These included visual tracking of faces and objects, orientation and reaching behaviors, coordinated eye–head movements, basic arm manipulation, and shared attention cues such as gaze-following or head nodding [Section 5.3.1]. While limited in scope, these behaviors demonstrated the feasibility of artificial learning and incremental development in robots through layered control architectures.

5.3 Emotions in Social Robotics

5.3.1 Designing a Social Robot

As we have seen, COG was conceived according to an intrinsically social view of intelligence (4): Cognitive competences were expected to emerge

not only from sensorimotor interaction with the environment, but also —and crucially— through structured exchanges with other (human) agents.

Among the socially-relevant capacities effectively implemented in COG, particular emphasis was placed on perceptual and motor routines enabling basic forms of social or proto-social behavior. Among the most notable examples of this was *joint attention*, realized through integrated systems for face and eye detection, combined with gaze stabilization and visual tracking. These allowed the robot to align its visual focus with that of a human partner, thereby establishing a shared reference to external targets in the environment. Other routines involved the imitation of head nods —such as those typically used among humans to indicate "yes" or "no". By detecting periodic vertical oscillations in the user's head movements and reproducing them with its own neck actuators, COG could mirror these common nonverbal signals, introducing a basic form of responsiveness that made the robot's behavior appear more socially engaged.

Thus, even in its early stage of development, COG incorporated functionally significant elements for the real-time coordination of action and attention with human partners, supporting the emergence of interactive behaviors grounded in shared perceptual cues [Brooks et al. 1998]. However, still several (and more complex) aspects of social behavior —as described in human psychology, remained at that stage to be actually implemented in the robot, being only envisioned within its design.

One such aspect concerned, for example, the refinement of joint attention mechanisms for facilitating the learning of a *Theory of Mind* (ToM)[94] [Brooks et al. 1998, 64-66]. More robust sensorimotor control, integrated face and gesture recognition, and context-sensitive attentional

[94]The *Theory of Mind* (ToM) —variously known also as *intentional stance* (see e.g., [Dennett 1971]), *folk psychology* (see e.g., [Churchland 1981; Stich 1983]) or *mentalizing* (see e.g., [Fonagy 1991])— is the ability to attribute, based on observable cues of attention and action, mental states such as intentions, beliefs, or desires to others (human or artificial) agents, and to use such attribution to interpret and predict their future behavior [Goldman 2012; Barlassina and Gordon 2017; Hutto and Ravenscroft 2021]. Two influential explanatory frameworks have been proposed to account for how ToM operates: the *theory-theory* —which suggests that individuals use a tacit, theory-like knowledge about mental states to infer others' minds [Gopnik and Wellman 1992], and the *simulation theory* —which posits that people simulate others' mental states by using their own mind as a model (see e.g., [Gordon 1986; Gallese and Goldman 1998]). A classic empirical paradigm illustrating ToM

systems were required for making a robot develop the ability to interpret others' mental and emotional states, engage in perspective-taking, and express its own goals and intentions through socially appropriate behavior, adapting dynamically to the observed reactions and affective cues of human partners.

Another aspect moreover regarded the implementation of an internal *motivational* dynamics as a regulatory system for social behavior [Brooks et al. 1998, 66–67]. Such a system was meant to keep the robot's drives within *homeostatic* bounds and to regulate the intensity of interaction with a human partner over time. Expressive cues had to serve as readable indicators of its internal state, allowing the human partner, in turn, to adjust both the form and intensity of the exchange accordingly. Through this process, the robot would gradually learn how to influence the partner's behavior in ways that contributed to the satisfaction of its own drives and needs.

Closely related to motivation, another aspect finally concerned the use of *emotional expressions* as regulatory signals within social exchanges [Brooks et al. 1998, 78-80]. An ability to express emotions, for example by gestures and facial expressions, was required as a means to render the robot's internal condition —namely the dynamic configuration of drives, perceptions, and cognitive states at a given moment— externally available to a human partner. This would allow the latter to interpret whether the robot was, for instance, overstimulated or under-stimulated, and to appropriately adapt the form and intensity of her interaction. Emotional expression thus had to play both a social and a regulatory role, supporting mutual adjustment and contributing to maintain the robot's overall homeostasis.

---

capacity is the *false belief attribution* task, exemplified by the Sally-Anne experiment (see e.g., [Wimmer and Perner 1983; Baron-Cohen et al. 1985]), where the ability to recognize that another agent can hold a belief that is false demonstrates understanding of others' mental perspectives. ToM has been studied in relation to animals (see e.g., [Premack and Woodruff 1978]), children within the autistic spectrum disorder (see e.g., [Baron-Cohen et al. 1985]), and is currently studied also in relation to robots and artificial agents (see e.g., [Banks 2020; Söderlund 2022] and [Section 7.1]).

5.3.2 Robotic Emotions and Intelligent Sociability: AC and the Kismet Project

The attention paid by the designers of COG to aspects such as ToM, motivation, and emotion [Section 5.3.1] —though not yet translated into concrete implementations— already made clear that artificial sociability, understood as an integration of emotion, cognition, and embodiment [Chapter 3], was becoming a relevant concern in robotics —as it was during the same years across computation in general through the emergence of AC [Chapter 4].

Such a concern had been indeed largely absent before, already in GOFAI robotics [Section 5.1.3]. Robots like Shakey were mostly devoid of any emotional awareness or responsiveness and, for this reason, also incapable of true social interactions: Their sociality only consisted in merely co-existing with humans in the same environment, or at best in being the target of unilateral human-controlled master-slave dynamics (see e.g., [Shkurti 2023, 55-56]). Even if later developments in Japanese robotics [Section 5.2.3] enabled machines to recognize gestures or respond to human presence, they still fell short of implementing any form of embodied intelligence that was intrinsically affective, and thus social in a really meaningful sense.

With the evolution of the debate on the role of emotions in computation and intelligence [Section 2.3], the 1980s saw the first attempts to design artificial emotions explicitly for robots. For instance, Sloman and Croucher [1981] —recalling the cybernetic intuitions of Toda [1962] [Section 5.1.2]— proposed a (very influential) theoretical architecture for robotic emotions centered on motivation, where these are conceived as "disturbances" caused by conflicting motives or external changes, and contribute to task prioritization and adaptive behavior in dynamic environments.

Thus, from this broader historical perspective COG marked a significant shift toward *social* (or *sociable*) *robotics* (see e.g., [Dautenhahn 1998; Duffy et al. 1999; Breazeal 2002; Hegel et al. 2009; Breazeal et al. 2016; Sheridan 2020]), conceiving the robot itself as an emotionally modulated agent, embedded in social dynamics and capable, in princi-

ple, of affect- driven interaction with humans[95]. As seen before [Section 5.3.1], however, although COG had introduced such a new perspective, it remained, for the most part, a theoretical construct. The actual implementation of motivational and affective mechanisms, along with their integration into real-time responsive social behaviors, was pursued in a different project, carried out again at MIT AI Lab by Brooks' student Cynthia L. Breazeal: the *Kismet* robot (see e.g., [Breazeal et al. 1999; Breazeal 2000; 2002; 2003a,b; Breazeal and Brooks 2005]).

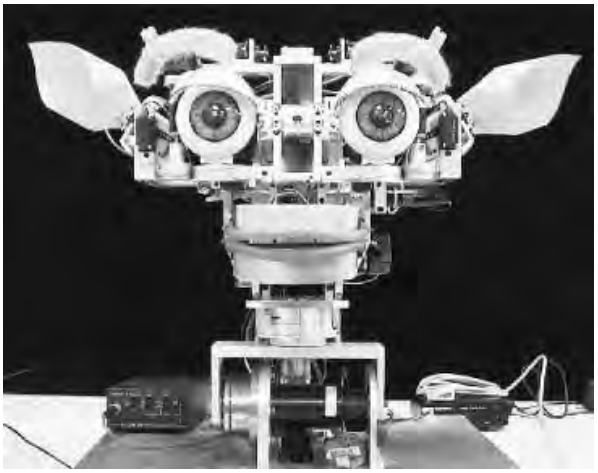


Figure 5.17: Kismet the robot —from [Breazeal 2002, 54].

Kismet [Figure 5.17] was the first robot to be explicitly designed for social interaction. Like COG, it served, on the one hand, as an epistemic tool and experimental platform for modeling and testing scientific theories of human social cognition, allowing comparisons with human behavior under controlled conditions and offering new insights into the mechanisms of our social nature. On the other hand, it explored the practical motivations and engineering challenges involved in designing robots capable of engaging naturally and effectively with human partners.

[95] A comprehensive definition of "sociable robot" has been given by Breazeal [2002, 1]: «What is a sociable robot? [...] For me, a sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It should be able to understand us and itself in social terms. We, in turn, should be able to understand it in the same social terms —to be able to relate to it and to empathize with it. Such a robot must be able to adapt and learn throughout its lifetime, incorporating shared experiences with other individuals into its understanding of self, of others, and of the relationships they share. In short, a sociable robot is socially intelligent in a humanlike way, and interacting with it is like interacting with another person. At the pinnacle of achievement, they could befriend us, as we could them».

In particular, Kismet modeled the simplest and most primitive human social interaction: the dyadic exchange between an infant and its caregiver. Building on COG's idea to connect cognitive development and social interaction [Section 5.2.3], it aimed to reproduce the learning processes that take place in this kind of relationship, especially by isolating and implementing basic interactive mechanisms characteristic of early communicative exchanges as identified in developmental psychology and ethology (see e.g., [Tinbergen 1951; Vygotsky and Cole 1978; Trevarthen 1979; Lorenz 1981; Bruner 1991]). These included the ability to direct attention toward salient stimuli in order to establish joint focus; to provide affective feedback that reinforces or inhibits behavior; to regulate arousal so that the interaction remains within an optimal range of intensity; to balance agendas between the two partners, alternating initiative and responsiveness; to repeat familiar patterns with slight variations that sustain interest and support learning; to ensure proper timing and contingency in responses; and to engage in simple game-like exchanges that provide structure and predictability [Breazeal 2002, 31ff.]

More generally, the design of Kismet was guided by a set of theoretical assumptions about what enables a robot to function as a competent social agent (see e.g., [Dautenhahn 1998]), a set "ingredients" of a sociable robot [Breazeal 2002, 6-11]. Primarily the robot must possess *life-like qualities* that promote natural interaction and *anthropomorphization* — qualities such as *autonomy*, understood as the robot's capacity to sustain its own activity and adapt to a changing environment, and *believability*, defined as its ability to project an "illusion of life" and to convey a coherent personality, allowing the human partner to interpret and predict its behavior in familiar social terms. Closely related to this is the requirement that the robot be endowed with a *ToM* [Footnote 94] —that is, the ability to infer intentions, beliefs, and emotions of a human partner from observable cues (gaze, expression, prosody, gesture, etc.). Correspondingly, it should also be capable of accessing, monitoring, and representing its own intentions, motivations, and affective states in socially meaningful terms, while ensuring that its outward behavior remains readily *readable* by a human observer. This means, the robot must produce a suitable series of responses that enable humans to attribute emotions to it (*affective responses*), to make their attention converge towards objects

and events that are interesting for it (*exploratory responses*), to notice signals of discomfort or distress in it in ways that would elicit protective behaviors (*protective responses*), and to adjust the overall rhythm and intensity of the exchange with it to maintain a mutually engaging interaction (*regulatory responses*) [Breazeal 2002, 36].

Beyond autonomy, believability, ToM, readability and responsiveness, a foundational property of any social robot is finally *embodiment*:

> «Humans are embodied and situated in the social world. We ground our experiences through our body as we interact with the environment and with others. As such, our bodies provide us with a means for relating to the world and for giving our experiences meaning [Lakoff 1987]. Brooks has extensively argued for the importance of embodiment and being situated in the world for understanding and generating intelligent behavior in animals and robots [Brooks 1990a]. Socially intelligent robots can better support these human characteristics if they are embodied and socially situated with people» [Breazeal 2002, 7].

Like already COG [Section 5.2.3], the robot's body is what structures and sustains social interaction. Embodiment enables the robot to participate in the social world by grounding its behavior in sensorimotor processes that unfold in real time. Through posture, gaze, facial orientation, and timing, the robot becomes intelligible and responsive within the interactional space it shares with human partners. The body is not a mere vessel for internal states but the medium through which the robots cognizes the world, organizes perceptions and behaviors, expresses emotions and coordinates social cues. Therefore, sociality cannot be decomposed into discrete computational modules, nor reduced to the manipulation of symbolic representations detached from context. Rather, it must be *enacted* —continuously, contingently, and corporeally— through the dynamic interplay between the robot's internal states and its environment.

This emphasis on embodiment in [Breazeal 2002] was clearly an epistemological emanation of the robotic paradigm advanced by Brooks et al. [1998] and the school of MIT AI Lab[96] [Section 5.2.2]. At the same time, however, it explicitly resonated with the broader reconceptualization of intelligence introduced by EC (see [Section 3.2.7] and [Section 3.2.8]): no longer conceived as an abstract inferential faculty, but as something that emerges from the agent's embeddedness in a world that is simultaneously

[96]The design of Kismet in particular, however, has been apparently less influenced by Brooks' approach based specifically on the subsumption architecture.

physical, affectively valenced, and intersubjectively structured. But, first and foremost, Breazeal's [2002] approach represented the most coherent fulfillment of AC's program [Chapter 4]. Its conceptual grounding clearly reflected the influence of Picard [1997] in multiple respects: the emphasis on the social and interrelational role of emotions [Section 3.3]; the idea of computational affects not as discrete states, but as internal processes regulating salience, attention, and behavior within situated interactions; and, not least, the institutional proximity to Picard's own MIT Media Lab. All these elements testified to the direct continuity between AC and Breazeal's model of social robotics. Already Picard [1997, 72] envisaged the centrality of artificial embodiment in shaping affective experience, conceiving affective robots as the highest achievement of affective computing —after all, «robots provide the most obvious example of a computer with a body». Yet, it was Breazeal [2002] who actually translated this vision into a *concrete* design framework, making embodiment not a peripheral feature, but a functional core of artificial sociability.

Thus, within this theoretical framework, Kismet came to be articulated as an integrated interactional, affective and embodied model of artificial agency. Currently housed at the MIT Museum, the robot consisted of a stationary head-like platform equipped with a rich set of sensors and actuators, designed to support expressive, face-to-face interaction. Its hardware architecture was deliberately anthropomorphic, incorporating moveable eyes, eyelids, eyebrows, lips, neck, and ears —each actuated by dedicated motors to convey nuanced facial expressions. Visual perception was enabled by four color cameras, providing both foveal and peripheral input, while auditory input was captured through a directional microphone. A speaker, embedded in the mouth area, allowed for synthetic vocalizations. These components were mounted on a three-DOF neck, allowing the robot to orient its head toward stimuli, perform gaze shifts, and express attention or disengagement —all crucial elements in infant-caregiver interaction.

The expressive capabilities of Kismet's hardware were supported by a dedicated software architecture —the *synthetic nervous system* (SNS) [Figure 5.18]— designed to coordinate perception, attention, emotion, and behavior in real time, thereby enabling the robot to produce co-

herent, responsive, and affectively modulated interactions with human partners.

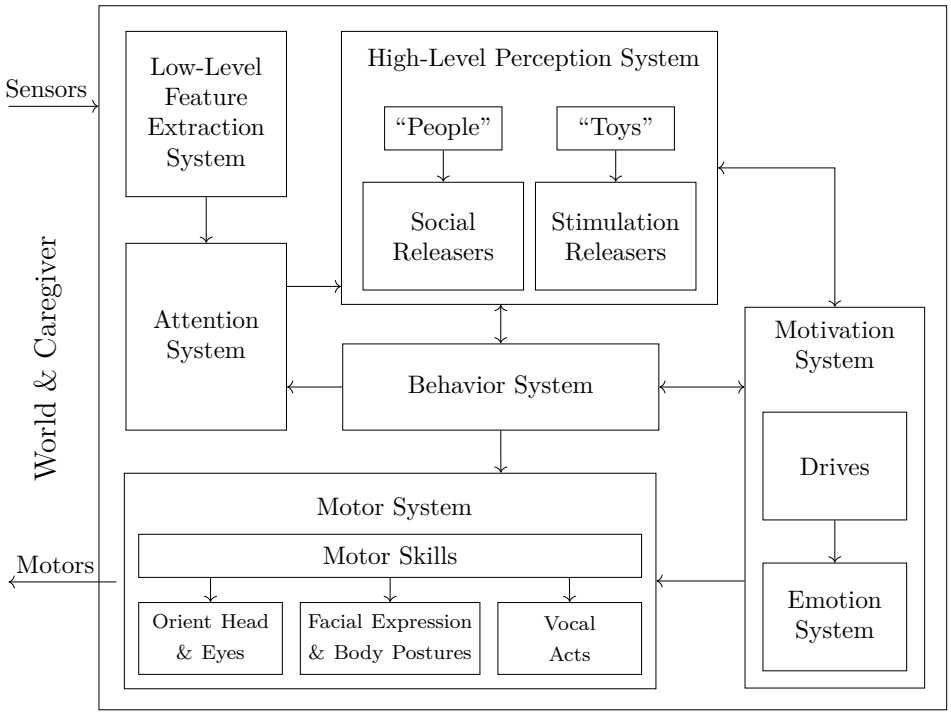


Figure 5.18: Overview of Kismet's *Synthetic Nervous System* (SNS) —adapted from [Breazeal 2002, 44].

The SNS is implemented as a *distributed* network of six highly *interconnected* subsystems for low-level feature extraction, attention, perception, motivation, behavior and motor control. Even if not formally structured as a properly defined ANN [Section 3.1], the system is nonetheless based on a *subsymbolic* logic of distributed activation and competitive integration between parallel modules (see [Lorenz 1981; Minsky 1985; Brooks 1986; Maes 1991], as well as [Section 4.2.1]). Each subsystem comprises multiple computational units whose activation thresholds determine their responsiveness to incoming signals —which may originate from sensory input, internal drives, or other modules. Behavioral output is thus not pre-programmed, nor selected by any central controller, but instead emerges from the dynamic equilibrium of activation patterns evolving within the system. In this respect, the SNS displays clear analogies with early connectionist models —especially in the way activation levels are computed, propagated, and modulated across units and sub-

systems (see [Equation 3.3] and [Breazeal 2002, 46]). At the same time, the absence of mechanisms such as synaptic weight adaptation or backpropagation, as well as the presence of functionally modular subsystems with specialized roles, distinguishes it from canonical ANN architectures. Rather than operating as a linear sequence of discrete processing stages, the SNS organizes interaction through continuous and recursive feedback loops between perception, emotion, motivation and behavior —a design that resonates not only with the principles of connectionism, but also with the cybernetic models of adaptive control systems [Section 5.1.2]. For these reasons, the SNS can be interpreted as a hybrid architecture situated between classical agent-based systems and connectionist networks —borrowing the modular organization of the former and the activation-based, emergent behavior of the latter. While Kismet does not perform learning or backpropagation as typical ANN models do, it nonetheless exemplifies a form of embodied AI that is deeply coherent with the principles of connectionism, AC, and EC alike.

The first processing layer of the architecture is the `Low-Level Feature Extraction System`, tasked with transforming *multimodal* raw sensory inputs (visual and acoustic) into perceptual quantities of behavioral significance. This system constitutes the crucial perceptual substrate upon which higher-level processes operate. On the visual side, the system includes modules for detecting motion, color saturation, luminance, and skin tone, as well as specialized routines for localizing facial features such as eyes and mouth. On the auditory channel, the system detects prosodic contours and amplitude dynamics to infer vocal affect (see also [Section 8.2.1] on this aspect). Broadly based on the computational mechanisms for sampling, feature extraction, normalization, etc. described in [Section 4.1.2.2] and [Section 4.2.1], all these routines operate in parallel and are implemented as independent feature maps, producing spatial distributions of salience that are later integrated by other subsystems.

The `Attention System` receives the low-level visual features and selects the most salient stimuli at each moment, establishing a dynamic focus that guides the robot's gaze and orients behavior. Saliency is computed by integrating multiple factors —such as sudden appearances (e.g., a hand entering the visual field), strong visual contrasts (e.g., bright colors or movements), or features that have been associated with important

events (e.g., a human face or a toy frequently used in interaction). These inputs are combined into a single winner-take-all saliency map that continuously determines where the robot should look.

The `High-Level Perception System` receives the features selected by the attention system and organizes them into behaviorally meaningful *percepts* —internal signals indicating that a relevant stimulus has been recognized in the environment. These percepts are detected via specific *releasers* —minimal sets of conditions corresponding to configurations of multimodal stimuli that signal the presence of objects or events of social or motivational relevance in the environment. Releasers are of two kind: `Social Releasers` —for features such as faces, eyes, and expressive prosody, and `Stimulation Releasers` —for non-social objects like colorful toys or moving items. Percepts formed in this system, as structured representations of the environment, are then fed into both the `Behavior` and `Motivation Systems`, enabling coordinated affective and interactive responses.

The `Motivation System` ensures that the behavior of the robot is not simply reactive to external stimuli, but organized around internal needs and affective tendencies that evolve over time [Section 5.1.2]. In Kismet, this system integrates two interdependent components —`Drives` and `Emotions`— whose interaction regulates attention, behavior, and expressive output within dynamic social exchanges. Firstly, `Drives` are modeled as *homeostatic* mechanisms of internal regulation and control [Section 5.1.2]. They represent three basic needs of the robot, in particular: `Social` —the need to engage in reciprocal interaction with human partners; `Stimulation` —the need to receive sufficient sensory input from the environment; and `Fatigue` —the need to disengage and recover after sustained activity or overstimulation. Each `Drive` is associated with an *activation energy* A_{drive}, which quantifies its current *intensity*. This value fluctuates within a bounded range $[A_{drive}^{-max}, A_{drive}^{+max}]$, defining its *homeostatic regime* —that is, the preferred operational window for the corresponding `Drive`. When A_{drive} falls below or rises above this range —due to either environmental or internal causes, the system experiences *under-stimulation* or *over-stimulation*, respectively, and generates compensatory behavior aimed at restoring *homeostasis*. For example, if the `Social Drive` drops below its homeostatic threshold, the robot becomes

increasingly motivated to seek interaction: The `Attention System` will be accordingly directed toward faces, voices and other social cues that may satisfy the need.

The robot's ability to satisfy its `Drives` is directly reflected in its *affective state*. The `Drives`, namely, continuously influence the `Emotion System` by modulating its activation levels: When a `Drive` remains within its homeostatic regime, for example, it reinforces affective states of positive valence and stable arousal, thus reflecting a condition of internal well-being. `Emotions` are hence modeled as *functional states*, whose role is to adjust the robot's relation to the environment, motivating the approach toward conditions that support internal well-being, and fostering the avoidance of those that threaten it.

From an operational perspective, `Emotions` in Kismet derive from an *affective appraisal* of a *precipitating event* that generates *multimodal characteristic expressions* (face, voice, posture), along with *action tendencies* motivating *behavioral responses* [Breazeal 2002, 112-114]. Dimensions of affective states concerning phenomenology and physiological activity are instead deliberately not operationalized in the model.

`Emotions` are generated by the `Emotion System` [Figure 5.19], whose routine is started by inputs received from the `High-Level Perception System` —that is, behaviorally and *affectively* meaningful *percepts*, organized by *releasers* as explained above. Importantly, the organization of raw data into percepts is also modulated by the currently active `Drive`, the status of ongoing behavior, and by the affective state of the system, so that, for example, a human face may elicit positive or negative valence depending on whether the `Social Drive` is currently within its homeostatic regime or not.

Then, once perceptual inputs are received, the `Emotion System` starts a subroutine for *affective appraisal*, in which percepts are evaluated according to their affective significance. This significance is modeled as a `Somatic Marker` (SM) [Section 3.2.7], a minimal representational format assigned to each releaser, which encodes the affective relevance of the percept as a triple-valued vector $[A, V, S]$. The first component, arousal (A), reflects the estimated intensity or activation level associated with the stimulus —with positive values indicating high energetic impact and negative values indicating passivity or low stimulation. The second, va-

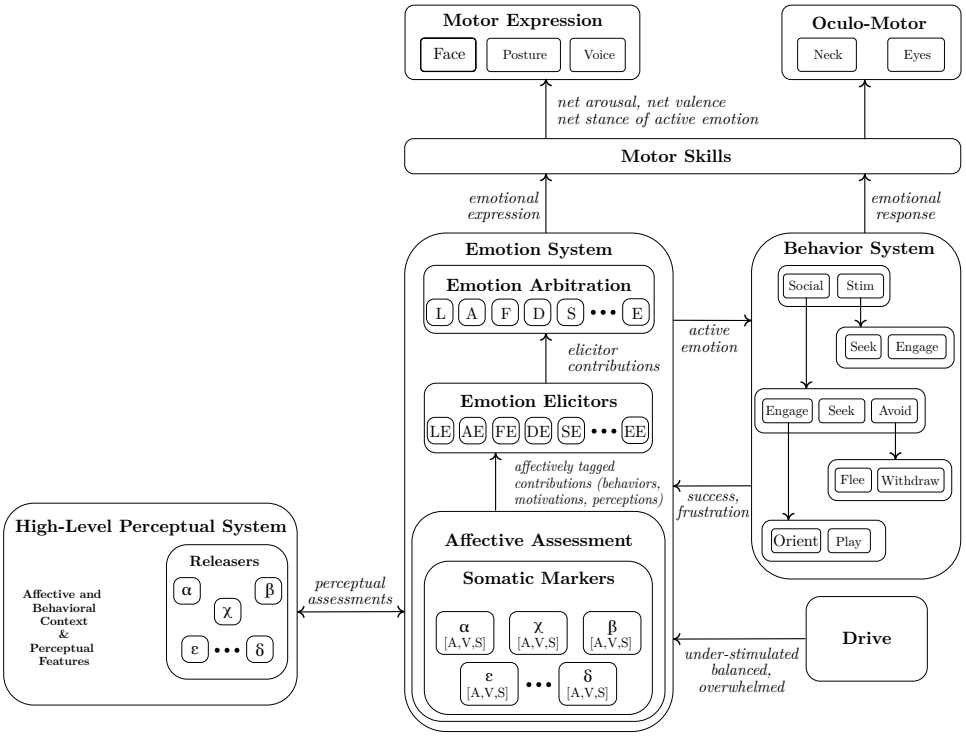


Figure 5.19: Overview of Kismet's `Emotion System` with direct connections to the `Behavior System` and `Motor System` —adapted from [Breazeal 2002, 113].

lence (V), signals the hedonic quality of the stimulus, i.e., whether it is conducive or detrimental to the robot's internal regulation —a positive V implies the stimulus is appraised as rewarding or beneficial, while a negative one marks it as unpleasant or harmful. The third component, stance (S), finally expresses the behavioral orientation prompted by the stimulus —with positive values associated with approach or engagement tendencies, and negative ones corresponding to withdrawal or avoidance. Of course, these values do not arise solely from the sensory features of the stimulus, but are also influenced by the robot's internal state.

Once a SM has been assigned, it is forwarded to a set of specialized computational modules, the `Emotion Elicitors`, whose role is to detect the emergence of an emotional state by evaluating whether the affective configuration encoded in the $[A, V, S]$ vector matches the characteristic profile of a specific emotion. Each elicitor corresponds to one of the nine discrete emotions implemented in Kismet's repertoire —namely, `interest`, `calm`, `boredom`, `joy`, `sorrow`, `fear`, `anger`, `disgust`, and `surprise`

[Ekman 1992]. They operate as continuously active filters, each selectively responsive to specific $[A, V, S]$ configurations that reflect the affective profile of the corresponding emotion as defined within a bounded region of a *multidimensional affective space* [Figure 5.20] (see also [Section 3.3]). The degree of match is not binary but graded, and it is quantified by the internal activation level of the elicitor process, denoted as E_{emotion}, which indicates how strongly a given percept resonates with the affective template of that emotional state. For example, a percept exhibiting high arousal, negative valence, and a retreat-oriented stance may raise the activation of the elicitor for `fear`, whereas one marked by low arousal, positive valence, and an approach-oriented stance may contribute to the activation of `interest`.

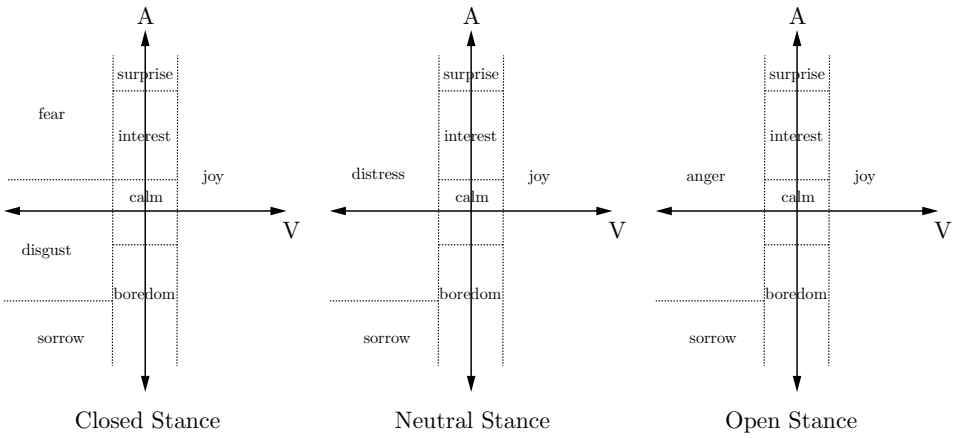


Figure 5.20: Visualization of the affective space used in Kismet's emotion model, where specific configurations of arousal, valence, and stance ($[A, V, S]$) are associated with discrete emotional states. The figure presents two-dimensional slices of the three-dimensional affective space, illustrating how distinct regions correspond to individual emotions based on their characteristic affective patterns —from [Breazeal 2002, 116].

Subsequently, the value E_{emotion} computed by elicitors is used in the activation process of the corresponding `Emotion`. Rather than functioning as discrete switches, emotions in Kismet are modeled as *continuous* processes whose degree of activation evolves over time, depending on multiple interacting variables [Section 4.1.2]. Each `Emotion` has thus an *activation threshold* A_{emotion}, updated at every computational cycle through the weighted sum of several components, as expressed by the following equation:

$$A_{\text{emotion}} = \sum (E_{\text{emotion}} + B_{\text{emotion}} + P_{\text{emotion}}) - \delta t \tag{5.9}$$

whereby E_{emotion} is the input provided by the elicitor, B_{emotion} is a baseline bias that reflect the sensitivity with which the `Emotion` can be activated, P_{emotion} reflects the level of its persistence and δt accounts for its natural fading over time.

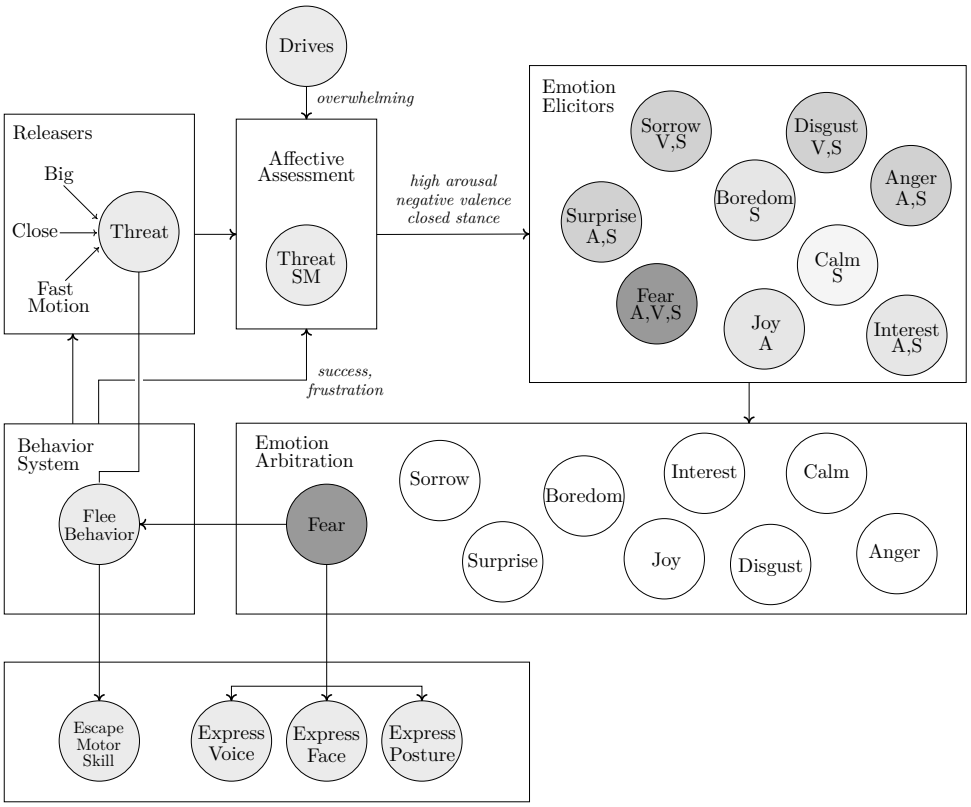


Figure 5.21: Diagram showing the emotion pathway for `fear` generation. A threat releaser is evaluated in the affective assessment stage and tagged by the `Somatic Marker` (SM) with high arousal, negative valence, and a withdrawal-oriented stance. This affective vector is then evaluated by all emotion elicitors. Only the `fear` elicitor (marked in darker color tone) fully meets all three criteria, resulting in maximal activation. The emotion `fear` is then chosen via the arbitration process and routed to both the `Behavior System` (initiating `escape`) and the `Motor System`, which controls the coordinated expressive outputs in facial, vocal, and postural channels —adapted from [Breazeal 2002, 119].

Finally, the set of activation values A_{emotion} associated with each emotion is evaluated by an *arbitration mechanism*, whose function is to determine which `Emotion`, if any, should become dominant at a given moment. Since multiple `Emotions` can be partially active in parallel, the

system compares their current activation levels and selects the one with the highest value exceeding a predefined threshold. The dominant `Emotion` then modulates the robot's expressive behavior and motivational priorities, while the remaining emotions remain latent. The arbitration process ensures that the emotional state expressed by the system is both coherent and contextually relevant, while also enabling flexible transitions between emotions as perceptual and internal conditions evolve.

`Emotions` are thus produced as *internal* outputs of the `Motivational System` [Figure 5.21]. Their emergence is shaped over time by the system's dynamic integration of perceptual, affective, and motivational factors, constrained by the characteristic profile of each emotion and the contextual contingencies of the interaction [Table 5.1].

Once selected by the arbitration process, these internal emotional states, along with active `Drives`, exert a top-down influence on the `Behavior System`. This subsystem serves as the central coordinator of the robot's behavioral repertoire, translating affective and motivational states into structured, context-sensitive action. `Behaviors` in Kismet (e.g., `Engage People`, `Reject Toy`, `Sleep`, etc.) are implemented as modular, hierarchically organized, routines that run in parallel and compete for expression based on relevance and salience [Section 4.2.1]. Each `Behavior` monitors its own enabling conditions —which may include perceptual inputs, motivational triggers, and affective cues— and dynamically signals its readiness to execute. At any given time, the system selects the most appropriate behavior via an arbitration process that ensures responsiveness, coherence, and adaptability in real-time interaction.

The selected `Behavior` is then passed to the `Motor System`, which converts it into specific expressive actions across multiple channels: gaze direction, head orientation, posture, facial expression, and vocal output. Each of these channels is modulated by a dedicated motor skill controller, allowing for fine-grained coordination between subsystems. Crucially, the robot's *expressive behavior* is not merely functional but deliberately designed to be readable by a human partner. Emotional states, intentions, and internal needs are rendered externally visible through coordinated facial expressions and prosodic contours [Figure 5.22]. For instance, the activation of `joy` elicits raised eyebrows, open eyelids, upturned lip cor-

Antecedent Conditions	**Emotion**	**Behavior**	**Function**
Delay, difficulty in achieving goal of adaptive behavior	anger, frustration	display-displeasure	Show displeasure to caregiver to modify his/her behavior
Presence of an undesired stimulus	disgust	withdraw	Signal rejection of presented stimulus to caregiver
Presence of a threatening, overwhelming stimulus	fear, distress	escape	Move away from a potentially dangerous stimuli
Prolonged presence of a desired stimulus	calm	engage	Continued interaction with a desired stimulus
Success in achieving goal of active behavior, or praise	joy	display-pleasure	Reallocate resources to the next relevant behavior (eventually to reinforce behavior)
Prolonged absence of a desired stimulus, or prohibition	sorrow	display-sorrow	Evoke sympathy and attention from caregiver (eventually to discourage behavior)
A sudden, close stimulus	surprise	startle	Alert
Appearance of a desired stimulus	interest	orient	Attend to new, salient object
Need of an absent and desired stimulus	boredom	seek	Explore environment for desired stimulus

Table 5.1: The table provides a synthetic overview of the typical conditions under which each of Kismet's emotions tends to be elicited, alongside the corresponding observable behaviors and their functional role within the system. The antecedent column specifies the perceptual configurations likely to give rise to a given emotional state; the behavioral column indicates the expressive or motor output associated with that state—ranging from facial expressions to full behavioral routines such as withdrawal. Finally, the functional description clarifies the adaptive purpose each emotion serves in regulating the robot's interaction with its environment —from [Breazeal 2002, 111]

ners, and a bright vocal tone, whereas `fear` triggers a backward head tilt, wide eyes, a frowning brow, and a high-pitched startled sound. These expressions are grounded in FACS [Ekman and Friesen 1978] (see [Section 3.3]), and mapped onto Kismet's motor schema to maximize their social recognizability and communicative effectiveness.

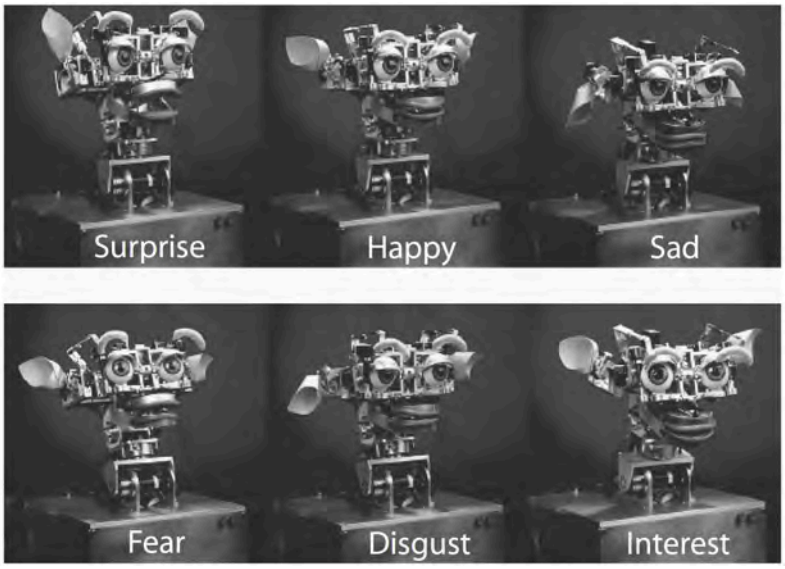


Figure 5.22: Facial expressions of Kismet —from [Breazeal and Brooks 2005, 281].

Crucially, expressive behavior in Kismet performs a twofold function. On the one hand, it serves to manifest the robot's internal states, thereby facilitating readable and affectively congruent social exchanges. On the other, it operates as a channel of continuous feedback for the human partner, modulating the intensity, tempo, and emotional tonality of the interaction. Expressions, in this sense, are not conceived as mere by-products or terminal outputs of internal processes, but as constitutive elements of the interactional loop itself: They orient the interlocutor's responses, draw attentional engagement, and sustain the rhythmic contingency of dyadic coordination. Through such an architecture —in which perception, motivation, affect, behavior, and expression are deeply interlinked— Kismet enacts a mode of embodied social interaction that is not only dynamic and emotionally grounded, but, above all, *adaptive*.

Far from being pre-defined or statically programmed, in fact, the robot's responses emerge through the recursive coupling of its internal states with the caregiver's actions, giving rise to a form of interactional plasticity that, while not grounded in representational memory or ANN-like weight adaptation, nonetheless supports a rudimentary, but functional, process of *socially situated learning*. Within the SNS architecture, learning unfolds as a non-symbolic modulation process: The caregiver's reactions operate as forms of social reinforcement, gradually

influencing Kismet's attentional biases, affective dispositions, and behavioral inclinations. As drives are either fulfilled or thwarted, and as emotional responses are reinforced or attenuated via expressive feedback, the robot continuously recalibrates the salience of environmental stimuli, the temporal structure of its reactions, and the motivational hierarchy of its internal dynamics. This process does not rely on stored representations or explicit updating algorithms. Rather, learning emerges from the ongoing regulation of synthetic nervous activity —a continuous fine-tuning of perceptual-affective linkages and interactional patterns that, in a rudimentary sense, parallels early social development in human infants. Thus, learning is not external to behavior, but embedded within the very operational logic of the system, allowing for the stabilization of successful interactional configurations through homeostatic regulation and emotional coherence.

It is precisely through this continuous, embodied negotiation that Kismet has realized its most distinctive contribution to the development of robotics: modeling sociality not as a set of symbolic competences, but as a developmental process —affectively regulated, temporally situated, and enacted through shared interaction.

Summary

Kismet marked a pivotal moment in the development of affective and social robotics, providing one of the first concrete implementations of a robot capable of leveraging emotions and cognition for engaging in real-time interactions with humans. It demonstrated how motivations, emotions, and perceptions could be computationally orchestrated to produce lifelike, socially meaningful behavior. Through its combined scientific and engineering value, it laid the conceptual and technical groundwork for a broad lineage of research in robotics and HRI, whose products —such as NAO, Pepper, iCub, and other social robots— are today widely employed in educational, therapeutic, and experimental settings [Section 6.1.1]. While these developments lie beyond the historical scope of the present work, they clearly reflect the enduring influence of the Kismet project on the evolution of social robotics as a research field.

From a broader historical-epistemological perspective, the discussion of Kismet closes the trajectory that has guided [Part I] of this thesis —a trajectory that has followed the progressive inclusion of emotions within the computational sciences, from their early exclusion as irrational residues of mind to their eventual integration as functional, embodied, and interactive components of artificial cognition. From the speculative theorizing of GOFAI pioneers to the more operational frameworks of AC and behavior-based robotics, emotions have gradually been redefined as computationally tractable, functionally significant, and architecturally indispensable to the development of machine intelligence. This transformation has unfolded across both symbolic and subsymbolic paradigms, encompassing hybrid models, synthetic architectures, and neuro-inspired systems.

As more than a chronological genealogy of models and technologies, [Part I] has attempted a reconstruction of epistemological frameworks and ideas, revealing how computer science has continually redefined its own notion of mind and agency through the progressive integration of affect. In this reconstruction, [Chapter 5] has shown how it is particularly within robotics that artificial affectivity has found one of its most fertile grounds, becoming a constitutive mechanism of adaptive and social behavior in physical systems. In this sense, the emergence of social robotics has represented the point of convergence between the theoretical lines opened by *Affective Computing* and the embodied paradigm inaugurated by the *Nouvelle AI* —a synthesis in which emotions operate not only as computational variables, but as infrastructural forces mediating between perception, cognition, and social interaction.

Yet this reconstruction has also revealed unresolved epistemological tensions in the implementation of artificial emotions. With the crystallization of AC as an autonomous discipline, emotions have become computationally tractable objects —measurable, classifiable, and reproducible through multimodal algorithms, sensors and actuators— but at the cost of a progressive simplification of their conceptual richness. This, more precisely, has led to a form of *physiological-behavioral reductionism* about emotions that still characterizes contemporary approaches to artificial affectivity, mostly overlooking its cognitive and intentional depth [Section 4.3].

Building on these historical and epistemological results, in the next [Part II] a philosophical analysis will be undertaken. This analysis is primarily aimed at analyzing and problematizing how such reductionist presuppositions have shaped, beyond CME, also implementational practices in contemporary social and affective robotics —an analysis that will be carried out in [Chapter 6]. Achieving a clearer understanding and contextualization of the physiological-behavioral reductionism about emotions in robotics, and of the interpretive and functional problems it generates, will constitute the first step toward developing broader conceptual frameworks of emotion for the discipline. In this respect, the advantages of reconsidering the cognitive dimension of emotion —particularly its propositional and inferential aspects— as a fundamental component for modeling, interpreting, and designing emotional interactions in HRI will be explored in [Chapter 7]. In turn, in [Part III], these theoretical insights will be translated into an experimental and design-oriented framework through the development of the *EEVAN Project*, where, in particular, the potential of music and sound to enhance machine affectivity will be investigated.

This articulation ultimately reflects the general goal of the present work: to understand, from a historical perspective, the emergence of current implementational practices in social and affective robotics, to critically reassess their limitations, and to open new paths toward more affective and sociable machines by moving beyond surface-level expressivity toward a richer and more nuanced account of robotic emotion —one that better captures the complexity of human emotional life and the epistemic conditions under which emotions are attributed to artificial agents.

Part II

Philosophy of Robotic Emotions

Chapter 6

Current Perspectives on Robotic Emotions

[Part I] has attempted to historically and critically reconstruct the developments that have brought emotions to become a computational object. Also, it has traced the broad theoretical and technical evolutions in the way of modelling emotions computationally between the 1960s and the 1990s. From GOFAI to affective and social robotics, it has shown how affectivity has been operationalized and artificially reproduced through symbolic programming, statistical learning, multimodal integration, embodied interaction, and intelligent sociability.

Furthermore, by connecting these developments to broader trends in philosophy, psychology, medicine, neuroscience, and computer science, it has highlighted how certain theoretical conceptions of emotion gradually consolidated within computation, shaping later attempts to build emotionally responsive machines. In particular, [Section 4.2.2] has shown how computational approaches to affect have progressively taken a tacit orientation toward a form of *physiological-behavioral reductionism* (PBR) about emotion, which seems even to persist in the field to these days [Section 4.3]: Despite a strong tradition in CME and early AC grounded in affective cognitivism, the modeling of emotion —both in terms of expression and simulation, and in terms of recognition— has been largely revolving around a conception of affectivity primarily linked to facial expressions, physiological changes, and observable behaviors.

In so doing, this historical effort has provided a solid understanding of the theoretical frameworks underpinning contemporary research on computational emotions, and of how and why they have developed and come to be adopted. What emerges from this critical reconstruction is not only a cumulative picture of the various technical strategies through which affect has been modeled artificially, but also a mapping of the different conceptual presuppositions that silently sustain such strategies.

Now, it is precisely at this juncture that a *philosophical inquiry* becomes necessary. By interrogating the assumptions (often implicitly) embedded in these frameworks, philosophy allows to assess their scope and limits, to highlight what dimensions of emotional life they tend to exclude or oversimplify, and to explore whether alternative conceptualizations might offer a richer basis for future research. Conducting such an inquiry is the major aim of [Part II]: By taking the historical reconstruction as its indispensable point of departure, it tries to critically reframe the genealogical and epistemological roots of current approaches in computational emotion modelling, individuating their present constraints and sketching ways for possibly moving beyond them.

Within this broader philosophical project, the present chapter focuses specifically on the domain of *social and affective robotics*. While the reductionist tendencies outlined in [Part I] concern computational approaches to affect in general, here the analysis turns to examine how these tendencies have concretely taken shape in robotics, and what consequences they bear for the design and interpretation of emotionally responsive robots[97].

Thus, drawing on the examination of relevant scientific literature, the first aim of this chapter is to show that emotional PBR in computation has also come to characterize present-day social and affective robotics. This is pursued in [Section 6.1] through the analysis of recent review papers surveying research in social and affective robotics over the past twenty-five years. What emerges from this analysis is that, across different platforms and implementations, emotions have been predom-

[97] Presumably, many of the philosophical analyses and conclusions developed with respect to social and affective robotics will also hold for affective computing more generally.

inantly conceived and operationalized in behavioral and physiological terms, thereby consolidating a reductionist orientation in the field[98].

Further considerations follow in [Section 6.1.1] concerning the historical reasons why the theoretical framework of emotions in social and affective robotics may have gradually taken shape in a PBR direction, including conceptual inheritances from affective computing as well as practical constraints related to design and implementation.

A further aim of the chapter is then to consider the implications of this reductionist orientation for social and affective robotics. As examined in [Section 6.2], the persistence of emotional PBR in robotics does not remain a merely theoretical concern, but has concrete consequences for the design, performance, and evaluation of robotic systems. By framing emotions primarily in terms of externally observable cues, the field has gained a practical operationalization of affect, yet at the cost of overlooking other dimensions that are still essential to emotional phenomena [Section 6.3].

Bringing these aspects to light will prepare the ground for [Chapter 7], where the role of propositionality in the design of affective robots and in the attribution of emotions to them will be examined as a possible broadening of the theoretical framework of emotions in robotics and HRI beyond its current reductionist orientation.

6.1 The Physiological-Behavioral Reductionism

Emerging from the combined influence of explicit design choices, theoretical biases, technological limitations, and practical constraints on real-world implementations, affective PBR has also come to characterize the field of social and affective robotics —and this from its very beginnings.

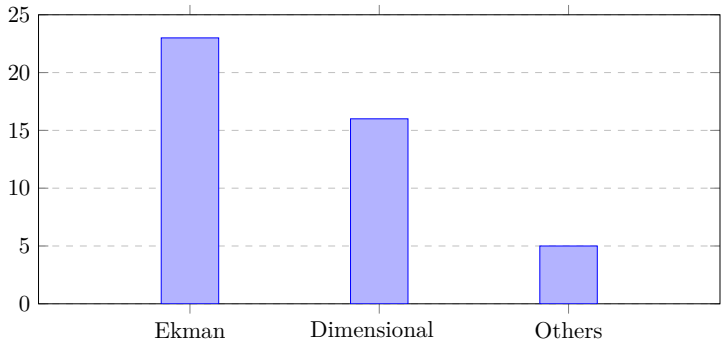
Already in Breazeal's Kismet [Section 5.3], indeed, emotions were framed mostly in behavioral terms. On the one hand, external stimuli — such as facial expressions, vocalizations, gaze patterns and gestures from the human interlocutor— were mapped onto arousal-valence-stance configurations corresponding to specific emotions from a finite and discrete

[98] An extended version of this review is currently in preparation for *BMC Psychology*.

affective spectrum. On the other, in turn, the same behavioral cues were used by the robot to generate and express those emotions. The focus on producing clear and readily interpretable affective signals left little space for modeling the cognitive and evaluative content of emotions, narrowing them just to a set of directly observable behaviors.

From these early implementations, then, as in the case of AC, such reductionist tendencies have never ceased to shape the way emotions are modeled in social and affective robotics. Across subsequent stages of research in robotics, an implicit "equation" between recognized, simulated, or attributed emotions and external, observable behaviors has consistently influenced the thinking and design practices of both researchers and engineers.

Until at least the late 2010s, for example, the predominant conceptual approach in emotion recognition for social and affective robots has been derived from Ekman's theory of emotions and FACS [Section 4.3], with cognitive and appraisal-based emotion theories appearing only in very few studies [Cavallo et al. 2018] —see e.g., the virtual robot *Roboceptionist* [Kirby et al. 2010] (see also [Perez-Gaspar et al. 2016]), able to recognize emotions from verbal inputs typed by hotel customers; or [Banda et al. 2015] and [Jing et al. 2015], partly accounting for the cognitive dimension of affectivity beyond facial expressions and similar cues (see [Figure 6.1, a]).



(a)

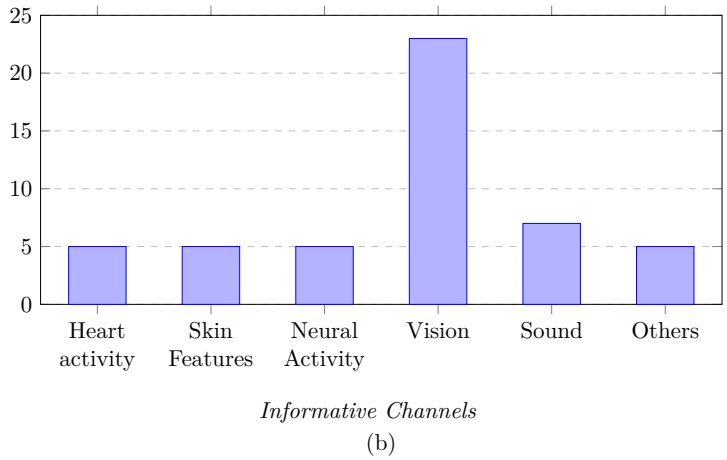


Figure 6.1: Distribution of *emotional models* (top) and *informative channels* (bottom) in emotion recognition for social and affective robotics, based on a review of 56 studies published between 2010 and 2017 (see [Cavallo et al. 2018]). Models based on Ekman's theory of emotion and FACS [Section 4.3] have formed the largest theoretical foundation in the field, with appraisal-based theories appearing in only a few studies (top chart). Additionally, vision has proven to be the most widely used informative channel for emotion recognition (bottom chart), emphasizing the dominant role of visual cues, particularly facial expressions. These findings underline the field's continued focus on observable and physiological cues for emotion recognition, with limited attention given to cognitive or evaluative dimensions of affectivity —adapted from [Cavallo et al. 2018, 195].

This trend has remained relatively stable until the present day. In contrast to Breazeal's Kismet, which did not yet incorporate physiological signals —beyond facial expressions, body language, and vocalizations— as relevant cues for emotional recognition and generation, latest developments have especially seen researchers and engineers increasingly including physiological signals and sensing modalities —electrocardiography (ECG) and photoplethysmography (PPG) for heart rate, electromyography (EMG) for muscle tension, electroencephalography (EEG) for brain activity, skin conductance sensors for electrodermal activity (EDA), infrared thermography (IRT) for autonomic activity— within the architectures of affective robots [Gervasi et al. 2023] (see [Figure 6.2]).

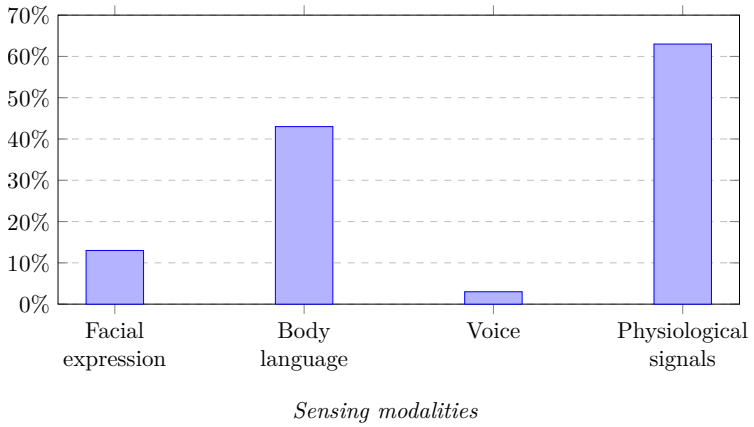


Figure 6.2: Sensing modalities employed for emotion recognition in recent HRI research, based on a review of 36 studies published between 2001 and 2022 [Gervasi et al. 2023]. The chart reports the relative frequency of four main modalities —facial expression, body language, voice, and physiological signals— showing that, within the reviewed studies, physiological sensing modalities emerge as the predominant approach for emotion recognition in affective robots, surpassing traditional behavioral modalities —adapted from [Gervasi et al. 2023, 825].

Even affective models developed for some of the currently most widely used social robotics platforms [Aymerich-Franch and Ferrer 2023] rely on a behavioral–physiological conception of emotion that is often implicitly presupposed by their designers. The robot *Pepper* [Figure 6.3, right], for example, instantiates the recent introduction of advanced AI techniques into social robotics (ANNs, k-NN, MLP, SVM, DL, BP, etc.). Despite this, however, the emotion recognition systems developed for this platform remain basically anchored to externally perceivable cues. This is the case of CNNs-based implementations for facial expression analysis [Corti and Cerrone 2020], affective states classification based on multimodal input analysis [Garcia et al. 2024], or DL-based visual models for advanced facial emotion recognition [Ahmed and Mishra 2024]. Still, another famous platform, the social robot NAO [Figure 6.3, left], has seen proposals for the extension of its affective capabilities through CNN-based modules for facial expression recognition and arousal detection [Filippini et al. 2021], or on vision-based systems linking real-time emotion recognition with head-motion imitation [Rayati et al. 2023].

Hence, across diverse application domains —from assistive and well-being-oriented systems [Spitale and Gunes 2022; Spitale et al. 2025]),

to infant–robot interaction [Filippini and Merla 2023]— emotion is today mostly operationalized through the joint mapping of behavioral and physiological correlates, reinforcing a reductionist conception in which observable parameters stand in for the full complexity of affective states, with cognitive and evaluative components —still central to emotions as far as known from human psychology and philosophy (see [Chapter 3] and [Section 7.2.2])— remaining largely neglected.



Figure 6.3: The social robots *NAO* (left) and *Pepper* (right). Developed by SoftBank Robotics Group Corporation in 2006 and 2014 respectively, both humanoids are currently among the most widely used platforms in assistive, educational, and HRI experimental research contexts —from the company website.

6.1.1 Emotional PBR in Robotics: Some Historical and Design Reasons

Among the possible reasons underlying this particular development there may be the need —crucial especially in the early stages of social and affective robotics— to design systems able to operate in real-world environments and to produce timely and unambiguous responses. Given the still limited processing capabilities and sensing technologies available

then, this constraint initially made it natural to privilege those affective models grounded in measurable, easily processable indicators. Such initial choices, while rooted in the contingencies of their time, may have nevertheless persisted in subsequent research efforts, carrying forward a preference for behavioral and physiological parameters even if the underlying technological limitations have progressively receded.

Another reason may be the (obvious and structural) emphasis of social and affective robotics on embodiment. The robot's physical presence —its morphology, movements, and sensory channels— constitutes both the medium through which it engages with human interlocutors and the primary vehicle for affective expression. In such a context, the body itself becomes the most immediate substrate for implementing emotional displays, naturally orienting design strategies toward the exploitation of externally perceivable behaviors and bodily signals, and further reinforcing a focus on the corporeal dimension of affect. This moreover resonates with the intrinsically embodied nature of emotions themselves. As previously examined [Section 3.2.1], affective states are profoundly linked to bodily changes. When transposed into robotics, this view tends to legitimize and reinforce the reliance on physical and physiological indicators as primary proxies for affect, often at the expense of less directly and externally observable dimension of emotion.

Often driven also by market and extra-academic pressures —aimed at producing emotional interactions with social robots that are immediately perceivable and readily assessable by observers, evaluators, and potential funders or customers— a form of affective reductionism in robotics has thus eventually crystallized. The high visibility and straightforward measurability of behavioral and physiological signals provide readily communicable proof of the robot's affective capability, entrenching their role as the default foundation for emotional modeling in the field —and this, arguably, even in a more pronounced fashion than in (non-embodied) AC [Section 4.3], where cognitivist perspectives maintain a somewhat greater representation due to the lesser constraint and shaping exerted by an intrinsic physical substrate.

6.2 Emotional PBR in Robotics: Practical Implications

The persistence of a reductionist orientation in the development of affective and social robots does not remain only a theoretical concern. Conversely, it has tangible implications for the purported effectiveness these systems are expected to demonstrate in real-world interactional contexts, especially as their deployment extends into an increasingly broad range of social applications.

In fact, robots are nowadays spreading in several domains of our society, from education and assistive care, to therapeutic interventions [Gonzalez-Aguirre et al. 2021; Aymerich-Franch and Ferrer 2023; International Federation of Robotics 2024]. In such contexts —as it might be readily acknowledged— the ability to sustain believable and nuanced emotional exchanges is a compulsory prerequisite for socially acceptable and functionally effective interactions. Therefore, detecting, interpreting, and responding to human emotions and intentions play a pivotal role for the robot with respect to the goal of enabling meaningful engagement and ensuring that the interaction aligns with human social norms and expectations (see e.g., [Yohanan and MacLean 2012]). In this sense, the integration of affect understanding has been central to the design of socially interactive robots, particularly in sensitive applications like those aimed at mental health or well-being (see e.g., [Kabacińska et al. 2021; Robinson et al. 2021; Axelsson et al. 2024]).

However, the very endeavor of building robots capable of recognizing, interpreting, processing, and simulating human emotions remains today far from being resolved. Adapting to the variability of human behavior or personalizing the interaction, in fact, involves technical challenges that are extremely difficult to address [Spitale and Gunes 2022] —not to mention the social and ethical considerations they rise (see e.g., [Boada et al. 2021; Hung et al. 2025]). But there are also deeper —and often overlooked— epistemological issues in present-day robotics that hinder the achievement of such ambitious goals.

Indeed, one of these exactly concerns the predominant presupposition of a reductionist behavioral–physiological conception of emotion [Section 6.1] as an implicit theoretical framework in scientific and en-

gineering research on affective and social robots. This presupposition namely creates a somewhat "paradoxical" condition. On the one hand, it enables the operationalization of emotions in artificial agents by framing them as directly observable, measurable, readily implementable, and easily recognizable parameters. Yet, on the other, as a side effect, it constrains the scope itself of what can be represented and processed as "emotion" in these systems. By reducing affect to a set of behavioral and physiological correlates, in fact, this reductionist paradigm tends to marginalize dimensions of affectivity that are central to our emotional life, ranging from the cognitive and evaluative aspects involved in emotional occurrences to the contextual, cultural, and societal factors shaping emotions themselves [Section 3.3]. As a result, the same framework intended to realize *deep* and *authentic* emotional interactions between humans and robots *simultaneously* limits the possibility to achieve such affective depth and authenticity[99]. This tension is further reinforced by the iterative nature of the technological development in the field: Design choices made under reductionist assumptions tend to propagate across generations of systems, embedding those assumptions ever more firmly into both research practices and technical architectures.

At the practical level, this reductionist paradigm gives rise to a cluster of recurring limitations that undermine the capacity of robots to sustain fluid, credible, contextually attuned, and ultimately *human-like* emotional exchanges with humans, especially in sensitive social domains like healthcare or education.

In what follows, a detailed overview of some (with each-other closely related) problems deriving from the adoption of PBR-based conceptual frameworks in robotics is presented. While distinct in their manifestation, these issues often operate in combination, shaping the overall interactional profile of current affective and social robots, mostly in limited and unauthentic ways[100].

[99] It may be questioned whether we truly need *deep* and *authentic* social and affective interactions with robots. This is a (fully legitimate) moral concern that lies however far beyond the more epistemologically oriented scope of the present work, and that will be therefore not accounted for.

[100] Of course, this list should not be taken as exhaustive, as further issues may be identified.

6.2.1 PB-Underdeterminacy

Problem 1 (PB-Underdeterminacy). *The same observable behaviors and physiological signals expressed by the robot can correspond to different emotions.*

The first problem is a long-standing one —at least from JLTE onwards— common to all conceptualizations of emotion as essentially behavioral or physiological states (see [Section 3.2.1.3], [Section 3.2.1.4] and [Section 3.3]): Identical outward behaviors and physiological responses can be associated with distinct emotional states. Same heart rate or skin conductance patterns hardly underlie discrete emotions, reflecting rather generic activation profiles that vary across contexts and individuals (see e.g., [Kreibig 2010; Shu et al. 2018]). Same postures, voice prosodies and gestures characterize qualitatively different emotional states (see e.g., [Barrett et al. 2019]); facial expressions are not *bijectively* associated to specific affects and are influenced by visual and situational factors, with body cues in high-intensity emotions often conveying valence more reliably than the face alone (see e.g., [Aviezer et al. 2011; 2012]).

If so, BP-patterns are not sufficient for validly inferring emotional states. Information about emotional states formally contained in emotional datasets and systems such as FACS' AUs [Footnote 47], for example, is not completely reliable when it comes to emotion recognition or generation in robots.

6.2.2 Interpretational Variability

Problem 2 (Interpretational Variability). *Different observers may infer different, and potentially conflicting, emotions from the same observable cues expressed by an affective robot.*

A second problem directly follows [Problem 1] as a corollary pertaining to *interpretational variability* of robotic emotions. If physiological and behavioral patterns are namely underdetermined with respect to a sin-

gle emotion, then they can be interpreted in multiple —and potentially conflicting— ways by different observers.

Such variability is conditioned, among other things, by cultural and situational factors, which shape both the production and the interpretation of emotional expressions [Section 3.3]. Different cultures, for example, emphasize distinct expressive cues as emotionally relevant, combining them in different ways, and regulating the contexts in which they are displayed or suppressed (see e.g., [Jack et al. 2012]). Even classical affective behavioral markers like the *Duchenne smile*[101] can vary with social distance, situational norms, and interaction goals, making them unreliable as unequivocal indicators of (in this case positive) emotional states (see e.g., [Bogodistov and Dost 2017; Krumhuber et al. 2014; 2023]).

Observers' interpretations of a robot's facial displays vary not only with the morphology of the robot's "face" and the technical rendering of its expressions, but also with the surrounding task and function of the robot, as well as the narrative context, and the broader cultural and social settings in which the interaction occurs. The *same* robotic expression, for example, may be read as friendly in a cooperative scenario or as competitive in a game, underscoring the degree to which context shapes both the recognition and the generation of emotions in social robots (see e.g., [Bennett and Šabanović 2015; Frederiksen et al. 2022; Ashok et al. 2025]).

Furthermore, personal background as well shapes emotion recognition and interpretation. Individual differences such as past experiences, personality traits, transient affective states (e.g., moods) or even individual physiological condition (see e.g., [Brown et al. 2017]) modulate how observers decode emotional cues. For instance, people with higher trait anxiety are more prone to perceive neutral or ambiguous expressions as threatening, while individuals with empathic dispositions are more in-

[101] Named after the French neurologist Guillaume-Benjamin-Amand Duchenne [1862], who first identified its distinctive muscular configuration, the *Duchenne smile* is a particular facial expression traditionally associated with genuine enjoyment. It is characterized by the simultaneous contraction of the zygomatic major muscle —which raises the corners of the mouth, and the orbicularis oculi muscle —which produces wrinkling around the eyes. Conversely, a *non-Duchenne smile*, in which the activation of the zygomatic major is not accompanied by the engagement of the orbicularis oculi, is often linked to posed or socially motivated displays rather than spontaneous affect.

clined to attribute positive emotions to the same display (see e.g., [Park et al. 2016]). Similarly, prior familiarity with robots, expectations about their role, or the degree of anthropomorphism assigned to them influences the way a robotic expression is read (see [Problem 5]).

6.2.3 Appropriateness and Misattribution

Problem 3 (Appropriateness and Misattribution). *The emotional state the robot is intended to express can be easily misinterpreted.*

As a further corollary to [Problem 1], closely related to [Problem 2], there is the issue of *emotional appropriateness and misattribution*. Especially within cognitivist approaches to emotion, emotional appropriateness is defined as the degree to which an experienced or expressed emotion fits the characteristics of the object, event or situation that elicited it, as well as the expectations that normally apply in the (social and cultural) context in which that elicitation occurs (see also [Section 7.2]). For example, fear is appropriate when one faces a real threat, while it would be not if it is displayed in a safe and familiar environment.

Since physiological and behavioral patterns are underdetermined with respect to emotions [Problem 1], the observable cues expressed by a robot may be easily misinterpreted, leading a human observer to attribute an emotion different from the one those very cues were programmed to convey in that context (see e.g., [Fiorini et al. 2024]). In such cases, in other words, observers may ultimately end up ascribing to the robot an affective response that is therefore contextually *inappropriate*.

6.2.4 Complex Emotional States

Problem 4 (Complex Emotional States). *BP-based affective robots struggle with composite, reflexive, or meta-emotional states.*

Another problem concerns the accurateness and scope of the implementation of robotic systems capable of recognizing and generating *complex emotions*.

"Complex emotions" is a label that denotes a broad and heterogeneous set of affective states. In the literature, multiple definitions have been proposed, alongside several accounts attempting to clarify their nature, internal categorization, and the conditions under which they arise (see e.g., [Lindquist et al. 2006; Hay and Diehl 2011; Grossmann et al. 2016; Berrios 2019]). Perhaps the most straightforward way to appreciate what complex emotions are is by contrasting them with the naïve, folk-psychological understanding of *basic emotions* (joy, sadness, fear, etc.) [Section 3.3]. Whereas the latter are often described as biologically hard-wired, discrete, and readily recognizable reactions to specific stimuli, the former typically manifest as nuanced, context-dependent states in which cognitive, social, and cultural influences converge within a dynamic interplay of affective and evaluative processes.

Commonly, complex emotions are taken to involve different kinds of affective states, each characterized by distinct patterns of experience, expression, and underlying psychological or neurocognitive processes[102]:

(a) *Mixed Emotions.* Mixed emotions —also variously discussed as dialectical or ambivalent emotions— are affective states in which one finds herself feeling emotions with opposite valence (positive and negative) simultaneously or in such close succession that they are phenomenologically apprehended as part of a single episode (see e.g., [Larsen and McGraw 2014]). In contrast to simple alternation between discrete emotions, here affective components of opposite

[102]This list is not intended to establish a definitive sub-classification of complex emotions, i.e., to provide necessary and sufficient conditions for an emotion to qualify as a particular instance of one category on the list. As noted, this issue remains an open debate in philosophy and psychology, and I do not take a position here on it. My aim is rather to offer a simple overview of some of the most commonly discussed complex affective states —that is, states whose occurrence and structure extend beyond an immediate, automatic affective response to a clearly identifiable stimulus. The discussion is pursued here namely for its relevance to the treatment of complex emotions in robotics (see [Section 7.3.2.4]), rather than for the purpose of engaging in detailed theoretical disputes about their logical or ontological status. For reasons of conciseness, therefore, this overview also includes cases such as recalcitrant emotions —which are seldom addressed in debates specifically focused on complex emotions. Furthermore, it acknowledges —here in a purely *descriptive* rather than philosophically critical sense— that a single complex emotion may at times be plausibly assigned to more than one category —for example, guilt can be considered a social emotion [Problem 4(c)] and, in certain contexts, a meta-emotion [Problem 4(e)] or an emotion of emotion [Problem 4(b)].

valence *coexist* within *one* coherent experiential frame. Mixed emotions can derive from the combination of both "basic" (feeling joy and sadness together upon, say, a graduation ceremony) and complex emotions themselves (feeling pride and guilt in achieving a goal through questionable means).

(b) *Emotions of Emotions* (*Second-Order Emotions*). Emotions of emotions —also referred to as second-order emotions— are affective states whose intentional object is not an external event or situation, but rather another emotion already being experienced by the subject (see e.g., [Mendonça 2013]). In these cases, the primary emotion (possibly basic or complex) becomes itself the target of a further evaluative or affective response, giving rise to a layered emotional episode. Examples of emotions of emotions are feeling joy upon one's own sadness, or shame about one's own anger.

(c) *Social Emotions*. Social emotions are affective states whose elicitation experience or expression depends on other (human) beings —i.e., in their directly lived, remembered, expected, or just envisioned actions, thoughts and feelings. They are shaped by social norms, interpersonal relationships or group dynamics, and often presuppose complex cognitive capacities such as perspective-taking, self-evaluation, or the attribution of mental states [Footnote 94]. Examples of social emotions are empathy, shame (say, in response to public criticism), guilt (say, for having harmed someone), schadenfreude (joy for another's misfortune), envy, pride, remorse, embarrassment, jealousy, love, etc. (see e.g., [Parkinson et al. 2005; Hareli and Parkinson 2008]).

(d) *Recalcitrant Emotions*. As a true paradox in the philosophical and psychological reflection on emotions, often invoked as an evidence for the untenability of cognitivist theories of affect (see e.g., [D'Arms and Jacobson 2003; Brady 2009; Benbaji 2013; Sacco 2024]), recalcitrant emotions are particular affective states that one continues to experience despite judging them to be unfounded or inappropriate in light of her own beliefs. In these states, there is a conflict between what the emotion presents as evaluatively signif-

icant and what the subject, upon reflection, takes to be the case —a conflict that bears interesting epistemological connections with phenomena like *self-deception* (see e.g., [Deweese-Boyd 2023]) and *delusion* (see [Bortolotti 2022] and [Section 2.2.4]). Examples include feeling fear during an airplane flight despite believing it to be perfectly safe, or feeling envy toward a friend's achievement while sincerely judging it to be well-deserved.

(e) *Meta-Emotions.* Meta-emotions are affective states whose intentional object consists in an emotional state experienced by the subject herself or by others (see e.g., [Gottman et al. 2013]). Unlike emotions of emotions [Problem 4(b)] —which can arise spontaneously in response to another emotion— meta-emotions typically involve a degree of self-awareness and cognitive appraisal of the appropriateness, intensity, or moral status of the primary emotion, they can reinforce, modulate, or conflict with. Typical of parental relationships, meta-emotional instances include, for example, feeling pride in a child's excitement about a new experience, feeling frustration at a someones's anger during a disagreement, or feeling joy for a future feeling of joy (what is known as "*Vorfreude*" in German, or "*prìsce*" in Barese dialect[103]).

(f) *Dynamic Emotions.* Dynamic emotions are affective states whose character and trajectory unfold over time, often through multiple phases that may differ in valence, intensity, or qualitative tone. Rather than being static, they are shaped by ongoing appraisal processes and by the interaction between internal affective dynamics and external situational changes. This temporal structure can include escalation, attenuation, transformation into other emotions, or recurrent cycles —in Plutchik's wheel, in other words, this would correspond to moving along the gradients between emotions or shifting across adjacent categories as the episode develops. Examples include the gradual shift from irritation to amusement during

[103] As far as I know, German and Barese are curiously the two only European languages to have a single-term word for this specific complex emotion —albeit with some nuanced differences in the way they are used. This confirms as even in relatively near social communities, emotions partly derive from cultural constructs [Section 3.3].

a challenging conversation, or the intensification of fear into panic as a threatening situation develops.

(g) *Machiavellian Emotions.* As a particular case of social emotions [Problem 4(c)], machiavellian emotions —named after the Italian Renaissance thinker Niccolò Machiavelli, known for his reflections on power and strategy— are affective states whose production and/or expression are finely attuned to the social and strategic context of the agent [Griffiths 2003]. They arise from an assessment of the potential social advantage in experiencing or displaying them: In some cases, the emotion is genuinely felt but its expression is strategically modulated or suppressed; in others, it is deliberately produced or simulated to influence another's affective state or to induce deceptive beliefs (see e.g., [Mahon 2016]). Examples include feigning sympathy to gain trust, amplifying indignation to deter aggression, suppressing joy to avoid provoking envy, exaggerating gratitude to secure future favors, simulating remorse to mitigate punishment, etc.

(h) *Reflexive Emotions.* Related to emotions of emotions [Problem 4(b)] and meta-emotions [Problem 4(e)], reflexive emotions are affective states that take the self both as subject and as object of the emotional stance [Fisch 2025]. They include, for example, emotions such as shame or pride.

(i) *Aesthetic Emotions.* Aesthetic emotions emerge as affective responses to objects, events, or experiences —e.g., artistic creations, natural landscapes, performances, etc.— appreciated primarily in their aesthetic dimension (see e.g., [Schindler et al. 2017a; Menninghaus et al. 2019]). They typically involve a complex interplay of perceptual, cognitive, and evaluative processes, and, unlike basic emotions, they do not have an adaptive significance, being rather elicited by symbolic and cultural expressive forms. Examples include being moved by a musical composition (see especially [Section 8.1]), feeling awe before a sublime landscape, experiencing poignancy while watching a tragic play, etc.

(j) *Epistemic Emotions.* Epistemic emotions are affective states whose intentional object concerns epistemic ends, such as truth, understanding, or explanatory adequacy (see e.g., [Morton 2009; Arango-Muñoz and Michaelian 2014; Elgin 2016]). Unlike enduring epistemic virtues or traits of character, they are occurrent states that arise in specific moments of inquiry and exert a motivational force on the pursuit of knowledge. They can sustain, redirect, or obstruct epistemic activity depending on their valence and intensity. Typical examples include curiosity, wonder, intellectual fascination, or disappointment when an expected explanation fails; but also the relief and satisfaction of resolving a puzzle, or the unease of encountering apparent incoherence. By acting as affective motors of cognition, epistemic emotions bridge the gap between mere instrumental reasoning and the lived experience of inquiry, shaping how agents engage with questions, assess alternatives, and commit to epistemic goals (see also [Craik 1943, 86]).

(k) *Emotional Granularity.* Emotional granularity denotes the complex affective experience in which one is able to differentiate emotions with fine precision —i.e., say, to distinguish even very close values-points in a PAD space [Section 3.3]. This capacity allows the subject to register as discrete and qualitatively distinct affective states that would otherwise appear identical (see e.g., [Smidt and Suvak 2015]). Examples include distinguishing irritation from frustration in conflict, parsing shades of sadness into melancholy, nostalgia, or grief, etc.

(l) *Emodiversity.* Emodiversity denotes the breadth and balance of the emotional life of a subject, understood as the extent to which a variety of distinct affective states are experienced by her within a given period of time. It can be even formally quantified through an adaptation of the Shannon index, defined as

$$\text{Emodiversity} = -\sum_{i=1}^{s} p_i \cdot \ln(p_i),$$

where s is the number of distinct emotions and p_i the relative frequency of each [Quoidbach et al. 2014]. High emodiversity is characterized by multiple, well-balanced emotions, whereas low emodiversity corresponds to a state of affective monotony or redundancy.

Such complex, higher-order affective phenomena hardly lend themselves to be accounted for *exclusively* in physiological and behavioral terms. Their defining features —layered appraisal, social embedding, and reflexive dynamics— extend beyond the bodily signals or expressive cues that PB-based models are mostly able to represent. Also for this reason, robotics has so far devoted little attention to them, concentrating instead on emotions that can be straightforwardly operationalized in physiological or behavioral terms.

Some initial attempts to model complex affective states for robots do indeed exist —for example, recent work about robotic social emotions [Problem 4(c)] (see e.g., [Hieida and Nagai 2022]). Nevertheless, the research in the field is still embryonic, conceptually narrow, and far from providing a comprehensive treatment of the full spectrum of affective phenomena that fall within the domain of complex emotions. In other words, the possibility for a robot to experience and express, say, recalcitrant emotions [Problem 4(d)] or epistemic emotions [Problem 4(j)] remains largely unexplored. Considering such possibilities would render its affectivity more human-like and, consequently, its interaction with humans more nuanced and credible, a line of inquiry that clearly requires moving beyond PBR.

6.2.5 Anthropomorphism and Uncanny

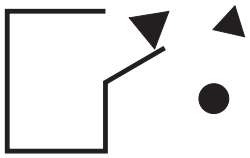
Problem 5 (Anthropomorphism and Uncanny). *Human-like robots whose affectivity is reduced to shallow PB-cues risk generating a gap between appearance and emotional depth, often eliciting uncanny reactions in the observer.*

A fourth problem arises when PBR-based emotional models are embedded into *humanoid robotic platforms*. Reducing affectivity mostly to

surface-level physiological and behavioral patterns may suffice in simpler mechanical or zoomorphic robots. As noted in the case of cybernetic creatures [Section 5.1.2], for example, even minimal architectures can generate affectively convincing behaviors —something true already for simple moving geometric shapes, as demonstrated by Heider and Simmel [1944][104]. In the case of humanoid robots, however, PBR-based emotional models begin to show severe limitations. Interacting with a robot with (at least purported or partial) human-like appearance, in fact, naturally invites observers to expect corresponding human-like emotional depth from it (beside human-like cognitive capabilities, communication, behaviors, etc.). Yet, since emotions are here reduced to physiology and behavior, these robots can only reproduce "shallow" affective displays, without the evaluative and contextual grounding that makes human emotions intelligible and credible.

This discrepancy between form and affective substance often results, for the interacting human, in awkward or unsettling responses, undermining the intended goal of fostering trust and natural interactions. In literature, this particular affective reaction has been referred to with a particular name, the *Uncanny Valley Effect* (see [Footnote 80]). As an object of psychological inquiry, the *uncanny* has a long history (see e.g., [Schwind 2015]). Drawing on E.T.A. Hoffmann's [1817] tale *Der Sandmann* (1817)[105], already the German psychiatrist Ernst

[104] In their classic experiment, Heider and Simmel [1944, 244] presented participants with an animation of two triangles and a circle moving around a rectangle. Despite the extreme simplicity of the shapes, most viewers spontaneously described the scene in rich social and emotional terms (e.g., attributing intentions such as chasing, hiding, or bullying). This showed how even minimal perceptual cues can trigger complex emotional attributions



[105] *Der Sandmann* by E.T.A. Hoffmann [1817] tells the story of Nathanael, a young man haunted since childhood by the terrifying figure of Coppelius, the "Sandman". He later falls in love with Olimpia, who is eventually revealed to be a mechanical automaton created by Spallanzani with the help of Coppola. The shock of this discovery proves unbearable for Nathanael, and the trauma ultimately drives him to his death. The tale exemplifies the unsettling ambiguity between the animate

Jentsch [1906], for example, described it (the "*Unheimlich*") as a feeling of unfamiliarity, estrangement, disorientation and displacement[106] that is caused by the *psychic uncertainty* generated from an event that is new, unusual and therefore difficult to intellectually assimilate on the basis of previous knowledge and experience[107]. Discussing the typical scenarios in which this feeling arises, he interestingly mentioned the confrontation with objects (e.g., wax figures and, namely, *humanoid automata*) that elicit the doubt as to whether an apparently living being is truly animate[108].

and the inanimate, a central motif in Jentsch's analysis of the *Unheimlich* as arising from uncertainty over whether something apparently alive truly possesses life.

[106] «Es scheint dadurch wohl zweifellos ausgedrückt werden zu sollen, dass einer dem etwas „unheimlich" vorkommt, in der betreffenden Angelegenheit nicht recht „zu Hause" nicht „heimisch" ist, dass ihm die Sache fremd ist oder wenigstens so erscheint, kurzum, das Wort will nahe legen, dass mit dem Eindruck der Unheimlichkeit eines Dinges oder Vorkommnisses ein M a n g e l a n O r i e n t i e r u n g verknüpft ist» [Jentsch 1906, 195], (En. tr.) «Without a doubt, this word appears to express that someone to whom something "uncanny" happens is not quite "at home" or "at ease" in the situation concerned, that the thing is or at least seems to be foreign to him. In brief, the word suggests that a *lack of orientation*» [Jentsch 1997, 8].

[107] «Prüft man die Psychologie des Alltags in dieser Richtung näher, so ersieht man unschwer, [e]s ist eine alte Erfahrung, dass den meisten Menschen das Althergebrachte, Gewohnte, Angestammte lieb und vertraut ist, und dass sie das Neue, Aussergewöhnliche mit Misstrauen, Missbehagen, selbst Feindseligkeit aufnehmen (Misoneismus). Dies erklärt sich zum grossen Theile aus der Schwierigkeit, die Ideenverbindungen, die das Object zu dem bisherigen Vorstellungsbereich des Individuums anstrebt, also die Herrschaft des Intellects über das neue Ding, rasch und vollständig herzustellen. Das Gehirn scheut es oft, die Widerstände zu überwältigen, die sich dem Hineinassociiren der fraglichen Erscheinung auf den ihr gebührenden Platz entgegenstellen. [...] Das Altgewohnte erscheint nun nicht nur als willkommen, sondern, mag es noch so wunderbar und unerklärlich sein, auch leicht als selbstverständlich» [Jentsch 1906, 195-196], (En. tr.) «If one takes a closer look at everyday psychology in this sense, it can easily be seen that [... i]t is an old experience that the traditional, the usual and the hereditary is dear and familiar to most people, and that they incorporate the new and the unusual with mistrust, unease and even hostility (misoneism). This can be explained to a great extent by the difficulty of establishing quickly and completely the conceptual connections that the object strives to make with the previous ideational sphere of the individual –in other words, the intellectual mastery of the new thing. The brain is often reluctant to overcome the resistances that oppose the assimilation of the phenomenon in question into its proper place. [...] That which has long been familiar appears not only as welcome, but also —however remarkable and inexplicable it may be— as straightforwardly self-evident» [Jentsch 1997, 8-9].

[108] «[D]er Zweifel an der Beseelung eines anscheinend lebendigen Wesens und umgekehrt darüber, ob ein lebloser Gegenstand nicht etwa beseelt sei», (En. tr.) The «doubt as to whether an apparently living being is animate and, conversely,

Later, Freud [1919] elaborated on this view. Highlighting other aspects of the phenomenon from a psychoanalytic perspective, he proposed that the *uncanny* is rather an affective response to familiar experiences that, once relegated to the subconscious, are at some point recalled, often returning in distorted, repressed, or estranging forms[109].

Building on these earlier psychological accounts, the notion of uncanny was later reformulated in a more applied framework by the Japanese roboticist Masahiro Mori [1970]. Shifting the focus from literature and psychoanalysis to technology and design, Mori proposed the now famous *Uncanny Valley Hypothesis* (UVH)[110]. According to this hypothesis [Fig-

doubt as to whether a lifeless object may not in fact be animate» [Jentsch 1997, 11].

[109] «[D]as Unheimliche sei jene Art des Schreckhaften, welche auf das Altbekannte, Längstvertraute zurückgeht» Freud [1919, 298], (Eng. tr.) «[T]he uncanny is that species of the frightening that goes back to what was once well known and had long been familiar» [Freud 2003, 124]; «Unheimlich sei alles, was ein Geheimnis, im Verborgenen bleiben sollte und hervorgetreten ist» [Freud 1919, 302], (Eng. tr.) «"[U]ncanny" (*unheimlich*) applies to everything that was intended to remain secret, hidden away, and has come into the open» [Freud 2003, 132]; «Eine „intellektuelle Unsicherheit" kommt hier nicht mehr in Frage: Wir wissen jetzt, daß uns nicht die Phantasiegebilde eines Wahnsinnigen vorgeführt werden sollen, hinter denen wir in rationalistischer Überlegenheit den nüchternen Sachverhalt erkennen mögen, und —der Eindruck des Unheimlichen hat sich durch diese Aufklärung nicht im mindesten verringert. Eine intellektuelle Unsicherheit leistet uns also nichts für das Verständnis dieser unheimlichen Wirkung» Freud [1919, 306], (Eng. tr.) «There is no longer any question of "intellectual uncertainty": We know now that what we are presented with are not figments of a madman's imagination, behind which we, with our superior rationality, can recognize the sober truth — yet this clear knowledge in no way diminishes the impression of the uncanny. The notion of intellectual uncertainty in no way helps us to understand this uncanny effect» [Freud 2003, 139]; «[D]ies Ängstliche [ist] etwas wiederkehrendes Verdrängtes [...]. Diese Art des Ängstlichen wäre eben das Unheimliche [...], denn dies Unheimliche ist wirklich nichts Neues oder Fremdes, sondern etwas dem Seelenleben von alters her Vertrautes, das ihm nur durch den Prozeß der Verdrängung entfremdet worden ist» [Freud 1919, 314], (Eng. tr.) «[T]he frightening element is something that has been repressed and now returns. This species of the frightening would then constitute the uncanny [...], for this uncanny element is actually nothing new or strange, but something that was long familiar to the psyche and was estranged from it only through being repressed» [Freud 2003, 147-148].

[110] Indeed, Mori [1970] published his UVH in Japanese in the relatively obscure journal *Energy*. This limited its initial circulation, preventing the idea from gaining international resonance at the time. In the west, the English expression "*Uncanny Valley*", as a translation of 不気味の谷 (*bukimi no tani*), was introduced during the late 1970s by the journalist Jasia Reichardt [1978] in a famous book about the cultural history of robots. Though not an academic study in the proper sense, this work played a key role in popularizing Mori's idea even among scientists. An official English translation of Mori's original paper however appeared only during

ure 6.4], the relation between the human-likeness (i.e., the degree of anthropomorphism) of an agent and the affective response it elicits in an

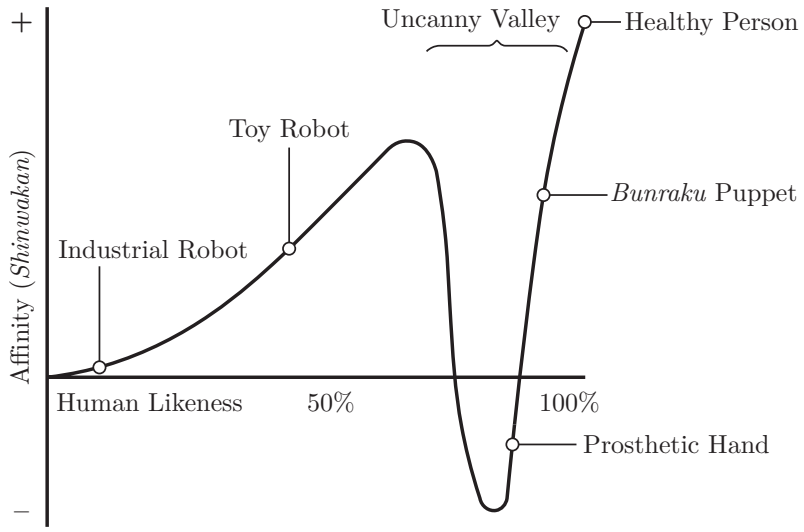


Figure 6.4: Graphical representation of Mori's [1970] *Uncanny Valley Hypothesis* (UVH). As human-likeness increases, affective affinity rises until a critical threshold, where too high anthropomorphism generates a sharp drop in positive response (namely, the "valley"). The examples along the curve illustrate this trajectory: An industrial robot (low human-likeness) evokes neutral reactions, while a toy robot (medium human-likeness) elicits greater affinity. By contrast, figures that are highly anthropomorphic but not yet fully human —such as the bunraku puppet, a traditional Japanese wooden marionette, or a prosthetic hand — fall into the valley. At the far end, a healthy human restores positive affective engagement —from [MacDorman and Kageki 2012, 99].

interacting observer can be defined in mathematical-quantitative terms as *nonlinear*. As the agent becomes more anthropomorphic, affinity and positive affective engagement increase. At a critical threshold, however, when the agent is *too* human-like, the affective response suddenly turns negative, producing the characteristic "valley" observable in the graph. Beyond this dip, then, as the agent increasingly approximates a real,

the 2010s thanks to MacDorman and Kageki [2012], whose editorial and scientific contributions on UVH (see also [MacDorman 2005a; MacDorman and Ishiguro 2006]) have been crucial in the process that has brought the hypothesis to become one of the most controversial topics in the whole contemporary HRI research (see e.g., [Wang et al. 2015; Palomäki et al. 2018; Zhang et al. 2020b]).

healthy person to the point of becoming indistinguishable from her, the the valence of the observer's affective reactions turns once again positive.

Despite its conceptual influence and its enduring relevance in design and robotics, UVH remains controversial from an empirical perspective (see e.g., [Kätsyri et al. 2015; Diel et al. 2021; Di Natale et al. 2023]). Several experimental studies have sought to confirm the existence of the curve, yet results remain inconsistent, and the hypothesis has never been conclusively validated: Some studies, for example, have found the nonlinear relation between human-likeness and affective reaction as originally predicted by Mori [1970] (see e.g., [Thepsoonthorn et al. 2021]); other studies, however, have found a linear relation (see e.g., [Diel and Lewis 2024]), subject- and culture-dependent quantitative relations of various kinds (see e.g., [Rosenthal-von der Pütten and Krämer 2014]), or even no relation at all between human-likeness and emotional response (see e.g., [Kätsyri et al. 2019]).

Other open questions concern the *qualitative-phenomenological nature* of the uncanny as an affective state (see e.g., [Schwind 2015; Benn 2025]), its *neurobiology* (see e.g., [Schindler et al. 2017b; Fitzpatrick et al. 2025]), and its *causes*. Possible explanations for the emergence of the phenomenon, for example, involve *categorical ambiguity*, i.e., the difficulty in assigning an entity to a familiar ontological class such as "human" or "non-human" (see e.g., [Cheetham et al. 2011]); *evolutionary accounts*, which trace the uncanny to pathogen-avoidance mechanisms or mortality salience responses (see e.g., [MacDorman 2005b; MacDorman and Entezari 2015]); and *perceptual mismatch theories*, according to which negative affect arises when features of an agent display inconsistent degrees of human-likeness (for instance, when highly realistic facial morphology coexists with unnatural motion or expression) —see e.g., [MacDorman and Chattopadhyay 2016]. A further and closely related explanation is the *expectation-violation hypothesis*: Given high anthropomorphism, the observer comes to expect fully human-like appearance, motion, and behavior. Even slight deviations from what would normally occur in an interaction with a real person —such as subtle irregularities in form, movement, or expressivity— can thereby elicit unease and estrangement instead of empathy and positive feelings (see e.g., [Saygin et al. 2012; Urgen et al. 2018]).

Finally, as a relatively less thematized (philosophical) aspect in current UVH research, there also arises the question of anthropomorphism itself as a parameter of the model (see e.g., [Cheetham et al. 2011; MacDorman and Ishiguro 2006]); whether anthropomorphism is restricted to physical and morphological traits or also extends to behavioral, cognitive, and emotional dimensions (see e.g., [Crowell et al. 2019; Kühne and Peter 2023]); whether and how it can be adequately quantified and measured (see e.g., [Von Zitzewitz et al. 2013; Bartneck et al. 2017; Torricelli et al. 2020] and in particular the *Godspeed Questionnaire*, used in current research for quantitatively assessing human-likeness [Bartneck 2023]); and whether it is to be regarded as an objective property intrinsic to an artifact or rather as a (cultural) construct dependent on the perceiving subject (see e.g., [MacDorman and Entezari 2015; Jastrzab et al. 2024; Roselli et al. 2025]). Connected to this, there is also the problem of agent–human familiarity, namely the possibility of operationalizing and quantifying it, and of understanding its influence on anthropomorphism and, consequently, on the emergence of uncanny responses[111] (see e.g., [MacDorman 2006; Kamide et al. 2014; Diel and Lewis 2022]).

Now, even if Mori's [1970] UVH quantitative model has not received consistent empirical confirmation, the phenomenon itself continues to be widely attested by anecdotal evidence (see e.g., [Tinwell 2014]), being also largely acknowledged in robot design (see e.g., [Faraj et al. 2021; Rojas-Quintero and Rodríguez-Liñán 2021; Yan et al. 2024]). For the purposes of the present discussion, therefore, the broader concern it raises with respect to the problem of affectivity remains central: Humanoid robots that display a high degree of anthropomorphism, but possess mostly PB-based affective competence, risk eliciting uncanny reactions in users.

[111] For example, in a recent study [Zanzarella 2025a], I showed how changes in the historical ways of perceiving automata directly affect familiarity, thereby modifying the perceived degree of human-likeness and consequently altering the conditions for the emergence of the uncanny. Historical sources on eighteenth-century humanoid musical automata indicate, in fact, that the public often experienced positive affective responses and attributed human-like qualities even to machines with relatively coarse features, precisely because their low familiarity reduced the threshold for anthropomorphic attribution. Such evidence suggests that Mori's model cannot be taken as universally valid, since it is not invariant with respect to historical and cultural coordinates. Following this line, I propose that the UVH model should be reformulated with three dimensions, introducing familiarity as a further parameter alongside anthropomorphism and emotional valence.

The mismatch between a human-like appearance and an affective competence limited to surface-level expressive patterns generates precisely the estrangement and unease described in the literature on the uncanny. From a design perspective, this implies that increasing the accuracy of PB expressive patterns up to a human-like level does not straightforwardly secure trust, empathy, or social acceptance. On the contrary, it may result in the opposite effect, producing alienation and discomfort whenever the expressive display fails to correspond to the depth of affective grounding that users spontaneously expect from human-like agents. The simulation of human emotions in artificial agents can amplify uncanny responses instead of mitigating them —which is also shown by several experimental findings in robot psychology and HRI (see e.g., [Mäkäräinen et al. 2014; Reuten et al. 2018]).

6.3 Beyond Reductionism

As seen in previous [Section 6.2], the persistence of a reductionist conception of emotion in social and affective robotics entails a series of theoretical and practical shortcomings that prevent the field from fulfilling its central aim —that is, the development of affective machines capable of interacting with humans in a context-sensitive, timely responsive, durable, and genuinely social manner.

Conventional methods for emotion recognition and generation based solely on vision, bodily cues, facial expressions, and physiological signals lack the accuracy and reliability needed to produce credible emotions and to ensure their consistent recognition. They namely depend heavily on age, gender, culture, social norms, pathological conditions, etc. [Filippini and Merla 2023, 399]. A PB-based robotic affectivity moreover tends to minimize the role of *context* (see e.g., [Lewis and Cãnamero 2013]), *background knowledge* and *memory* in shaping emotional states, both in themselves and in their social and interactional significance [Cavallo et al. 2018, 191]. Nor does it adequately allow for building and sustaining *long-term relationships* between humans and robots (see e.g., [Kirby et al. 2010]), which is a prerequisite for systems that are meant to take part in human social life —especially in sensitive social domains— and to support forms of affective interaction that go beyond episodic exchanges.

For these reasons, there emerges a pressing need to *broaden* the theoretical framework of emotion in social and affective robotics beyond affective PBR. A new conception of robotic affect is required that —*alongside* vision, bodily cues, facial expressions, and related signals— integrates mechanisms enabling robots to display emotions in ways that are both more credible and more recognizably "human-like". Such a framework should help avoid ambiguous emotional expressions, sustain the representation of complex affective states, and support the generation of emotions that are appropriate to the context. It should likewise enable them to more accurately recognize human affective states, thereby overcoming the inherent underdeterminacy of purely physiological and behavioral cues. Also, this new conception should allow technical implementations that facilitate human partners in discerning the robots' emotional states with greater clarity and reduced noise, allowing them to *attribute* correct emotions more readily within ongoing interactions, while avoiding uncanny effects.

In other words, what is needed is a new theoretical framework of emotion in social and affective robotics that acknowledges the richness of human emotional life as a complex, dynamic, and multifaceted phenomenon, and that aligns more closely with the ways emotions are actually experienced and expressed by humans —as highlighted through centuries of inquiry in philosophy, psychology, medicine, and related disciplines. In particular, a new conception of robotic emotion should be advanced that accounts for the representational and evaluative dimensions involved in affectivity, its sensitivity to context, as well as its anchoring in memory and background knowledge. Only such a conception of emotion can underpin the development of more authentic and genuinely "human-like" robotic affective capacities, thereby sustaining richer forms of human-robot interaction within nuanced social contexts.

One possible way of developing such a new broader theoretical framework is to consider the function of propositionality with respect to robotic emotions and the attribution of emotions to robots, a possibility I shall explore in the next [Chapter 7].

Summary

This chapter has examined the current state of robotic emotion research, focusing on the persistence of a physiological-behavioral reductionism in both theoretical frameworks and implementational practices of the field. Through the analysis of its historical and design roots, it has been shown how the reliance on observable bodily and expressive cues, inherited from earlier AC models, has shaped the development of robotic architectures privileging outward expressivity over internal evaluative or representational depth. While these approaches have contributed to the realism and social effectiveness of robotic interaction, they have also revealed critical limitations —ranging from the underdeterminacy of bodily patterns and the variability of interpretation to the frequent misattribution of complex emotional states and the recurrent emergence of uncanny feelings.

The discussion of these issues has emphasized the need for a conceptual shift beyond mere simulation or recognition of behaviors and physiological signals, toward frameworks capable of grounding emotions in intentional, semantic, and context-sensitive dimensions. The next [Chapter 7] takes up this challenge by exploring the role of propositions in robotic emotions and emotional attributions to robots, with the aim of accounting for their formal structure, cognitive implications, and relational meaning beyond the physiological-behavioral paradigm.

Chapter 7

Robotic Emotions Beyond PBR: A Propositionalist Approach

The previous [Chapter 6] has shown how, in contemporary affective robotics, emotions are still predominantly conceived within a PBR-based framework. While this paradigm has provided a solid technical ground for the design and evaluation of robotic systems, it tends to overlook some of the essential aspects of emotional life —namely, its cognitive and intentional dimensions.

Building on the critique developed there, the present chapter aims to move beyond the limits of this reductionism by introducing a propositionalist account of robotic emotions. Here, emotions are approached not merely as patterns of behavior or physiological activation, but as states that can be attributed, interpreted, and socially constituted within human-robot interaction. This shift allows to rethink the epistemology of robotic emotions, situating them within a broader framework that connects their design and attribution with propositional content.

In [Section 7.1], I examine the social constitution of robotic emotions, showing how emotional attributions arise from the interplay between the robot's design features and the interpretative responses of human interactants. This section clarifies the interdependence between design and attribution in the emergence of emotional meaning and discusses its implications for human-robot interaction.[Section 7.2] develops the theoret-

ical foundations of the propositionalist approach, outlining how emotions can be understood as intentional states with propositional content. It reconstructs the main notions of attribution, intentionality, and evaluative representation, offering a pluralist conceptual model for integrating these dimensions into the study of robotic emotions. Finally, [Section 7.3] rethinks the epistemology of robotic emotions in light of the propositionalist framework. Here, I show how this account can address the issues discussed in [Section 6.2] by grounding emotional design on propositional and evaluative structures beyond physiological or behavioral correlates.

7.1 The Social Constitution of Robotic Emotions

In [Chapter 6], I highlighted the fact that most approaches to emotions in present-day affective and social robotics rely on a form of *physiological-behavioral reductionism* (PBR), as I called it. This model conceives of emotions essentially as sets of external and observable patterns — physiological arousals, bodily postures, vocal modulations, stereotypical facial expressions, etc.— that can be isolated, codified, and then engineered into a robot. To make a robot express fear, for example, a collection of fear-related PB-cues is typically selected and then implemented mechanically and computationally in it, as the external output of internal computational processes for emotion generation, which may vary in typology and degree of complexity. This, it is assumed, is sufficient for the robot to simulate that emotion. Yet, such an assumption, as I have argued, is problematic in several respects, since it reduces the richness of emotions as ordinarily experienced by humans to a narrow repertoire of signals and behaviors, overlooking dimensions that are central to our understanding of emotional life especially within social contexts.

7.1.1 From Design to Attribution

Indeed, in contexts like these, emotions do not emerge only from internal generative mechanisms and external expressive cues —which applies to both artificial agents and humans alike. Let $\mathcal{Q}$ be a particular *social context*. As an event b occurs, a series of internal mechanisms and processes

$I = \{i_1, i_2, \ldots, i_n\}$ (physiological changes, cognitive appraisal, etc.) lead a subject $s_i \in S$ in $\mathcal{Q}$ to the *generation* of a certain emotion $e_i \in E$ in response to b. These mechanisms and processes can be realized in different ways according to the nature of s_i: physiologically and neurally in a biological organism, or computationally and electro-mechanically in an artificial agent. Then, a set of characteristics $C = \{c_1, c_2, \ldots, c_n\}$ (physiological signals, behaviors, mental states, phenomenological conditions, etc.) instantiated by s_i function as the *expression* of e_i.

Now, neither $I(s_i)$ nor $C(s_i)$ alone are sufficient for e_i to be constituted as an emotion of s_i in $\mathcal{Q}$. Of course, both $I(s_i)$ and $C(s_i)$ are necessary for an emotional occurrence. Yet, something is still required —in particular, that another subject s_j in $\mathcal{Q}$ *recognizes* e_i in s_i. The act of recognizing the emotion e_i expressed by s_i consists, in this case, in an *attribution* of e_i to s_i made by s_j on the basis of $C(s_i)$:

$$C(s_i) \rightsquigarrow e_i \tag{7.1}$$

whereby the relation between $C(s_i)$ and e_i may take for s_j the form of a logical (deterministic, probabilistic, counterfactual, bayesian, etc.) inference[112] (see e.g., [Ong et al. 2019; Becker et al. 2021; Goel et al. 2024]), but also of other mechanisms such as neural association (see e.g., [Bastiaansen et al. 2009; Jeon and Lee 2018]), embodied simulation (see e.g., [Gallese 2005; 2007]), ToM (see [Footnote 94]), affective empathy (see e.g., [Singer et al. 2004; Bernhardt and Singer 2012]), etc. through which s_j comes to attribute e_i to s_i. Importantly, however, the relation between $C(s_i)$ and e_i is not absolute. It is conversely mediated by a number of cultural, social, and personal factors (see [Section 3.3] and [Section 6.2]) that are part of an interpretive code shared by all S in $\mathcal{Q}$, through which emotional attributions become possible —that is, through which S can *make sense* of C as relevant affective information to attribute emotions to others.

A subject s_i may instantiate all C coherent with an emotion e_i. In $\mathcal{Q}$, however, s_i can be said to have e_i only insofar as other S recognize —i.e., attribute— e_i to it on the basis of those C. Without such at-

[112] Of course, this may be the case also for humans. Yet, remember that I am talking here not necessarily about human S, but also about artificial agents (like robots).

tribution, s_i may still generate e_i, but this e_i does not constitute itself *socially* and *interactionally* as an emotion in $\mathcal{Q}$, because it would be *not recognized* as such. So, for example, a robot may be designed with mechanisms for expressing a state of fear by patterns such as retreating movements, elevated pitch in vocal output, rapid blinking lights, etc. From the perspective of the designer, who developed its internal computational architecture, one could say that, when those patterns occur, the robot is in a state of fear. Yet, unless (human) observers interpret these patterns as signs of fear (and not e.g., of bugs or motor malfunctions) —thereby (correctly, [Problem 3]) attributing fear to the robot according to a shared interpretive code about those patterns— that state does not acquire any social or interactional meaning as fear, remaining merely a programmed reaction to given stimuli. Only through an emotional attribution does the robot's internal state and expressive behavior *become* socially constituted as an emotion within the interaction.

Thus, from an interactional point of view, robotic emotions cannot be reduced to the mere outcome of implementation. They are *not only* the result of designing in the robot computational processes for emotion generation and systems for emotion expression, *but also* the product of the human interlocutor's act of recognition and attribution. In a social context, design and attribution are two inseparable and complementary sides of the same coin. They are jointly constitutive of what counts as an "emotional" state in human-robot interaction.

7.1.1.1 Objections and Replies

Against this claim —interdependency and complementarity between emotion design and emotion attribution in robots, some objections may be raised. In the following, I attempt to respond to the most significant of them.

(O1) ***Ontology.*** An emotion exists as an internal state even if it is not recognized by others: A person who feels fear in an empty room still experiences it, just as a robot with mechanisms functionally corresponding to fear would have that state regardless of external attribution. To reduce emotion solely to the social dimension risks turning it into a mere interpretative construct.

(R1) I do not claim that an emotion is *solely* dependent on the fact that it is attributed by someone. I claim just that «without such attribution, s_i may still generate e_i, but this e_i does not constitute itself *socially* and *interactionally* as an emotion in $\mathcal{Q}$, because it is not recognized as such». In other words, I do not deny that the subject may still continue to have an emotion even in the absence of an attribution. Rather, my focus here is on the social and interactional dimension of this emotional occurrence. In a social context, the mere generation and expression of fear by s_i is not sufficient: Others must also be able to recognize that s_i is afraid, and to do so they must know how to (appropriately) interpret its expressive cues.

Consider, for instance, the case of an ancient temple. On its walls there are signs, apparently mere scratches: Nobody realizes they are the letters of an ancient alphabet until an old missionary's booklet is discovered that identifies them as such. Ontologically, these signs have always been writing (even without us), and they remain so even if no one recognizes them as such. Yet, they acquire meaning as writing for us only once an interpretative code allows us to recognize them not as random marks but as alphabetic symbols —and this, even before asking questions about their semantic content.

I take here emotions in a social context to work nearly in the same way. From an ontological point of view, in absolute terms, the robot's emotion always exists[113]. Yet, for me, it acquires social meaning only if I am able to recognize it —and its expression— as such, even prior to identifying which particular emotion it is.

(O2) ***Functionalism.*** According to functionalism in philosophy of mind, what matters is the causal and functional organization of internal processes, not their attribution by external observers. If a system produces input–output patterns consistent with the emotion "fear" then it has fear (or a functional equivalent) independently of social recognition. In this view, attribution appears secondary and almost epiphenomenal.

[113] In the present work, I do not enter the debate about whether they exist as substantial states or functional states.

(R2) As in [R1], my focus is here on emotions in a social and interactional context, where, as explained, attribution plays a constitutive role. Moreover, anthropological, ethnological, sociological, philosophical and historical research has provided strong evidence that there is no direct correspondence between input–output patterns and particular emotions [Section 3.3]: The same processes and expressions can carry very different emotional meanings across cultures, societies and historical periods.

(O3) ***Evolutionary Continuity.*** In nature, many non-human animals possess emotions (see e.g., [Darwin 1872] and [Section 3.2.1.2]), even when these are not recognized or correctly interpreted by humans or by other animals. An attribution-based view seems to risk making such emotions *invisible* as though they had no reality unless explicitly acknowledged, thereby raising the charge of anthropocentrism.

(R3) I do not deny that animals —like robots— may have emotions independently of recognition. My claim concerns only the social and interactional domain: Here, emotions acquire significance as such only when they are recognized. Thus, if an animal's emotion is not recognized, it remains ontologically present but socially "invisible", without direct consequences for how we relate and respond to it (see [R1] and [R2]).

(O4) ***Pragmatics.*** From an engineering perspective, tying the very existence of robotic emotions to human recognition may complicate design. Engineers would then need to account not only for internal processes and expressive outputs, but also for the cultural and interpretative dynamics of observers. This could shift the goal from developing autonomous emotional systems to merely optimizing the readability of expressions for humans, potentially reducing innovation at the internal level.

(R4) My claim concerns precisely this idea. If our aim is to make robotic emotions (socially) more truthful and effective, then engineers must consider in their design not only internal processes and expressions,

but also the cultural and interpretative frameworks of human observers, i.e., how the implemented emotions are *actually* recognized. As anthropological research has demonstrated ([R2], [Section 3.3]), these dimensions shape human affective life and cannot be disregarded if the aim is to design robots whose emotional lives are comparable to ours. The support of correct attribution, the facilitation of recognition, and the prevention of misattributions [Section 6.2] thus have to become central concern also in robotic design.

(O5) ***Phenomenology.*** An emotion, particularly for a conscious being, has a subjective experiential dimension (*qualia*). Pushed to the extreme, the present perspective may appear to disregard this dimension entirely. For robots this may not be problematic, since there is no reason to assume they possess qualia [Footnote 113], but in comparison with humans the distinction is crucial: If human fear were only what others attribute, the lived experience of fear —the internal "what it feels like"— would be lost.

(R5) I do not make assumptions about the ultimate intrinsic nature of emotions, restricting my analysis only to certain aspects of it —primarily physiology, behavior, and cognition. Also, my interest is not in human emotions as such but in robotic emotions. Here, namely, wether robots and computers have qualia is still debated [Footnote 113], which is the case also for humans themselves (see e.g., [Turing 1950, 446ff.], [Nagel 1974], etc.).

(O6) ***Misattribution.*** If social attribution is taken as constitutive of robotic emotion, the risk arises that an observer may attribute an emotion even when no corresponding internal process is present. Random or meaningless movements could, for instance, be interpreted as signs of sadness or fear, thereby creating the appearance of an emotion where none exists internally. This would expose the concept of robotic emotion to complete arbitrariness, reducing it to the mere projections of observers and undermining its theoretical robustness.

(R6) As previously argued ([Section 7.1.1] and [R1]), attribution is not a sufficient condition for a robot to have an emotion. Processes of emotion generation and expression must necessarily sustain such attribution. The problem is therefore not whether emotions can be constituted solely through external recognition, but how to design robotic emotions in such a way as to foster correct and appropriate emotional attributions [Problem 3]. As shown, basing the design of robotic emotions on PBR [Chapter 6] heightens the risk of emotional misattribution [Section 7.1.2]. For this reason, it is necessary to incorporate into robotic emotional design additional mechanisms —for instance of cognitive–propositional nature— that may foster the avoidance of misattributions of (particular) emotional states (see also [R4]).

(O7) ***Relativism.*** If robotic emotions depend on a "shared interpretative code", the same emotion might exist in one social context but not in another. This would seem to imply an excessive relativism: The robot's "fear" would no longer be a stable emotion, but a phenomenon varying across cultures and interpretative frameworks.

(R7) This consequence is not problematic but rather consistent with what we know about human emotions in social and cultural life As anthropological research has demonstrated ([R2], [R4], [Section 3.3]), emotional expressions and their recognition are shaped by interpretative frameworks that differ across contexts, and robotic emotion design must therefore take such variability into account, implementing measures to accommodate it.

(O8) ***Anthropomorphism.*** Closely related to [O1], [O3], and [O6], one might object that humans can attribute emotions or other mental states to almost anything —animals, objects, or artifacts— as extensively discussed in the literature on anthropomorphism (see [Footnote 59], [Footnote 93] and [Problem 5]). Such attributions, however, do not in themselves constitute evidence that the corresponding states exist on the side of the robot.

(R8) The objection correctly highlights that humans are capable of attributing emotions to almost any entity, even in the absence of a corresponding internal state (as in the case of naïve anthropomorphism). This confirms that attribution alone is not sufficient to constitute a robotic emotion. According to the present perspective, the social constitution of an emotion in a robot requires the coexistence of two *necessary*, and *alone not sufficient* conditions:

(a) *Internal condition* (*generation and expression*). The robot must possess mechanisms capable of generating emotional states (however computationally implemented) and of producing expressive patterns coherent with those states.

(b) *External condition* (*attribution*). Interlocutors must recognize those expressive patterns as manifestations of an emotion, based on shared cultural and interpretative codes.

Attribution without the internal condition remains an unfounded misattribution (as when an observer interprets the trajectory of a vacuum cleaner, say, as "sadness"). Conversely, internal generation alone is not sufficient to constitute an emotion in the social domain unless it is also recognized as such. A robotic emotion is thus constituted only where generation/expression and attribution integrate, thereby distinguishing it both from mere anthropomorphic projections and from internal states lacking social relevance.

7.1.2 From Robotics to HRI

From a disciplinary standpoint, the interdependence between emotional design and attribution is mirrored by the relation between robotics and HRI (*Human-Robot Interaction*). Whereas the former focuses namely on the design, engineering and implementation of affective architectures in robots, the latter investigates the processes of recognition, attribution, and interaction through which those architectures acquire emotional significance for humans in the concrete practice of a social environment in which the robot happens to be engaged.

HRI has emerged as a relatively recent field of inquiry at the intersection of multiple disciplines, including AI, robotics, cognitive science, psychology, neuroscience, philosophy, linguistics, and HCI (see [Bell 2018] and [Section 4.1]). As an autonomous discipline, it has taken shape since the 1990s, when the need to study systematically how humans engage with robots has started to become increasingly compelling. The IEEE RO-MAN (*International Symposium on Robot and Human Interactive Communication*) conference series —initiated in 1992 in Japan and since then established as a major international venue in the field— has played a central role in consolidating HRI as a research community and in defining its research agenda. Yet, the cultural roots of this new scientific discipline can be traced back to the science fiction literature of the 1940s-1950s [Section 1.2.3], for example with Asimov's [1950] early reflections on the role of robots in human society[114].

HRI investigates how humans perceive, interpret, and interact with robotic systems. Central research questions concern, for instance, how humans attribute mental states to robots (see e.g., [Manzi et al. 2021; Goldman et al. 2025; Maier et al. 2025], as well as [Gray et al. 2007; Thellman et al. 2022; Jastrzab et al. 2024; Nääs et al. 2025]), how different modalities of robotic expression (e.g., facial, vocal, or gestural) and degrees of anthropomorphism influence such attributions, how cultural and contextual factors shape human responses to robotic agents, how design choices affect the perceived naturalness, trustworthiness, and effectiveness of interactions, etc. (see e.g., [Dautenhahn 2007; Sheridan 2016; Bartneck et al. 2024]). In this respect, HRI exhibits a twofold orientation: (1) *descriptive* —in that it conducts empirical analyses of the

[114] In Asimov's [1950] famous collection *I, Robot*, the short story *Runaround* (1942) is contained, which first introduced Asimov's formulation of the *Three Laws of Robotics* Asimov [1950, 37]:

1. *The First Law*: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. *The Second Law*: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. *The Third Law*: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws will be partly the basis on which both the ethical reflection about robots in society and HRI studies will later develop.

processes by which humans actually relate to robots, attributing to them intentions and social roles; and (2) *normative* —in that, also drawing on empirical evidence from (1), it formulates principles and design guidelines for working roboticists in order to foster the development of robotic technologies capable of sustaining more effective, meaningful, natural, and less frustrating interactions with humans. For this reason, of course, HRI inevitably maintains a close and reciprocal relation with robotics engineering.

Of course, given the increasing presence of robots in social environments [Section 6.2], understanding *affective interactions* between humans and robots has become a central concern for robot psychologists and HRI specialists. Emotions enhance the perceived transparency of these interactions, enabling individuals to interpret robotic behavior through familiar social signals learned from prior human-human exchanges [Embgen et al. 2012]. As social robots progressively approximate human form and conduct, their emotional expressions increasingly facilitate and enrich social engagement with humans (see [Parkinson et al. 2005; Butler et al. 2003; Haslam 2006], and also [Stock-Homburg 2022, 389]). Accordingly, affective HRI is directly tasked with supporting affective robotics «to make robots more understandable, likable, intuitive, and predictable (or "believable") by using patterns that allow people to apply mental models and heuristics from interactions with humans to infer a robot's internal states and intentions» [Jung 2017, 263].

7.1.2.1 PBR in HRI

Now, the (empirical) investigation of how humans attribute emotional states to (humanoid) robots has emerged as one of the primary ways in which HRI responds to this task. Until now, indeed, there has been a discrete effort devoted to human-robot emotional attributions[115] (see e.g., [Jung 2017; Fischer et al. 2019; Stock-Homburg 2022]). Yet, the way in which investigation on the topic has been approached raises im-

[115] Yet, there is some overlapping between the research about *mental states attribution* and the research about *emotion attribution*, since emotions are often treated as mental states [Section 7.1.2]. Allegedly, in fact, research effort *specifically* and *explicitly* targeting human-robot emotional attributions as an *autonomous* and *delimited* phenomenon is minor compared to research on mental state attribution in general.

portant epistemological concerns, calling into question the adequacy of the theoretical frameworks and assumptions guiding current research.

To begin with, emotions have usually been absorbed into the broader research question of mental states attribution [Footnote 115]. In particular, they have been considered alongside beliefs, intentions, or decisions indistinctly as "mental states", without regard for their complex nature, or for the (possibly) specific processes through which they are attributed (see e.g., [Thellman et al. 2022]). As a result, their evaluative, expressive, and embodied dimensions —as well as their dynamic interplay with other non-affective mental states— have been systematically overlooked, preventing the acknowledgement of their unique role in shaping human-robot interaction.

Moreover, paralleling the same tendency recognizable in robotics [Chapter 6], HRI research on emotional attributions has likewise typically been conducted within a PBR-based theoretical framework, often implicit and uncritically adopted (see e.g., [Stock-Homburg 2022; Rawal and Stock-Homburg 2022; Rosenthal-von der Pütten et al. 2025]). A conception of emotion has largely been presupposed that identifies emotional states primarily with physiological signals and behavioral cues, and that consequently considers PB-related characteristics C of the robot as the only (causally) relevant factors triggering and directing human emotional attributions to it. This view has accordingly shaped the design of experiments in the field, the observational and measurement methodologies employed, the interpretation of empirical data, and the evaluation of findings.

Generally, observable (multimodal) behaviors exhibited by the robot are taken as the *independent variable* —in most cases, for example, facial expressions (see e.g., [Konijn and Hoorn 2020; Lombardi et al. 2024; Mishra et al. 2024]), verbal and non-verbal sounds (see e.g., [Liu et al. 2023; Gasteiger et al. 2024; De Marchi et al. 2024]), lights (see e.g., [Marmpena et al. 2018; 2020]), gaze (see e.g., [Cucciniello et al. 2023; Fiorini et al. 2024]), position of robotic ears (see e.g., [Eimler et al. 2011]), mouth, antennas, and appendices (see e.g., [Lakatos et al. 2014]), touch (see e.g., [Sawabe et al. 2022]), color and temperature changes (see e.g., [Peña and Tanaka 2020]), breath (see e.g., [Klausen et al. 2022]), etc.

The *dependent variable*, i.e., the attribution of an emotion to the robot by a human, is sometimes understood in terms of a neural and/or physiological response to the robotic behavior —for example, particular patterns of brain activity, measured by fMRI (see e.g., Rosenthal-von der Pütten et al. 2013; 2014) and EEG (see e.g., [Suzuki et al. 2015; Shao et al. 2020; Kiilavuori et al. 2022; Staffa et al. 2023]), of facial muscle activation, measured by surface EMG on the zygomatic and corrugator regions (see e.g., [Kiilavuori et al. 2022]), of electrodermal activity (C-tactile afferents), measured through skin conductance level sensors (see e.g., [Sawabe et al. 2022; Kiilavuori et al. 2022]), of cardiovascular responses, measured by heart rate monitors and arousal indices (see e.g., [Grundke et al. 2022]), of body temperature variations, measured by thermal imaging systems (see [Peña and Tanaka 2020]), of ocular behavior, measured through eye–tracking technology registering gaze direction and fixation (see e.g., [Fiorini et al. 2024]), of kinematic parameters of motor behavior, measured through motion-tracking systems capturing velocity, acceleration, or height during interaction (see e.g., [Lombardi et al. 2024]), etc.

In addition, attribution is operationalized through *self-report instruments*, for example *forced-choice* (FC) *emotion recognition tasks* in which participants are asked to classify the robot's behavior within a restricted predefined set of basic emotions (see e.g., [Lakatos et al. 2014; Peña and Tanaka 2020; Mishra et al. 2024]). Such approaches are typically grounded in universalist theories of emotions —prevailing in HRI [Jung 2017; Fischer et al. 2019; Stock-Homburg 2022; Vagnetti et al. 2024; Rosenthal-von der Pütten et al. 2025], drawing on Ekman and Friesen's [1978] taxonomy of basic affective categories formalized in the FACS (see e.g., [Lakatos et al. 2014; Peña and Tanaka 2020; Cucciniello et al. 2023; Liu et al. 2023; De Marchi et al. 2024; Mishra et al. 2024]). Hereby, attribution is framed within the assumption that emotions are discrete, cross-culturally stable, biologically determined universals, univocally correspondent to specific PB-patterns [Section 3.3].

In other cases, attribution is instead studied along continuous affective dimensions —most prominently using VAD (valence-arousal-dominance) and related models [Section 3.3]— with operationalizations typically relying on standardized protocols that combine the *International*

Affective Picture System (IAPS) as stimulus set with the *Self-Assessment Manikin* (SAM) as a tool for measuring responses (see e.g., [Marmpena et al. 2018; 2020; Coyne et al. 2020; Kiilavuori et al. 2022; Sawabe et al. 2022; Prati et al. 2022; Staffa et al. 2023; Fiorini et al. 2024; Heinisch et al. 2024]). Yet, even the adoption of continuous scales for assessing emotional attributions does not overcome the reductionism, since PB-related elicitors remain the independent variable[116].

The foundational role of evaluative, contextual, propositional and, more generally, cognitive dimensions in human-human emotional attribution is widely acknowledged within social psychology (see e.g., [Harris et al. 1989; Hadwin and Perner 1991; Ruffman and Keenan 1996; Doorn et al. 2015; Wu and Schulz 2018; Scherer et al. 2021]). Nevertheless, even when some HRI studies display a more or less explicit awareness of these dimensions as integral to affective experience, the exploration of their role in human-robot emotional attribution remains largely unaddressed. In some cases, the employment of tools such as IRI (*Interpersonal Reactivity Index*), AMS-Q (*Attribution of Mental States Questionnaire*), GQS (*Godspeed Questionnaire Series*), etc. (see e.g., [Weiss and Bartneck 2015; Tobis et al. 2023]) has allowed to connect attribution to empathic and mentalizing dispositions, or to highlight participants' occasional appeals to reasons and justifications as part of their attribution process (see e.g., [Herzog et al. 2022; De Marchi et al. 2024]). Yet, these dimensions, while recognized, have been rarely formalized and systematically operationalized in experimental designs and testing practices, which continue to be predominantly framed within a PBR-based research paradigm.

Besides neglecting the contribution of the cognitive dimension in emotional attributions, present-day affective HRI research is further constrained by other intrinsic limitations of the experimental methodologies on which it relies [Rosenthal-von der Pütten et al. 2025]. Also by systematically inflating the role of PB-related cues and marginalizing alternative explanatory factors, these methodological practices make current findings narrow in scope and of questionable reliability.

[116] A more extensive review of the epistemological and methodological tendencies in current affective HRI, and in particular in the study of emotional attribution to robots, is forthcoming [Footnote 98].

Often implicitly, for example, the very structure of experiments on emotion attribution and recognition in HRI introduces systematic biases. Recognition rates, in fact, are not a neutral and objective reflection of participants' perceptual capacities, but a contingent product of specific choices made by the experimenter about the structure itself of the experiment and the response formats to be employed. Whether participants are exposed to a single stimulus or to multiple ones (*within-* or *between-subject* experimental design), and whether their responses are constrained through closed formats (e.g., FC test) or instead elicited via scales and open-ended answers (e.g., SAM), has a direct impact on the experimental outcomes (see also [Russell 1994]). FC tests, for instance, artificially raise recognition rates by constraining responses, whereas open formats (free text, graded scales) produce lower but more results outcomes (see e.g., [Schiano et al. 2000; Frank and Stennett 2001; Winters Fisher 2005; Romashov 2018]). Also, the widespread reliance in affective HRI on universalist conceptions of emotions reinforces this methodological bias, in that standardized and pre-selected expressions —often posed or artificially programmed in robots— facilitate recognition in the lab, but at the cost of ecological validity. Such settings namely neglect the contextual and ambiguous nature of real-life emotional expressions (see e.g., [Read and Belpaeme 2014]), thereby frequently overestimating the transparency of robotic behaviors, that appear "clear" under experimental conditions, but possibly less interpretable in real interactions.

The neglect of the cognitive dimensions of human-robot emotional attribution seems thus to be paralleled by methodological choices that directly bias HRI research toward PBR. By structuring experiments in ways that artificially boost the rates of PB-related emotional attributions and suppress the ambiguity typical of real-life affective exchanges, current methodologies reinforce the very reductionism that sidelines the cognitive and contextual components which are, by contrast, constitutive of actual emotional phenomena.

Therefore, a urgent need emerges to integrate such dimensions into the theoretical framework of empirical HRI research on emotional attribution. Empirical protocols must be developed that treat non-PB components of emotion as constitutive elements of attribution, rather than as secondary or peripheral "by-products" of the process. This re-

quires the conception of adequate test formats that explicitly and robustly operationalize the cognitive dimension of human-robot emotional attribution. These may include for example, tasks and questions designed to assess how participants ground their emotional attributions in the reasons they ascribe to the robot's expressive behavior, the justifications and explanations they consider plausible for it, the contextual framing of the interaction, and the evaluative beliefs and non-affective mental states they attribute to the robot along with the emotion. Experimental designs have to be developed that consistently and actively integrate the robot's propositional attitudes (desires, beliefs, intentions, etc.) into the set of characteristics C under investigation with respect to their possible (causal) relevance in emotional attributions. Measuring practices, accordingly, should not be restricted to physiological signals or forced-choice classifications of basic emotions, but extend to indicators of cognitive evaluation —such as participants' reactions to explanations provided by the robot, the degree of their preference for agents that justify their actions, or their propensity to trust and empathize depending on the reasons they perceive as guiding the robot's emotional behavior. It is not sufficient merely to acknowledge the presence of cognitive dimensions in robotic emotions and human-robot emotional attributions in HRI. These dimensions must be incorporated as genuine independent and dependent variables within HRI experimental designs *together with* the traditional PB-dimensions.

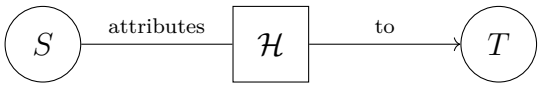
Now, concentrating on HRI and on the empirical problem of emotional attribution, rather than directly on affective robotics and design, offers a privileged standpoint for confronting the limitations of the PBR paradigm. This does not imply that robotic emotions *per se* are irrelevant, nor that PBR cannot be addressed at the level of design. Yet, in the present work, the focus is deliberately placed on attribution. As argued above [Section 7.1.1], design and attribution are interdependent; but it is through attribution that this interdependence becomes empirically accessible, since attribution determines the very criteria by which robotic emotions are recognized, evaluated, and ultimately deemed successful. Moreover, since HRI offers robotics normative reference points to robotics, rethinking its epistemological assumptions exerts a consequential influence on the direction of robotic design.

On this basis, the next step in [Section 7.2] and [Section 7.3] is to outline a framework in which attribution is no longer confined to PB-dimensions but integrates cognitive —and specifically *propositional*— components as constitutive of robotic emotions and of the processes through which they are attributed to them by humans.

7.2 The Propositional Dimension of Emotions

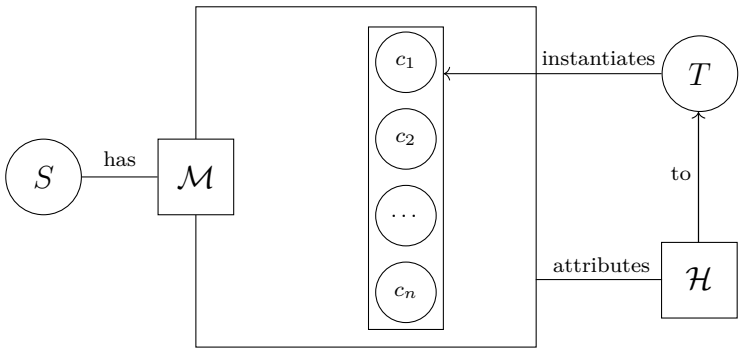
7.2.1 The Concept of Attribution

Attribution can be understood as a complex process, grounded in neural and physiological mechanisms, as well as in higher-level cognitive operations (see e.g., [Fiske and Taylor 1991; Bricker 2020; Malle 2007; 2011; 2022; Quilty-Dunn and Green 2023; Quesque and Apperly 2024] and [Section 7.1.1]). In the last case, it is typically described as a process through which individuals interpret observed events or other's behaviors by organizing them into coherent and understandable patterns based on contextual features, personal background and expectations. As such an act, attribution is still debated in philosophy, cognitive science, and social psychology with regard to its exact nature, its components, phases and the conditions under which it occurs. Put differently, it is still controversial what is precisely meant, for a certain subject S, to attribute a certain state $\mathcal{H}$ to something or someone T:



—whereby $\mathcal{H} = \{E, B, D, ...\}$ (i.e., emotions, beliefs, desires, etc.) and T ("target") can be *the same or another* subject S, an (inanimate) object, an event, a situation, *a robot*, etc. Yet, for the purpose of the present section, I take the (cognitive) act of attribution to be the adoption by S of a *theory* or *model* $\mathcal{M}$ by which S attributes a state $\mathcal{H}$ to T. Indeed, referring to the formalism introduced in [Section 7.1], $\mathcal{M}$ is

what allows S to make sense of those (implicit or explicit) characteristics $C = \{c_1, c_2, \ldots, c_n\}$ of T that $\mathcal{M}$ itself selects as necessary and/or sufficient for the occurrence of $\mathcal{H}$:



In a given sense, thus, $\mathcal{M}$ *defines* for S what a given state $\mathcal{H}$ is, so to enable her to recognize it in T. In a given social context $\mathcal{Q}$, it is $\mathcal{M}$ that, given the presence of expressed characteristics C —produced by I— in T, selects and organizes (possibly only some of) them in a recognizable pattern (the state $\mathcal{H}$) to attribute to T, whereby $\mathcal{M}$ is influenced by a number of cultural, social, and personal factors relative to S [Section 7.1.1] that act on which C are taken as relevant and characteristic of $\mathcal{H}$ within the attribution process.

Now, attributable states also encompass emotions E. This means that certain patterns of characteristics C, when selected and organized through $\mathcal{M}$, are recognized as emotions and thereby attributed as such to the target T. In this respect —also in light of the interdependency between design and attribution discussed earlier [Section 7.1.1]— the question of what an emotion is, namely which C may be considered necessary and sufficient for (a given) E, can be approached as a question concerning the nature, structure and functioning of $\mathcal{M}$.

As seen in the previous chapters, several theories of emotions in philosophy, psychology and medicine have tackled this question, proposing different kinds of answers. Physiological theories of emotions like JLTE [Section 3.2.1.3] or CBTE [Section 3.2.1.4], for example, have highlighted physiology-related C as relevant in E and therefore in $\mathcal{M}$ —say, George sees Anna's face turning pale and her hands trembling, and so he recognizes she is afraid. Behavioral theories (see also [Section 3.2.1.5], among

others) have instead focused on outward expressions and actions as the defining characteristics C of E, which are also those primarily considered in $\mathcal{M}$ for emotional attributions —say, George notices Anna shouting and striking the table, and so he takes her to be angry. Cognitive and appraisal theories [Section 2.3], in turn, have emphasized evaluative characteristics C as relevant in E and in the model $\mathcal{M}$ for the attribution of emotional states —say, George learns that Anna has just lost a competition that mattered to her, and so he understands she is sad. Within the cognitive approach, then, further theoretical articulations specify the relation between emotion, cognition and physiology. *Judementalists* (see e.g., [Solomon 1976; Nussbaum 2001]), for example, claim that emotions *are essentially* cognitive evaluations of events and situations. Sustainer of *two-factor theories* (see e.g., [Schachter and Singer 1962]) that physiological arousal is a necessary, but undifferentiated component of emotion, which acquires affective meaning only through subsequent cognitive labeling. Proponents of *appraisal theories* (see e.g., [Arnold 1960; Lazarus 1984; Frijda 1986; Ortony et al. 1988; Scherer 1999]), finally, defend a view in which cognitive evaluations antecede and elicit both the physiological activation and the experiential quality of emotions, typically through structured appraisal dimensions such as goal relevance, accountability, and coping potential.

There is still an ongoing debate (see e.g., [Scarantino and Sousa 2021]) on which kinds of C —physiological, behavioral, cognitive, etc. — hold a logical and ontological priority in defining what counts as an emotion E —that is, which kind of C can be regarded as necessary and/or sufficient for its occurrence, and thus as primarily relevant in $\mathcal{M}$ for determining its attribution. Indeed, it is beyond the scope of the present work to take a position in this debate. For the purposes of the present analysis, I rather adopt a *pluralistic* stance on the nature of E and $\mathcal{M}$: Components of different kinds —physiological, behavioral, cognitive, etc. — can be legitimately considered as characteristic of the nature of emotions and relevant in their attribution, while the logical and ontological priority (the "dosage") among them may vary depending on the particular emotional state, on contextual circumstances, and on a number of different cultural, social and personal variables (see also [Section 7.2.2.3] and [Section 7.2.3]) —when Anna jumps and screams at a snake, George

attributes fear primarily on the basis of her automatic reactions; Anna smiles during her performance review, but only when George learns she has been denied a promotion does he attribute polite disappointment rather than satisfaction; Anna raises her voice in a meeting, but George does not take it as anger, knowing that such expressiveness is typical of her.

7.2.2 Emotions and Intentionalty

Emotions, thus, include several different components [Section 7.2.1]. Some of them are *external* (such as the physiological and the behavioral ones), while others *internal*, that is, related to the individual's mental life (such as phenomenology, appraisal, evaluation, etc.). Suppose George suddenly learns that someone intends to kill him. His heart rate increases, he cries, cannot stop moving, informs his family and the police, experiences unpleasant subjective feelings (apprehension, dread), and forms a series of cognitive judgments evaluating the event (death threat) as something negative and undesirable for him. On the basis of such different features (C) instantiated by George, we ordinarily say that George is in a certain emotional state (E), namely, fear. Sharing his same social environment ($\mathcal{Q}$) and cultural background, that is, a similar attributional model $\mathcal{M}$ linking those features to fear, we therefore ascribe to him this particular emotional state.

7.2.2.1 A Taxonomy of Mental States

Now, undoubtedly, since emotions have mental components, they are in part *mental states*. Mental states are traditionally divided *intentional* and *non-intentional* ones (see e.g. [Searle 1983; Crane 2009; Jacob 2023]). In general, intentionality is the property possessed by those states which have a *content*, i.e., are "directed at", or "about", objects and states of affair internal or external to the mind. For example, in "George loves his dog", the dog is the intentional content of George's mental states (loving).

Sometimes, *representational* and *non-representational* mental states are distinguished within the realm of intentionality (see e.g., [Pitt 2022]). Representational states are those whose content involves presenting ob-

jects or situations as having certain features or properties. For example, in "George thinks that his dog is cute", George represents the intentional content (the dog) of his mental state (thinking) as being in a certain way (cute). In other words, his mental state represents a certain object in the world as having a certain property, whereby representational mental states can be assessed commonly for truth, accuracy, or correctness (it may be the case that George's dog is not actually cute). Non-representational mental states are instead directed at objects or situations without attributing to them any specific representational property —e.g., in "George focuses on his dog", George's mental state (focusing) is directed to an intentional object (the dog) which is however not represented in any particular way, nor is his state assessable for truth, accuracy, or correctness.

Representational mental states can be further divided into *conceptual* and *non-conceptual* ones (see e.g., [Bermúdez and Cahen 2024]). In the former, the subject can entertain the state only by possessing the concepts that structure its content. "George thinks that his dog is cute", for instance, is conceptual because it presupposes George possessing the concepts "dog" and "cute". In the latter, on the other hand, the state represents objects and features in the world without requiring the subject to possess the corresponding concepts. A common example of intentional, but non-conceptual mental states are perceptions: George, for instance, does not have a concept for each shade of gray he can perceive of his dog's fur.

Moreover, it is possible to categorize conceptual mental states into *propositional attitudes* and *non-propositional* (or *objectual*) *attitudes* (see e.g. [Grzankowski 2013; 2015; Grzankowski and Montague 2018]). Whereas the content of the former is expressible in propositional form, that of the latter is an object or situation that cannot be captured by a proposition. For instance, George's love for his dog has an intentional content (the dog), which is however not a proposition —nor, at least on the main account[117], susceptible of propositionalization. Instead, in "George

[117]The nature of propositions and propositional (doxastic) attitudes is indeed highly controversial (see e.g., [Kripke 2011; Elgin 2017; McGrath and Frank 2024; Schwitzgebel 2024]).

thinks that <*p*>", George's mental state (thinking) has an intentional content <*p*> of propositional form, say <*the dog is cute*>.

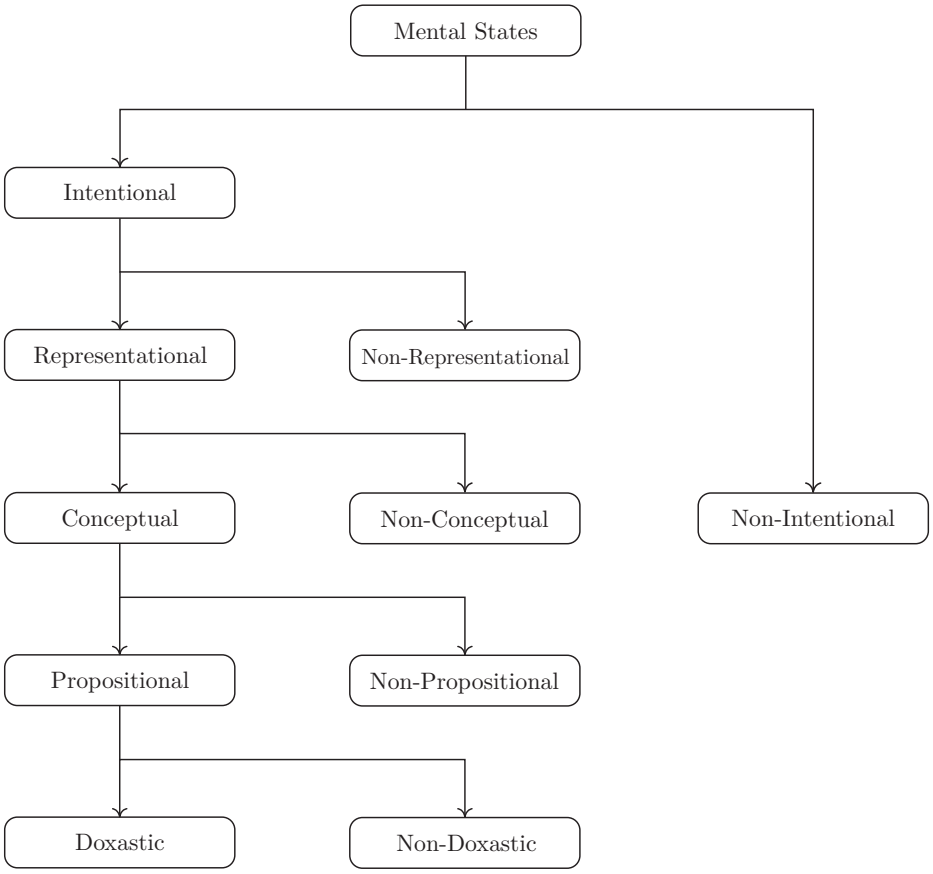


Figure 7.1: An approximate (*not epistemologically nuanced*) taxonomy of mental states.

Doxastic attitudes (from the Ancient Greek δόξα, *dóxa*, meaning "opinion") are then a particular subclass of propositional attitudes containing mental states that express a subject's commitment to the *truth* of a given proposition (see e.g., [Shah and Velleman 2005]). Traditional doxastic attitudes are considered beliefs, disbeliefs, knowledge imagery, judgments, etc. In "George believes that <*p*>", for instance, George is taking the proposition <*p*> (<*the dog is cute*>) as true —of course, independently from the corresponding fact actually obtaining in the world, and whereby it is not possible for someone to believe something false, because in this case one would normally not have the belief either (see [Footnote 117]). Other *non-doxastic* propositional attitudes moreover involve

conative and volitional states such as desires, hopes, etc. —e.g., "George wants that <*p*>", as well as *aliefs* [Gendler 2008], i.e., particular belief-like attitudes standing in contrast with a subject explicit beliefs (see [Problem 4(d)]).

Finally, *non-intentional* mental states (intrinsically non-representational, non-conceptual, non-propositional and non-doxastic) consist in purely qualitative or phenomenal aspects of experience that are *not* directed at particular objects, events, or situations in the world —such as, in general, sensations of pain, bodily discomfort, etc.

7.2.2.2 Emotions as Mental States

The taxonomy of mental states [Figure 7.1] just exposed in [Section 7.2.2.1] is far from uncontroversial. Numerous borderline cases and competing classifications have been advanced, with plenty of discussions and accounts ranging, say, from the intentionality of non-conceptual mental states to the propositionality of perceptions and the truth-aptness of propositions, making the epistemology of mental states a notoriously intricate and contentious research field.

This outline above, however, shall be sufficiently clear and comprehensive to introduce the very problem of this section, namely the nature of emotions as mental states. Indeed, there is presently a lively debate about this topic, concerning how and where emotions should be situated along this taxonomy.

A long-standing line of thought has taken emotions as one of the major and most direct instances of non-intentional mental states (see e.g., [Lamb 1987; Goldstein 2002; Whiting 2011]). Taken in themselves, at least in some cases, emotions are best understood just as elementary hedonic feelings, defined exclusively by their qualitative tone of pleasantness or unpleasantness. In this perspective, the intentional component that is often attributed to emotional life does not properly belong to the emotion as such, but is instead conferred upon it by other co-occurring mental states, such as beliefs, desires, or evaluative judgments, which provide the emotion with a direction and an object. Stripped of these ancillary contributions, thus, emotions reduce to mere sensations without intrinsic aboutness. In "George fears the dog", for example, George's

very emotional state (fear) consists merely in a feeling of unpleasantness, while its apparent aboutness is derived from other (non-affective) mental states (e.g., perceptions of the dog, beliefs about its dangerousness, etc.).

Apart from such radical positions, however, emotions are today regarded in epistemology, philosophy of mind and psychology as typically intentional states, essentially directed toward objects or situations in the world (see e.g., [Gunther 2003; 2004; Tye 2008; Deonna and Teroni 2015; Grzankowski and Montague 2018; Ratcliffe 2019; Tappolet 2020], etc.). So, for instance, the aboutness of George's fear is not derived from other mental states. On the contrary, it is essentially built into his emotional state from the outset, which therefore cannot be reduced to a mere hedonic feeling of unpleasantness.

While the intentionality of emotions is broadly acknowledged, contemporary philosophical debate is largely concerned with the precise characterization of this intentionality. In this respect, some scholars claim that emotions are non-representational states (see e.g., [Gunther 2003; Hutto 2012; Mitchell 2020; Tappolet 2020]). On this view, emotions do not present the world as being in a certain way. Rather, they are best understood as embodied patterns of affective response that are intrinsically action-guiding. The intentionality of emotions, in this perspective, does not consist in representing the world but in orienting the subject toward specific forms of practical engagement. In "George fears the dog", for example, George's fear is not a representation of the dog, say, as dangerous, but an affective disposition that primes him for avoidance or flight.

By contrast, others argue that emotion are representational, maintaining that they are intentional precisely insofar as they present the world under certain aspects (see e.g., [Prinz 2004; Colombetti 2017; Hufendiek 2015; 2018]). In particular, many affective representationalists take emotions to be grounded in bodily responses that function as representations of so-called *Core Relational Themes* (CRTs) [Lazarus 1991]. CRTs are recurrent patterns in the way the relation between an organism and its environment is appraised. They capture the adaptive significance of objects and situations for the individual by attributing a distinctive emotional meaning to them —e.g., fear to the presence of danger, anger to the experience of offense, joy to the recognition of benefit, and so forth.

In this view, the physiological and affective patterns that constitute an emotion amount to (embodied) representations of CRTs, and it is through this representational dimension that emotions acquire their intentional character beyond immediate enactive and action-guiding functions. In "George fears the dog", for example, George's bodily arousal does not merely dispose him to avoidance, but intrinsically represents the situation as one of danger. This representation confers intentionality upon his fear and can be assessed for adequacy insofar as his bodily arousal appropriately corresponds to the circumstances.

Then, for many representationalists, representations involved in affective experience are non-conceptual (see e.g., [Gunther 2003; Prinz 2004; Colombetti 2017; Hufendiek 2015; 2018; Mitchell 2020; Tappolet 2020]). In some cases, they are conceived as *perceptions of values*, representing external objects as bearers of evaluative properties and thereby assessable for correctness or justification, while remaining non-conceptual insofar as they do not presuppose possession of the relevant value concepts (see e.g., [Tappolet 2016]). In other cases, emotions are understood as intentional states with an intrinsically action-oriented content, representing the world in terms of possibilities for action that carry a motivational force for the subject without conceptual mediations (see e.g., [Murillo-Lara and Muñoz-Serna 2019]). In "George fears the dog", for example, George's emotion can be construed as a perception of value: His fear represents the dog as dangerous, a property evaluative in nature, even though George need not possess or apply the concept of danger for the experience to arise. Alternatively, the same emotion can be interpreted in action-oriented terms, as representing the encounter with the dog directly as a possibility for avoidance or fight, thereby motivating George's behavior without requiring conceptual mediation.

On the contrary, other affective intentionalists and representationalists (often implicitly) defend a conceptualist view of emotions, according to which their content essentially depends on the subject's possession of the relevant evaluative concepts occurring in the representation, independently of what such representation specifically consists in (see e.g., [Solomon 1976; Tye 2008]). On this account, to experience fear, anger, or joy toward something requires the ability to conceptualize it under the corresponding evaluative description —e.g., "dangerous", "of-

fensive", "beneficial", and so forth. In "George fears the dog", for instance, George's fear involves not only representing the dog as dangerous but also possessing the concepts of "dog", "dangerous", and the like.

While conceptualist accounts insist on the dependence of emotions on the subject's possession of evaluative concepts, this still leaves open the issue of how such concepts are articulated within the intentional (representational) content of emotions. According to several authors —including, of course, some non-intentionalists, non-representationalists, and non-conceptualists— this content is best understood as non-propositional (see e.g., [Gunther 2003; Grzankowski 2013; Grzankowski and Montague 2018; Mitchell 2020; Tappolet 2020]). Although emotions may presuppose evaluative concepts, their intentional content does not take the form of a proposition that can be assessed for truth or falsity. Rather, emotions are directed at objects or situations under evaluative modes of presentation without thereby representing that some evaluative fact obtains. Thus, in "George fears the dog", George's emotion does not amount for him to represent his appraisal of the dog as dangerous in the form of a proposition (<*the dog is dangerous*>). Instead, it consists in his apprehending the dog under the evaluative and objectual guise of danger, in a way that is conceptually mediated, yet not propositionally articulated.

Conversely, other philosophers defend a propositionalist account of emotional intentionality, according to which emotions are individuated by the propositions they represent and can therefore be evaluated in terms of truth or falsity (see e.g., [Pitcher 1965; Gordon 1973; Kerner 1982; Crane and Farkas 2022; Logins 2024]). On this view, for example, George's fear for the dog has the proposition <*the dog is dangerous*> as its intentional content.

Another discussion finally concerns the articulation of the propositional content of emotion as doxastic or non-doxastic. Doxasticists maintain that emotions are fully reducible to evaluative judgment or belief —so that, George's fear of the dog would consist in his judging or believing that <*the dog is dangerous*> (see e.g., [Pitcher 1965; Gordon 1973; Solomon 1976; Kerner 1982]). Non-doxasticists, on the other hand, argue that even when emotions are propositionally articulated, they cannot be reduced to beliefs, judgments and other doxastic attitudes (see e.g., [Crane and Farkas 2022; Logins 2024]).

7.2.2.3 Dynamic Evaluative Pluralism about Emotions

As emerges from the overview of the previous [Section 7.2.2.2], the very nature of emotions as mental states remains highly controversial. No account has so far succeeded in capturing the richness and variety of affective life in its entirety. Some positions articulate effectively the representational, conceptual and propositional dimension of affective phenomena, while leaving nearly unaddressed the role of physiology or phenomenology in determining and characterizing emotional occurrences (or vice versa). Also, depending on how one understands the very notions of "intentionality", "representation", "concept", or "proposition" —topics on which philosophers still disagree [Section 7.2.2.1]— the characterization of emotions as mental states shifts accordingly.

Moreover, from an epistemological point of view, the purported generality of proposed theories of emotions continues to face serious difficulties once confronted with issues concerning problems such as infant and animal emotions, or recalcitrant affective states [Problem 4(d)]. Embodied, physiological and behavioral theories of emotion, for example, tend to downplay the role of cognition —representations, concepts, propositions — in affectivity, thereby struggling to explain the emergence of complex emotional states [Problem 4]. Conversely, representationalist, conceptualist and propositional approaches risk over-intellectualizing emotional phenomena, rising thereby several theoretical problems: whether genuine intentional or propositional content can exist in the absence of linguistic or conceptual capacities —as namely in the case of infants and non-human animals; how such emotions might be assessed for correctness or rationality when no reasons can be articulated; or how to account for recalcitrant emotions that persist against explicit judgments and resist rational revision. *Anomalies* like these continue to occupy a central place in contemporary philosophical debate on affectivity, precisely because they force any theory either to dilute its explanatory criteria to the point of vacuity or to exclude paradigmatic phenomena of affectivity, thereby undermining the claims to generality.

Now, it is beyond the scope of the present chapter to put forward and philosophically defend a new, alternative theory of emotion that might resolve these tensions, to definitively provide necessary and suffi-

cient conditions for a (mental) state to count as an emotion, and settle long-standing debates in the philosophy of emotion —such as those concerning infant, animal, or recalcitrant emotions. Given the complexity and multifaceted character of the phenomenon, it seems unlikely —at least in my humble opinion— that any single theory could ever accomplish such a task. Nevertheless, I find it worth at least suggesting that the different accounts examined above may be regarded as *complementary*, each illuminating distinct aspects of the same phenomenon from different perspectives. None of these aspects, taken in isolation, is *necessary* for defining what emotions are. Yet, they can be seen as *sufficient* to constitute emotional occurrences under certain conditions, provided that they are integrated into a more general framework that illustrates at least the "symptoms" [Goodman 1968] distinguishing emotions from other mental states. In this respect, what emotions seem invariably to involve is a form of *evaluative structure* or *evaluative relationship* with their intentional content: They present an object, event, or situation under a value-laden guise. How this evaluative relational structure is construed may vary across cases: Some emotions manifest primarily through physiological or enactive responses with minimal cognitive mediation; some are grounded in conceptual representations, while others in non-conceptual ones; some involve propositional contents, some others are propositional but non-doxastic, etc.

On such a comprehensive view, these different dimensions of affectivity are not, *a priori*, *discrete* and *mutually exclusive*. Instead, their balance *dynamically* changes according to the *degree* of *emotional complexity*. Simpler emotions tend to be less cognitively demanding and more physiological, while complex emotions (e.g., social, aesthetic, epistemic, etc.) typically require propositional, if not doxastic, articulations. The emotions of infants and animals remain genuine emotions, though simple ones, even in the absence of sophisticated conceptual or linguistic capacities[118]. In recalcitrant emotions, non-cognitive components of affect,

[118] And yet, as previously highlighted, much depends on the notion of *cognition* one adopts. If cognition is understood in representational and linguistic terms, then infants and non-human animals are unlikely to be capable of complex emotional states. By contrast, if one also accepts the possibility of non-representational, pre-linguistic and pre-conceptual cognitive faculties (which has been indeed even proved by empirical findings in infant psychology (see e.g., [Cesana-Arlotti et al.

such as those related to physiology and phenomenology, may outweigh the still present cognitive components, etc.

At the same time, the structure of emotions, and the way affective components are balanced in an emotional occurrence are also shaped by social norms, cultural frameworks and personal habits [Section 3.3]. What counts as salient, appropriate, or even intelligible in an emotional response thus varies across communities and individuals, further modulating how different elements of affectivity emerge in particular circumstances. Whether an evaluative relation takes, say, a conceptual or non–conceptual form is not determined solely by the intrinsic architecture of cognition, but is also implicitly mediated by shared frameworks of meanings, that establish which evaluative patterns are expected to be explicitly conceptualized and linguistically articulated, and which instead remain tacit, embodied, or enacted. At the same time, individual histories and habits sediment selective dispositions that favor either propositional articulation or immediate, pre-conceptual appraisal, thereby shaping the mode through which evaluative structures are instantiated in emotional life.

Taken together, these considerations suggest a tentative stance that I shall call *Dynamic Evaluative Pluralism about Emotions* (DEPE) —see also [Section 7.2.1]. As a position in the philosophical debate on affectivity, it would certainly require a more extensive and rigorous defense than the one that has been possible to provide here. Yet, I regard it as one of the few viable compromise perspectives that can account for the diversity, complexity and (*anomalous*) heterogeneity of emotions within a single dynamic and comprehensive framework, which is neither reductionist nor over-intellectualizing, and which further recognizes —as often overlooked in pure epistemology and philosophy of mind— the decisive role of cultural, social, historical and personal factors in shaping the structure of emotions as mental states[119].

2018; Bohus et al. 2023]) and animal ethology (see e.g., [Ferrigno et al. 2021]), then it may well be that infants and animals also possess a complex emotional life that goes beyond mere motivations or purely physiological and behavioral reactions.

[119] Even if the present exposition of DEPE is deliberately informal and provisional, it can be articulated in a more rigorous and falsifiable form. The theory may be framed by relying on two definitional criteria for emotions.

7.2.3 Emotions and Propositions

A view such as DEPE [Section 7.2.2.3] may accommodate the full heterogeneity of emotional life, showing that the balance of affective components and the articulation of the evaluative structures at the basis of emotional occurrences adapt to the degree of emotional complexity. As emotions become more complex [Problem 4], in fact, their evaluative structures tend to demand more explicit forms of articulation and normative assessment. For the purposes of this chapter, therefore, I now concentrate on the propositional dimension of emotions, as it provides a particularly fruitful lens through which such articulations and assessments can be examined, especially in the case of higher-order emotions in social contexts.

(1) *Evaluation*: An emotion is always an affective episode that presents an object or event under a certain evaluative aspect —however minimal or non-representational— and that implies a distinctive phenomenological and motivational disposition.

(2) *Grades and Dynamics*: The way in which such evaluation takes shape — whether as a bodily reaction, a conceptual mental state, a propositional or even doxastic attitude— varies in degree with the complexity of the emotion, such that more complex emotions typically involve increasingly articulated representational components, whereas simpler emotions may rely almost entirely on attitudinal and bodily patterns.

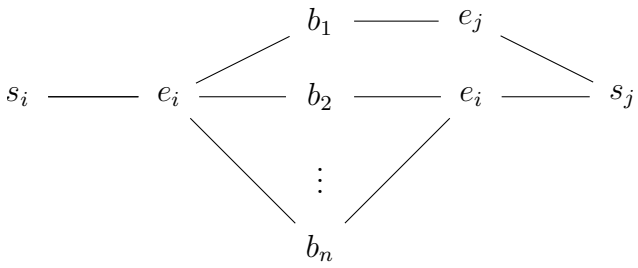
On the basis of these criteria, the theory would be falsified if (1) instances of emotions were found that do not derive from any form of evaluation, or (2) if simple emotions were shown to involve representations, propositions, concepts, or doxastic states, or conversely, if complex emotions were found that do not involve them. In this form, however, DEPE has not been empirically falsified yet, finding instead multiple lines of corroboration in contemporary psychological and neuroscientific research. For example, studies on conditioning and subliminal threat perception show that subcortical systems can trigger fear in the absence of conscious representation, indicating that simple emotions may manifest as purely attitudinal patterns (see e.g., [Gelder et al. 1999; Öhman et al. 2007; LeDoux 2014; Celeghin et al. 2015]). By contrast, higher-order emotions such as guilt or shame depend on complex representational resources, as confirmed by imaging and lesion studies (see e.g., [Tangney et al. 2007; Takahashi et al. 2008; Mendez and Shapira 2009]). Neuropsychological dissociations further reveal that affective dispositions and cognitive representations can come apart (see e.g., [Hadj-Bouziane et al. 2013; Fiacconi et al. 2014; Okon-Singer et al. 2015]). Finally, large-scale imaging demonstrates that emotions rely on distributed and partly distinct neural networks, a plurality consistent with DEPE framework (see e.g., [Lindquist et al. 2012; Pessoa 2017; Geng et al. 2024]).

7.2.3.1 Particular and Formal Content of Emotions

As previously mentioned [Section 7.2.2.2], most emotional occurrences have an intentional content. Generally, if s_i has a certain emotion e_i towards b, such b is said to be the intentional content of e_i.

In characterizing emotional intentionality, often a distinction is made between the *particular content* and the *formal content* (or *object*) of an emotion (see e.g., [Kenny 1963; Teroni 2007; Scarantino and Sousa 2021]). In "s_i has e_i towards b", for example, b is the *particular content* of the emotion, the *individual* object or event towards which s_i's emotional state e_i is directed. So, in the case of George, the particular death threat —say, that one he received yesterday evening in a phone call from a notable criminal of the town— is the particular object of his fear.

There may be the case, however, that George fears not only the death threat, but also Anna's pit bull, the number "4" (George is a very superstitious Chinese), and the little squirrel that sometimes strolls on his garden. Here, *different* objects or events b_n function as the particular content of the *same* emotional state e_i of s_i. Nevertheless, the *same* objects or events may possibly constitute the intentionality of a *different* emotional state[120] e_j for a *different* subject s_j (or for s_i at different times and in different contexts):



Anna feels joyful when she sees her dog, has nothing against the number "4" (she is a very rationalist American), was not bitten by a squirrel when she was a child, but is scared of clowns, which George instead finds funny. This means, in other words, that the particular content b of affect —which may vary across several instances of the same emotion, or be the intentional object of different emotions— can act as a *principle of individuation* only across several instances of the same emotion, but not for different emotional states $e_i, e_j \in E$ (see e.g., [Deonna and Teroni

[120] Or even of another non-affective mental state (belief, desire, etc.).

2015]) —the pit bull and the squirrel individuate two different instances of George's fear, but not George's fear *and* Anna's joy.

Rather, the principle of individuation for individual emotions is the *formal object* of affect, generally understood as the *property* ϕ of b in virtue of which s_i has e_i toward b. What makes George's emotional states towards the death threat, the pit bull, the "4" and the squirrel all instances of fear is the fact that all of them share the same formal content, that is, a property like "*dangerousness*" that George somehow sees in those particular objects and in virtue of which he fears them. Conversely, towards the same particular object —a clown— George and Anne have two different emotions because their emotional states have different formal content: They namely see two different properties in the clown —"*funniness*" and "*frightfulness*"— typically associated with two different emotions —fear and amusement.

7.2.3.2 Evaluative Properties and Emotional Attitudes

Properties like ϕ are called *evaluative*, and the different evaluative properties generally relate to the different emotions (see e.g., [Deonna and Teroni 2015]) —whereby, within a DEPE-based framework [Section 7.2.2.3], emotions of course remain not reducible just to the (representational) act of attribution of evaluative properties to particular objects, and are *additionally* individuated by integrated patterns of physiological responses, behaviors, motivations, phenomenological states, etc.

Against reductionist approaches such as judgementalism (E *is* the judgement that $<b$ *is* $\phi>$) [Solomon 1976; Nussbaum 2001] or perceptualism (E *is* the perception that $<b$ *is* $\phi>$) [Tappolet 2000; Johnston 2001; Deonna and Teroni 2008], and more in line with quasi-judgementalism [Roberts 2001; Solomon 2003] and quasi-perceptualism [Goldie 2000; Zagzebski 2003; Döring 2007], this view partly aligns with Deonna and Teroni's [2015; 2024] understanding of *emotional attitudes*: Alongside judging, believing or perceiving, fearing, rejoicing, being ashamed, etc. are distinct, *sui generis*, attitudes having a specific phenomenology (patterns of felt bodily arousal and felt readiness to act) with an *individual* evaluative structure —i.e., with each particular emotional attitude being correlated to a particular evaluative property.

Attitudinalist approaches move the evaluative property from the object of the emotion to the attitude itself, so that no representation is required. On this view, E is an emotional attitude already linked to ϕ, and to have E towards b is not to represent b as ϕ (in the form of a judgement or perception), but simply to take up that attitude towards that object. In this way, it is argued, it is possible to have different emotions towards the same object: Simply, it is all about a change in the emotional attitude.

Hereby, however, a problem remains in explaining how exactly the property ϕ applies to b (see e.g., [Dokic and Lemaire 2015; Müller 2017; Bonard 2024; Kauppinen 2024; Grzankowski 2025]): If the evaluative property belongs to the attitude rather than to the content, it is unclear why, in experience, it is nevertheless the object b that appears as ϕ. Nor moving the evaluative property from the object to the attitude excludes from itself that, at least in some cases, emotional attitudes may involve representations: Having attitude e_i and e_j towards b may still involve some kind of representation of b as ϕ_i and ϕ_j, which does not indeed collide with the possibility of different emotions towards the same object, and still remains consistent with the attitudinalist elimination of the particular object of the emotion as a principle of individuation for different emotional attitudes. Evaluative representations may be not sufficient for E to be an emotional state, but may be still involved in it, along patterns of bodily arousal and felt readiness to act, as characteristic of that emotional attitude.

7.2.3.3 Formal Objects and Correctness Conditions

Moreover, consider also that formal objects of emotion —evaluative properties ϕ directly inherent to emotional attitudes— are often taken by attitudinalists still as the *correctness condition* of emotions [Deonna and Teroni 2015; Scarantino and Sousa 2021]. S's E towards b is *correct* only if b actually instantiates ϕ. Thus, for instance, George's fear of the death threat would be inappropriate if the threat were issued, say, by Bugs Bunny —that is, if it were not actually dangerous. This means, however, that attitudinalism cannot entirely dispense with the idea that evaluative properties must, at some level, belong to the ob-

ject (see e.g., [Rossi and Tappolet 2019; Ballard 2020; Bonard 2024; Kauppinen 2024; Grzankowski 2025]). For, if the appropriateness of fear depends on whether the object is actually dangerous, or the appropriateness of anger on whether the object is actually offensive, then evaluative properties appear once more as somehow predicated of the intentional object. The risk, then, is that attitudinalism reintroduces through the back door the very representational commitments it sought to throw out from the window: In order to make sense of correctness conditions, evaluative properties, at a certain point, must be somehow ascribed to objects, not only built into attitudes. In this way, the account seems caught in a dilemma: Either evaluative properties are internal to the attitude, in which case correctness conditions remain unexplained, or they are properties of the object, in which case the view collapses back into a representationalist picture.

Hence, apart from the theoretical advantage of non-reductionism — emotions are not reducible to evaluations and have a *sui generis* attitudinal character marked by heterogeneous patterns of bodily arousal and felt readiness to act— it seems that such internal limitations make attitudinalism a rather untenable position for a general theory of emotions, and that, consequently, the role of representations in emotions has to be reconsidered.

7.2.3.4 Evaluative Representations

Let us return to George's fears. One day George meets Anna's pit bull at her home. The dog welcomes him by showing its teeth while barking and growling. This, of course, provokes George to fear the dog when suddenly confronted with it. In this case, it seems plain to assume that George's fear may be explained primarily in attitudinal terms: His body is aroused, his muscles tense, and he is prepared to flee, etc. Such phenomenology is evaluatively structured toward danger in a manner that may well be even biologically and evolutionarily hardwired, as the product of autonomic adaptive mechanisms enabling rapid responses to threats without the mediation of explicit representational processes (see [Section 3.2.1] and [Footnote 119]).

Consider, however, the case of the death threat. When George receives a phone call from the criminal announcing the intention to kill him, his fear can hardly be adequately explained in purely attitudinal terms. Unlike the immediate reaction to the dog, this episode involves an evaluative structure that unfolds across time, draws on background beliefs about the credibility of the source, and sustains anticipatory reasoning about possible outcomes. Here, in this sense, it seems indispensable to recognize some role to representations: George does not merely assume an attitude of fear, but represents the situation as dangerous. Without such representation, the persistence, intensity, and cognitive depth of the fear would remain inexplicable. One might object that an attitudinalist could still account for the difference between the two instances of fear by pointing to variation in the particular content of the attitude or in the associated bodily and motivational patterns. Yet, this strategy underestimates the degree to which the very occurrence of the emotion depends on how the situation is construed. If George believes the call is genuine, for example, then fear arises. If he takes it to be a prank, it does not. This dependence of the emotion on representation is not incidental but constitutive of it, marking a fundamental limit of the attitudinal view especially in cases of complex or socially mediated emotions.

This is not say, however, that George's fear towards the death treat can be entirely reduced to the representation of the latter as dangerous: The phenomenology of the attitude "fear" still remains present, with all its characteristic patterns of felt bodily arousal and felt action-readiness. It is only that such attitudinal features, while necessary to the experience, are not sufficient on their own to explain the distinctive character of the emotion in this specific complex case, which crucially depends on the representational construal of the situation as genuinely dangerous.

Therefore —especially from a DEPE-oriented perspective (pluralist and degree-based) [Section 7.2.2.3]— emotions are not reducible to representations, but neither can they be explained just in terms of non-representational attitudes. Rather, they are likely to involve representations of varying strength and complexity depending on the particular intentional object considered —which, as illustrated before, does not logically contradict even pure attitudinalism.

Now, once representations are readmitted, however, the task is to clarify what kind of representations are those involved in emotions and how they work. Representation is a long-standing and highly debated topic in philosophy of mind, cognitive science, philosophy of science, linguistics and aesthetics (see e.g., [Brentano 1874; Goodman 1968; Pitt 2022; Shea 2018; Frigg and Nguyen 2021]). One common understanding of the notion of representation stresses its *attributional* nature (see e.g., [Frigg and Nguyen 2017; Burge 2018; Nanay 2020; Quilty-Dunn and Green 2023]): In representations, an object is *presented* to the subject under certain aspects, namely as *endowed with* specific properties.

Attributed properties can be of different kinds, for example, perceptual (shape, color, movement, etc.) or structural (weight, roughness, etc.). To a certain extent[121], properties of this kind are often regarded as *intrinsic*, that is, grounded in the object itself and not dependent on its relations with the environment and the (mind of the) representer. By contrast, *relational properties* —e.g., "being next" (to another item), "being owned" (by someone), "being more valuable" (than another item), etc.— are *extrinsic* properties that do depend on the way the object is placed within a broader network of relations, including those with representing subjects[122] (see e.g., [Lewis 1983; Meinertsen 2024]).

Attributable properties also include *evaluative* ones, such as ϕ (dangerous, offensive, shameful, admirable, etc.). These are often regarded as a specific subclass of relational properties, namely *response-dependent* properties (see e.g., [Roseman 1996; Dokic and Lemaire 2015; Tappolet 2016; Müller 2020]). The latter are properties that objects possess in virtue of the characteristic responses they elicit in subjects interacting with them under given conditions (see e.g., [Yates 2008; Kauppinen 2022]). For instance, consider the property of being tasty. An apple may be said to be tasty insofar as it typically elicits a pleasurable gustatory response in normal subjects under standard conditions. Likewise, some-

[121] Of course, the "objective" (understood as subject- and relation-independent) status of such properties has been questioned in philosophy (see e.g., [Marshall and Weatherson 2023]).

[122] Crucially, the two distinctions —intrinsic and extrinsic, relational and non-relational— do not coincide: Some relational properties are namely intrinsic (e.g., having a hydrogen atom as a part), while some extrinsic properties are non-relational (e.g., being the only object in the universe).

thing is frightening if, under ordinary circumstances, it tends to provoke fear-like reactions in those who encounter it. In both cases, even if possibly grounded in the structure of the object, the evaluative property does not reduces to intrinsic feature of it, consisting rather in the patterned responses it gives rise to in suitably placed observers within a specific context.

Evaluative properties ϕ, thus, do not ontologically inhere in the object b, but are attributed to —that is, represented of— it by a subject S who happens to encounter it. In this sense, the attribution of evaluative properties is context-dependent, since it takes place within broader symbolic systems (see also [Goodman 1968]) that shape how objects are construed and appraised. What counts as admirable, offensive, or amusing is never fixed once and for all, but varies across cultural frameworks, linguistic practices, and individual perspectives.

Of course, this carries strong implications for a theory of emotions. If *aspectual representations* of —i.e., attributions of evaluative properties to— b are context- and subject-dependent, this means that the emotional states such representations ground are themselves variable and relative, shaped by the symbolic and contextual conditions under which the evaluative property is attributed. Emotions, in other words, cannot be accounted for as mere reactions to intrinsic features of objects, but must be understood as responses mediated by representational frameworks that determine under which aspects those objects are apprehended —a conclusion which seems to be consistent with empirical findings about emotions in psychology, anthropology, sociology and history [Section 3.3].

Second, this means that criteria of correctness and appropriateness for emotions cannot be defined in absolute or purely objective terms. For example, if George perceives a little innocent squirrel as dangerous and experiences fear accordingly, his emotion should not be deemed as inappropriate based on the ontological assumption of harmlessness in the squirrel. Rather, it has to be considered as adequate and correct within the symbolic and contextual framework in which George attributes that evaluative property to the squirrel. In other domains of representation —see especially the case of science (see e.g., [Elgin 2017])— it is increasingly acknowledged that standards of adequacy need not be conceived as resting on a strict correspondence between a representation and the

object it represents, but rather on the fittingness and consistency of the representation within a broader system of practices and commitments. So, too, the appropriateness conditions of emotions are best understood as determined by their place within a broader contextual framework, rather than by appeal to fixed, "objective" ontological features of the object —a point already noted in psychology and meta-ethics (though less so in philosophical emotion studies, see e.g., [D'Arms and Jacobson 2000; Warner and Shields 2009]), with significant implications for applied ethics, including the ethics of affective AI and social robotics.

7.2.3.5 Propositionality

Apart from such interesting discussions —which we cannot however pursue further here— it is worth stressing that the evaluative representations involved in emotions may also assume a propositional form. In general terms, S having a certain emotion E towards b involves, as we know, S's evaluation of b as ϕ, whereby such evaluation can take the propositional form $<p>$ of

$$< b \text{ is } \phi > \tag{7.2}$$

In this respect, S has $<p>$ as a part of the intentional content of her E. So, for example, George's fear toward the death threat involves $<$*The death threat is dangerous*$>$ —proposition containing the evaluation of the death threat as as dangerous, on which the fear rests— as a component (or dimension) of its intentional content.

Crucially, this is not to say that having an emotion corresponds to having a (not phenomenologically specific) propositional attitude among having a belief, a desire, etc. —a view that, as pointed out above, is problematic in many ways; nor that, like in judgementalism (all kinds of) emotions can be entirely reduced to propositions. Rather, the point is that propositionality should here be understood in a *weak* and *aspectual* sense (see also [Mitchell 2019]): Emotions present their intentional objects under evaluative aspects that can be articulated in propositional form, without thereby being constituted by propositional attitudes. In this way —within a DEPE-driven framework [Section 7.2.2.3]— propositionality is not the defining essence of emotions, but one possible *dimen-*

sion or *mode* of their evaluative structure, whose degree of salience varies with the complexity of the affective occurrence —minimal or absent in basic, embodied reactions, and more explicit in complex and socially mediated emotions, where the evaluative structure depends on background beliefs, unfolds over time, and requires articulation in intersubjective contexts of justification, communication, and normative assessment[123].

Now, involving propositional components, emotions are integrated within wider networks of cognitive attitudes. In particular, they can be implicitly or explicitly related to, sustained by, or even depend on, background beliefs, judgments, items of knowledge and other (intrinsically propositional [Section 7.2.2.1]) doxastic states the subject may derive from the context or has already stored in the memory[124]. Thus, while an emotion E toward b with formal object ϕ cannot be reduced to the judgment $<b$ *is* $\phi>$, its intentional content may nonetheless depend upon, or be modulated by, implicit (inferential) connections to further propositions contained in the subject's cognitive system.

If S's emotion E about b involves the proposition $<b$ *is* $\phi>$ as a component (or mode of presentation) of its intentional content, such a proposition is often not detached from other cognitive commitments, but resulting from inferential processes relying on S's background knowledge and beliefs about b, ϕ, the relation between the two, etc. In this sense, $<b$ *is* $\phi>$ may function as the propositional articulation of an evaluative con-

[123]The introduction of propositionality as a dimension of the intentional content of affects entails that emotions may be assessed in terms of correctness or appropriateness according to the *truth* of the propositions they involve (see [Mitchell 2019]): S's E about b is appropriate only if $<b$ *is* $\phi>$ is true; Georges's fear toward the death threat is correct only if the corresponding fact holds. Yet, as argued above [Section 7.2.3.4], also in case of propositionality, such correctness conditions cannot be conceived as absolute: They remain mediated by broader symbolic and contextual frameworks that determine which evaluative construals count as justified or fitting in a given setting. Propositionality therefore functions as a criterion of adequacy, but one whose normative force is always situated and context-dependent.

[124]More specifically, in her *declarative* memory, which is divided into episodic memory (the memory one has of specific events or object in the own personal life) and semantic memory (the memory one has about general states of affairs and concepts, formulated as propositional knowledge) —see e.g., [Michaelian and Sutton 2017]. Beliefs and items of knowledge stored in memory arguably constitute a significant part of a subject's cognitive identity and *personality* (see also [Section 1.2.2]), shaping the patterns of evaluation through which she apprehends the world and thereby modulating the specificity of her emotional responses in comparison to those of other individuals (see e.g., [Schechtman 2011; Klein and Nichols 2012]).

strual that S can draw from already available cognitive resources. Thus, for instance, it may emerge as the conclusion of a reasoning (explicit or implicit) of the following kind (for $b \in B$):

$$\forall x(B(x) \rightarrow \phi(x)) \tag{7.3}$$

$$B(b) \tag{7.4}$$

$$\therefore\ \phi(b) \tag{7.5}$$

whereby, thus, S having E towards b (also) consists in S's believing (7.5) relying on a more general background knowledge (7.3) of propositional form. It seems in fact difficult to think that —at least for complex emotional states— S can have emotions E towards b without having previous structured knowledge about b (whereby knowledge can be but propositional). It seems difficult to maintain that George could fear the death threat without possessing knowledge about what death threats are, about their (maybe socio-culturally mediated) dangerousness, as well as about the person who issued it, the context in which it was delivered, etc. His evaluation of the situation as dangerous namely results from an implicit reasoning drawing upon his background knowledge —stored in his declarative memory —concerning, for example, death threats, dangerousness, and the relation between the two[125].

[125]Several possible objections may be raised against the idea that emotions involve a propositional dimension. Critics might argue, for example, that this view (i) risks generating an infinite normative regress, insofar as each propositional articulation would itself require further higher-order propositional articulations; (ii) downplays the immediacy of phenomenological feeling; (iii) cannot account for animal and infant emotions, which evidently occur without propositional or conceptual capacities; (iv) assumes a semantic precision that affective contents often lack; or (v) underestimates the pragmatic and social functions of emotions by giving priority to their propositional structure. Yet, these concerns are already addressed once (1) the weak and graded sense of propositionality at stake here is taken into account [Section 7.2.3] —namely, as a form of evaluative articulation that emerges only at certain levels of complexity; and (2) once this view is situated within the broader framework of DEPE [Section 7.2.2.3], in which propositionality is neither necessary nor exhaustive, but rather one possible mode through which, especially in complex social emotions, affective life becomes structured and socially intelligible.

7.3 Rethinking the Epistemology of Robotic Emotions

7.3.1 The Model $\mathcal{M}$ in HRI

As previously shown, in both robotics [Chapter 6] and HRI [Section 7.1.2.1], current research on artificial emotions remains essentially constrained by a physiological-behavioral conception of affectivity. In robotics, this translates into the design of systems that seek to reproduce, or recognize, affective states by means of observable physiological and behavioral patterns, while, in HRI, it results in experimental methodologies that assess human-robot emotional attribution almost exclusively on the basis of such external cues.

Attributing an emotion E to a target T can be conceived as a process in which a subject S makes sense of characteristics $C = \{c_1, c_2, \ldots, c_n\}$ of T through a model $\mathcal{M}$ that specifies the conditions under which those C's count as constitutive of E [Section 7.2.1]. When the target is a robot, in other words, HRI implicitly operates mostly by presupposing that emotional attributions between humans and robots are mediated by a kind of $\mathcal{M}$ that takes only physiological and behavioral C's as constitutive of an emotion [Section 7.1], and therefore relevant in emotional attributions.

In the case of human-human interactions, however, the attribution of emotions cannot be adequately explained by restricting $\mathcal{M}$ to physiological and behavioral conditions alone. As argued in [Section 7.2], emotions are intentional states that can involve a cognitive dimension, including relations to representational, conceptual, and propositional contents. They can be directed toward objects under specific evaluative descriptions, which may take propositional form. Consequently, the model $\mathcal{M}$ that governs emotional attributions between humans also includes propositions among the characteristics C that structure an affective episode and are relevant for emotional attribution. A human subject s_i attributes a complex, socially situated emotion E to s_j relying on a model $\mathcal{M}$ that considers both external and internal characteristics C of s_j as relevant for ascribing E to her, that is, evaluative and conceptual components with propositional content related to E, which s_i hypothesizes are a part of the emotional state E of s_j.

It is precisely by integrating these components that human attribution becomes sensitive not merely to what is externally displayed, but also to what is taken to be thought, meant, or valued by the target. When attributing fear to George, Anne considers not only his external condition, but also assumes that he entertains an evaluative orientation toward the death threat, structured in conceptual, propositional or even doxastic form. In fact, it is often the case that Anne can attribute to George (exactly) the fear of the (specific) death threat (i.e., not a fear in general, or of death threats in general) only by asking him to externalize his evaluative stance toward it —something that can occur only in linguistic, conceptual, and propositional form— whereas just looking at his behavior (shared with many other emotions) is insufficient for that precise emotional attribution.

That considered, it would be unwarranted for HRI to continue disregarding the possible role of propositional factors in the attribution of emotions to robots. What needs to be examined is whether, and how, emotional attributions relate to the parallel ascription of evaluative proportional attitudes to a robot, which types of such attitudes exert the strongest influence on the attribution process, and whether their inclusion results in more accurate, consistent, and contextually appropriate attributions than those based exclusively on physiological and behavioral cues.

In other words, the model $\mathcal{M}$ of emotional attribution in HRI should be expanded to include propositional C's as well, treated as experimental *variables* to be systematically tested for their significance in shaping processes of robotic emotion attribution. This has to be performed by suitable experimental designs that can allow, for example, the crossed manipulation of cues —where physiological-behavioral signals and propositional contents are varied independently and systematically combined (e.g., presence/absence of one or both); the control of situational contexts —by testing attribution across scenarios (e.g., danger, cooperation, conflict, support, etc.), to assess the generalizability of propositional influences; the use of multiple measures of attribution —including self-reports, forced-choice tasks, reaction times, and implicit measures, to strengthen the validity of the findings; the systematic variation of propositional content —by introducing different evaluative orientations (e.g.,

threat, injustice, loss, etc.) to determine which exert the strongest impact on attribution; the control of expression modality —by comparing the effects of propositional contents when conveyed verbally, textually, or through symbolic signals, to assess whether the mode of expression modulates attribution; and the implementation of comparative analyses —contrasting robots, virtual agents, and humans, to establish whether the effect of propositionality is specific to humans or general across different types of agents.

Several considerations support the inclusion of propositionality in the study of emotional attribution within HRI. A first reason is of course epistemic: If emotional attributions in human-human interaction involve propositions and attribution of propositional attitudes, a model that excludes them in human-robot contexts risks conceptual inadequacy. A related reason concerns construct validity, since the consideration of propositional factors prevents HRI from relying exclusively on PB-proxies as relevant for emotional attributions, which are often ambiguous. From a predictive perspective, moreover, models that integrate propositionality are expected to capture user responses with greater accuracy, particularly in novel, complex, and socially embedded real-world scenarios, where the absence of controlled conditions and the lack of stereotyped, unambiguous emotional expressions make correct attributions more difficult in the absence of (namely propositional) evaluative and contextual information. Finally, incorporating propositionality also opens the way to a systematic investigation of how culture, memory, social norms, etc. shape emotion attribution processes, since these factors typically operate through evaluative, conceptual and representational mediations that cannot be reduced to physiological or behavioral signals alone.

7.3.2 Propositionality and the Design of Robotic Emotions

As discussed in [Section 7.1.2], the study of emotional attributions in HRI has not only a descriptive aim but also carries direct normative implications for the design of artificial agents. Moreover, as seen in [Section 7.1.1], the relationship between design and the attribution of emotions is essentially one of interdependence, especially when it comes to robots in

social contexts: The expressive and communicative resources embedded in a robot's design shape the kinds of attributions it invites, while the patterns of attribution ascertained in actual human-robot interactions feed back into the criteria by which robots are evaluated and further developed.

In this sense, clarifying the role of propositional components in human-robot emotional attributions is not only a task aimed at better understanding how such attributions actually work. Constituting a bridge from the explanatory level of attribution processes to the prescriptive level of robotic design, it also provides the basis for informing how robots could —or should— (not) be designed to elicit, sustain, and regulate appropriate emotional attributions by humans. In this way, empirical findings about the conditions under which humans attribute specific emotions to robots can be directly translated into design criteria for affective systems, with the goal of fostering less frustrating, as well as more authentic and effective, human-robot social interactions.

Now, it seems plausible that enriching robotic emotional design with a propositional component could prove advantageous, in particular, insofar as it offers a way to overcome the problems sketched in [Section 6.2] —PB-underdeterminacy, interpretational variability, appropriateness and misattribution, complex emotional states, anthropomorphism and uncanny— which social and affective robotics faces if constrained by a PBR-based design framework. In what follows, I will reconsider these problems from the theoretical and philosophical standpoint outlined in this chapter, showing how the inclusion of propositional components in the design of robotic emotions may provide a way to mitigate or resolve them.

7.3.2.1 PB-Underdeterminacy

Problem 1. *The same observable behaviors and physiological signals expressed by the robot can correspond to different emotions.*

The difficulty posed by PB-underdeterminacy lies in the fact that the same set of physiological or behavioral cues —such as agitation, raised

voice, or increased motor activity— can plausibly be interpreted as expressing very different emotions. Fear, anger, and excitement, for instance, often share overlapping external manifestations, making it impossible to determine with certainty which emotion is being expressed if one restricts the analysis to such cues alone. This ambiguity undermines the reliability of emotional attribution in HRI, since nothing in the surface behavior of the robot guarantees the correctness of the observer's inference.

The introduction of propositional components provides a way to address this ambiguity. By linking an emotion to an explicit evaluative stance, the attribution becomes anchored to intentional content rather than floating among possible interpretations of the same behavioral display. Consider, for example, a domestic robot that, during a sudden power outage, begins to move erratically and emit alarm sounds. Without further cues, one observer may interpret this as fear, another as malfunction, and another as frustration. If, however, the robot simultaneously conveys in some form the evaluation $<$*This situation is threatening*$>$, the attribution of fear becomes not only more plausible but also more stable across observers. Conversely, if the same outward agitation is coupled with the stance $<$*This situation is unjust*$>$, then anger becomes the more appropriate attribution.

Crucially, the propositional layer does not need to be expressed verbally alone. It can be rendered through speech, written messages, symbolic signals, or multimodal combinations designed to make the evaluative content transparent. What matters is that the observer has access to an intentional cue that specifies the evaluative orientation of the robot's state.

In this sense, propositionality narrows the range of admissible interpretations: It provides the observer with a principled criterion for distinguishing among emotions that would otherwise remain indistinguishable at the level of physiology and behavior. The ambiguity inherent in PB-signals is thus reduced not by eliminating those signals, but by supplementing them with intentional content that clarifies their evaluative meaning. As a result, the process of emotional attribution in HRI gains both in precision and in conceptual adequacy, since it reflects more closely the structure of human emotional life, where external displays are

typically interpreted in light of the beliefs and evaluations they are taken to express.

7.3.2.2 Interpretational Variability

Problem 2. *Different observers may infer different, and potentially conflicting, emotions from the same observable cues expressed by an affective robot.*

The problem of interpretational variability is closely related to PB-underdeterminacy. Whereas underdeterminacy highlights the difficulty of distinguishing among emotions that share the same external manifestations, interpretational variability emphasizes the fact that different observers may reach divergent, even conflicting, conclusions when confronted with the same set of cues. A sudden withdrawal by a robot, for instance, may lead one person to ascribe fear, another to ascribe shame, and yet another to interpret the behavior as mere fatigue. Because PB-signals lack intrinsic intentional content, they leave open a wide space for subjective projection, and nothing in the observable behavior alone guarantees convergence in interpretation.

The introduction of propositional components offers a way to counter this variability. By associating the behavioral cue with an explicit evaluative stance —such as <*This situation is threatening*> or <*This situation is embarrassing*>— the attribution process is no longer left entirely to the idiosyncratic assumptions of the observer. Instead, the propositional layer provides a common reference point that orients different observers toward a shared interpretation of the emotion. In practice, this means that a robot's withdrawal accompanied by the stance <*This situation is threatening*> will be consistently interpreted as fear across subjects, whereas the same withdrawal linked to <*This situation is embarrassing*> will converge on shame.

From a computational perspective, this implies that the software architecture of the robot must integrate evaluative structures that represent such stances as propositional contents, and link them to expressive behaviors in a systematic way. These structures can then be externalized

through different modalities —verbal utterances, textual messages, symbolic icons, or multimodal combinations— ensuring that the evaluative orientation is accessible to human observers. In this way, propositionality does not merely enrich the expressive repertoire but functions as a stabilizing mechanism, reducing the dispersion of attributions and fostering intersubjective agreement.

Thus, while PB-based models remain vulnerable to interpretational variability because they rely exclusively on ambiguous external cues, the inclusion of propositional components grounds emotional attribution in intentional content. This not only improves the accuracy of individual interpretations but also secures consistency across observers, addressing one of the key weaknesses of purely PB-based approaches in HRI.

7.3.2.3 Appropriateness and Misattribution

Problem 3. *The emotional state the robot is intended to express can be easily misinterpreted.*

The problem of appropriateness and misattribution arises when the emotional state a robot is designed to express is not the one that observers actually infer from its outward cues. Because PB-signals are inherently ambiguous, even carefully engineered displays of affective behavior can lead to systematic mismatches between what the designers intend and what users perceive —especially in noisy real-world scenarios. For instance, a robot programmed to express fear in the face of an imminent shutdown may, through its agitation and vocalizations, be read instead as frustrated, malfunctioning, or even indifferent, depending on the observer's assumptions.

The integration of propositional components addresses this difficulty by anchoring, again, the intended emotion to explicit evaluative content. If the robot, while displaying signs of agitation, also conveys in some form the evaluative stance $<$*This shutdown is dangerous*$>$, then the attribution of fear becomes not only more precise but also more evidently appropriate to the situation. Without such a propositional anchor, the same expressive behavior risks being misinterpreted, leading to a gap

between the emotion intended by design and the emotion actually attributed by the user.

In this way, propositionality enhances the appropriateness of emotional attributions and reduces the likelihood of misattribution. By explicitly tying expressive cues to evaluative beliefs, it ensures that observers interpret the robot's displays in line with the intended design, thereby strengthening the coherence, reliability, and intelligibility of human-robot emotional interaction.

7.3.2.4 Complex Emotional States

Problem 4. *BP-based affective robots struggle with composite, reflexive, or meta-emotional states.*

The problem of complex emotional states points to the difficulty of capturing, within a purely PB-based framework, emotions that are composite, reflexive, or meta-emotional. States such as pride, shame, or even emotions about emotions (for instance, being ashamed of one's fear) resist reduction to simple behavioral or physiological patterns. Their attribution requires reference to evaluative and conceptual components that specify both the intentional object of the emotion and the subject's evaluative stance toward it. Without such components, robotic emotions remain confined to flat, surface-level expressions unable to reflect the layered complexity of human affective life.

A particularly significant aspect of this complexity concerns recalcitrant and conflicting emotions. Human affectivity is often marked by tensions between what we explicitly believe and what we nevertheless feel. A paradigmatic case is the fear of flying: I may sincerely believe that the plane is safe, and yet continue to feel afraid when turbulence occurs. This recalcitrance does not stem from holding contradictory beliefs, but from the interplay between different cognitive and affective layers: Stable propositional evaluations on the one hand, and, on the other, implicit associations, bodily markers, or memory traces that can override those evaluations in lived experience.

Integrating propositionality into robotic design provides a way to model such complexity without requiring robots to maintain logically

inconsistent belief sets. A robot's cognitive architecture could, for example, include a layer of stable propositional evaluations —symbolically structured beliefs about the world— together with other sub-symbolic or associative layers responsible for affective reactivity, bodily simulation, or implicit orientations. Under certain conditions, these latter components could override the evaluative layer, producing behavior that appears emotionally conflicted or recalcitrant, even while the robot's explicit belief remains unchanged.

By linking emotional attribution to this multi-layered architecture, propositionality makes it possible to ascribe to robots richer forms of emotional life. Observers could interpret the robot as experiencing not only basic emotions, but also second-order, reflexive, and even conflicting ones, insofar as its behaviors and propositional stances reflect the dynamic interaction of distinct evaluative and affective processes. In this way, propositionality directly addresses the limitations of PB-based models, which cannot account for the non-unified, internally conflicted character of many human emotions, and it moves robotic affectivity closer to the complexity of human affective life.

7.3.2.5 Anthropomorphism and Uncanny

Problem 5. *Human-like robots whose affectivity is reduced to shallow PB-cues risk generating a gap between appearance and emotional depth, often eliciting uncanny reactions in the observer.*

The problem of anthropomorphism and uncanny may emerge from the structural gap between the external appearance of humanoid robots and the actual depth of their affective repertoire. When a robot's affectivity is confined to shallow PB-cues, its human-like morphology invites observers to expect a degree of emotional complexity that the system cannot genuinely deliver. The result is a dissonance between form and substance, which often manifests in uncanny reactions: The more the robot resembles a human agent, the more disturbing it becomes when its expressive signals reveal themselves as empty, ungrounded, or inconsistent with the intentional depth and contextual awareness normally presupposed in human emotions.

From a propositional standpoint, the way to mitigate this uncanny effect is to enrich robotic affectivity with evaluative content that grounds external displays in intentional meanings. If a robot's emotional expressions are systematically accompanied by propositional stances that specify their evaluative orientation —such as <*This event is unfair*> or <*This situation is threatening*>— the expressive behavior could no longer appear as a hollow simulation but as the manifestation of an intelligible emotional stance. In this way, the observer is not confronted with a mismatch between anthropomorphic appearance and affective shallowness, but rather with a coherent integration of form and content that renders the emotional life of the robot more credible and contextually appropriate.

Crucially, this does not mean that propositional components must be verbalized in explicit speech alone. They may be conveyed through multimodal strategies —iconic, symbolic, or otherwise— so long as they provide transparent access to the evaluative layer underlying the robot's displays. What matters is that the attribution of emotion is stabilized by reference to intentional content, thereby reducing the risk of dissonance and estrangement. In this sense, propositionality does not merely add a cognitive layer to robotic affectivity but functions as a structural mechanism for aligning human expectations with robotic expression. By narrowing the gap between appearance and emotional depth, it mitigates the uncanny and fosters a more authentic, intelligible, and trust-enhancing forms of human-robot interaction.

Summary

This chapter has proposed a reconsideration of robotic emotions beyond the limits of affective physiological-behavioral reductionism. In that, it has outlined a non-reductionist and dynamically plural understanding of robotic emotions that explores the role of propositionality and intentional content in accounting for the evaluative and representational nature of affective states. By examining attribution, intentionality, and the evaluative structure of emotional states, it has argued that robotic emotions can be conceived not merely as observable reactions but as cognitively

and semantically meaningful configurations that shape human-robot interaction. Such a shift reinstates the cognitive and epistemic dimension of affect within artificial systems, paving the way for a more comprehensive philosophy of robotic emotions, one that enables a richer conceptual and operational framework for designing machines capable of context-sensitive, coherent, multifaceted and better interpretable affective behavior.

The exploration of the propositional dimension of affect thus opens new perspectives for the design and assessment of artificial emotions, where expressivity is no longer confined to bodily performance but grounded in shared interpretative contexts. From the same need of finding new solutions to the problem of affective PBR in robotics, in order to enhance affective human-robot interactions, the last [Chapter 8] of this work now turns to the description of an actual robotic prototype, the humanoid EEVAN. Developed in collaboration with an industrial partner, it was conceived as a (material) philosophical platform for analyzing how computational and robotic models of emotions mediate between philosophical and psychological theories of affectivity and concrete implementations. In particular, it has served as tool for exploring the potential of music and sound to face PBR, so to extend and enrich the emotional repertoire of social robots. In the following, I describe the context of the project, as well as its development and testing.

Part III

Application

Chapter 8

The EEVAN Project: Musical Enhancement of Robotic Emotions

In [Part I] and [Part II], I reconstructed the historical and conceptual development of a physiological-behavioral reductionism that still seems to largely characterize research in affective computing, affective and social robotics, robot psychology, and Human-Robot Interaction. In [Chapter 6], I highlighted some problems arising in robotics from conceiving emotions predominantly as reducible to physiological responses and external behavioral patterns. Thus, in [Chapter 7], a possible enrichment of the notion of robotic emotions was proposed on the basis of a propositionalist account of affective states —a theoretical model that may in the future be translated into a computational architecture implementable in social robots and serve as a framework for designing experiments in HRI.

Building on the perspective of extending the scope of the concept of emotion in robotics and HRI beyond PBR, the EEVAN (whimsically, *Enhancing Emotions in Vision-Acoustic Neurorobotics*) project was carried out at Cube Ltd. Bari as part of the PNRR (*Piano Nazionale di Ripresa e Resilienza*, National Recovery and Resilience Plan), a national investment and reform program fostering technological innovation in Italian public administrations that funded this doctoral research.

EEVAN is a humanoid robot based on the open-source platform *InMoov*, developed by the French sculptor and designer Gaël Langevin since

2012. It is grounded in the general idea that complex musical stimuli, endowed with strong symbolic and cultural significance, can be integrated into artificial affective processing, thereby expanding the expressive and relational possibilities of social robots. In particular, EEVAN is designed as an experiment to examine how the integration of musical input with embodied interaction modulates the affective dynamics of human-robot interaction, in contrast to systems involving music without embodiment or embodiment without music.

This case study, besides providing a concrete application of the theoretical reflections elaborated in [Part I] and [Part II], aims to show and exemplify how the dialogue between philosophy and technology can translate into design practices capable of influencing concrete domains of public and institutional life.

In [Section 8.1], the chapter considers the historical and theoretical relationship between emotions and music, emphasizing thus its relevance for affective computing and robotics. [Section 8.2] situates this discussion within the broader field of musical robotics, reviewing some (historical) experimental approaches that have relied on sound, music and embodiment for improving affective interaction. [Section 8.3] is devoted to the EEVAN project, describing its building process [Section 8.3.1], programming architecture [Section 8.3.2], and experimental evaluation with corresponding results [Section 8.3.3]. Finally, [Section 8.4] outlines possible future developments and potential applications of music-based affective processing in social robotics.

8.1 Emotions and Music

Since remote times, music has been closely associated with the sphere of emotions. From ritual chants to sophisticated artistic compositions, it has constantly appeared as one of the most effective means through which human beings express, communicate, and regulate their affective life. More than verbal language, music addresses affectivity directly, shaping moods, evoking memories, and creating bonds among individuals and communities of even distant cultures (see e.g., [Merriam 1964; Sacks 2008; Juslin et al. 2011]).

8.1.1 A Historical Perspective

The nature of music and its profound connection to human emotions have long been subjects of philosophical and scientific inquiry. In Ancient Greece, for example, philosophers like Plato [1997c] and Aristotle [1984b] delved into the relationship between music and emotions, emphasizing its moral and educational significance (see also [Ravasio 2025]). Plato believed that music, through its specific modes, had the power to cultivate virtues such as courage and moderation, thereby shaping both individual character and societal harmony —or, conversely, leading to perdition and discord when improperly employed. Aristotle, while agreeing that music influenced moral development, focused on its ability to imitate and evoke genuine emotions in listeners, rather than merely replicate external emotional expressions, thereby stressing too its social role as a regulator of social order and emotional well-being. The educational and moral value of music in arousing (christianly good) emotions and influencing human character was recognized also in the Middle Age, when (Western) music itself began to be theorized and defined as a mathematical science within the *Quadrivium* (see e.g., [Boethius 1990]).

During the Early Modern period, the relationship between music and emotions was increasingly explored both artistically and scientifically. Baroque music, with its rich ornamentations and harmonic structures, gave composers powerful tools for expressing complex emotions (see e.g., [Grout and Palisca 2001; Spitzer 2020]). The relation between specific musical structures, patterns, keys, etc. and emotions was even formalized through the so called *Affektenlehre* (Eng. *Theory of Affects*, see e.g., [Glüxam 2020] and [Section 5.1.1]) —the key of C major is cheerful and bold [Charpentier 1971]; a large interval expresses joy, while a small interval expresses sadness [Mattheson 1739]; the "diabolus in musica" (augmented fourth interval) has to be avoided [Fux 1725], etc. Composers could thus employ the different elements of musical syntax as a true *affective language*, stirring particular emotions in the listener by employing appropriate modulations, chords and tempos. This theory found its scientific underpinnings also in the research on music carried out during this time, which sought mechanical explanations for how mu-

sic influences emotional experiences (see e.g., [Descartes 1649; 1650], and also [Cohen 1984]).

With the rise of Romanticism during the 19th century, the relationship between music and emotions became even more pronounced, with composers using music as a powerful vehicle for personal expression (see e.g., [Cook and Dibben 2013; Thormählen 2020]). Expressed affects were no longer the socially and culturally standardized ones of Baroque *Affektenlehre*, but now understood as reflections of the most deep dimensions of the composer's subjectivity. Philosophers such as Schopenhauer [1819] and Nietzsche [1889] saw music as the purest expression of human emotion, as it bypassed rational thought and directly communicated the inner experiences of the individual. On the contrary, musical formalists such as Hanslick [1854] denied any substantial connection between music and emotions, claiming that the former finds its meaning in its syntactic form, whereas the latter are merely subject-dependent byproducts of it. Parallelly, scientific approaches rooted in mechanism and physicalism addressed music, analyzing the acoustic properties of sound, the physiology of hearing, and the psychology of perception in relation with affective arousal (see e.g., [Helmholtz 1862; Stumpf 1883]).

8.1.2 Insights from Science and Philosophy

Since the 20th century, the affective potential of music started to become the object of systematic investigations in psychology, neuroscience, musicology, anthropology, philosophy and many other related disciplines. Moving beyond old purely speculative accounts, more empirically oriented studies were now trying to shed light on the actual psychological mechanisms through which music can evoke, express, or regulate emotions (see e.g., [Seashore 1938]). Research informed by the emerging fields of AI and cognitive science, on the other hand, already aimed to explain musical emotions as arising from processes of expectation, tension, and resolution (see e.g., [Meyer 1956; 1957; 1967]).

At the same time, the development of affective neuroscience was providing new tools for investigating the biological foundations of musical emotions. Connections between music listening and the activation of brain regions associated with reward, motivation, and emotion were

found (see e.g., [Peretz 1990; Sloboda 1991; Panksepp 1995; Krumhansl 1997; Blood and Zatorre 2001]) —in particular those involving the limbic system and the dopaminergic pathways [Section 3.2.1.6]. Neuroimaging studies further revealed that musical experiences involve both subcortical structures responsible for basic affective reactions and cortical areas engaged in higher-order appraisal, memory, and prediction (see e.g., [Brown et al. 2004; Koelsch 2005; Koelsch et al. 2006]). In particular, this physiological-cognitive research —still ongoing today— began to clarify why music can simultaneously trigger visceral reactions (chills, tears, etc.) and complex reflective states.

Anthropological and ethnological studies also started to emphasize that the emotional significance of music is not fixed, but mediated by cultural, social, historical and personal contexts (see e.g., [Merriam 1964; Nettl 1983; Balkwill and Thompson 1999; Finnegan 2003; Bowling 2013; Susino and Schubert 2019]): Identical musical structures convey different affective meanings across different communities, historical periods and individuals. Some acoustic cues, such as tempo or intensity, may elicit broadly similar reactions across cultures, yet their affective interpretation is largely shaped and transformed by cultural learning and contextual conventions. The emotional significance of musical structures is historically and individually plastic: The same musical intervals or modes may acquire divergent connotations across different epochs, or even among listeners within the same cultural community.

These studies introduced a form of relativism that reignited philosophical debates about the problem of musical meaning. In the wake of Hanslick's [1854] formalism, the question of how music could be expressive at all became central again in the middle of the century. Susanne Langer [1941; 1979], for example —drawing on Cassirer's [1923] philosophy of symbolic forms— argued that music is not a denotative system akin to language. Rather, it functions as a *presentational symbol* that conveys the logical form of feeling ostensively and non-propositionally (see also [Wittgenstein 1922]). Nelson Goodman [1968] developed this semiotic line in a different direction: Music does not denote emotions, but expresses them through metaphorical *exemplification*, whereby formal features serve as symbols of affective qualities within a particular notational system.

Later positions even marked a partial return to formalism. Peter Kivy [1980; 1989; 2002; 2007], for instance, advanced an *enhanced formalism*: While rejecting any semantic, representational, or affective content of music, he maintained that music can still elicit emotions, since expressive properties (e.g., a passage sounding melancholic or joyful) are *audible* qualities instantiated in musical structure alongside purely musical ones (e.g., a chord being major, an interval being diminished, etc.). In contrast, Jerrold Levinson [2011; 2018] defended a *persona theory of musical expressiveness*, according to which listeners experience music as if it were the expressive voice of an imagined persona whose emotions are apprehended through sound.

On the anti-formalist side, Stephen Davies [1994; 2003] has argued that music's meaning lies in its *appearance emotionalism*: Its objective expressiveness consists in presenting the behavioral appearance of emotions through dynamic, gesture-like qualities, without denoting them or depending on the actual feelings of composers, performers, or listeners. Similarly, Jenefer Robinson [1997; 2005] has contended that music is meaningful precisely in virtue of its capacity to express and *arouse* genuine, cognitively complex emotions, which listeners experience empathetically and imaginatively, making emotional engagement a primary key to musical understanding.

8.2 Experiments in Musical (Affective) Robotics

As seen in the previous [Section 8.1], the relationship between music and emotions is deep and multifaceted. Music exerts a significant emotional influence, impacting not only our physical responses but also shaping our thoughts, moods, and memories. Its ability to evoke, regulate, and express emotions transcends language, engaging both psychological and physiological processes. Whether through the direct physiological effect of sound or the cultural and social significance embedded within musical structures, music has a unique power to resonate with human experience, creating connections that unite individuals, cultures, and even historical epochs. It is in fact no coincidence that, from the earliest stages of their

development, automation and robotics have used music as a key medium for enhancing the emotional capabilities of machines.

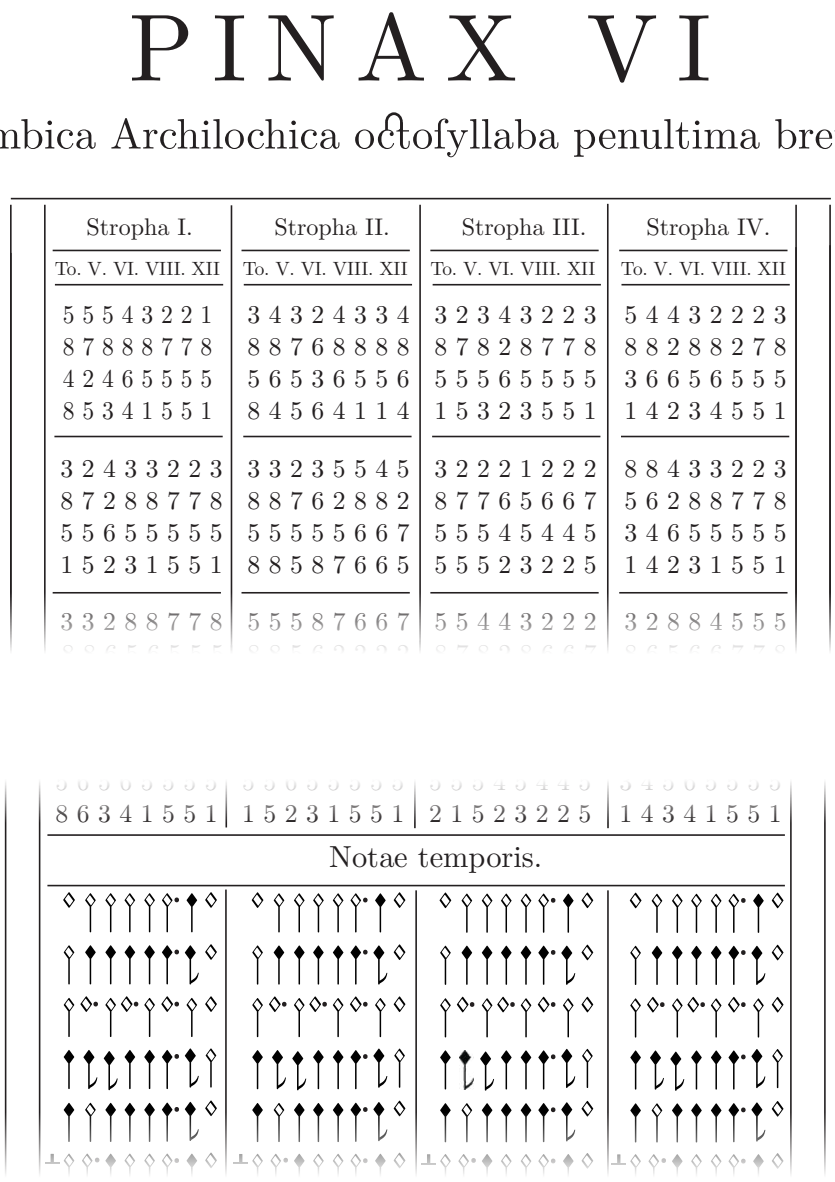


Figure 8.1: Example of a *pinax* from Kircher's *Musarithmica mirifica* —in this case *Pinax VI* for iambic archilochean octosyllables with a short penultimate syllable. The upper section contains numerical figures for the harmonization, while the lower section provides the *notae temporis*, i.e., rhythmic schemes to be applied to the chords —adapted from [Kircher 1650, VIII, 83] (image cropped to fit the page).

8.2.1 From Algorithmic Composition to Robots

During the Baroque era, for example, the German-Roman polymath and Jesuit Athanasius Kircher [1650] invented a system for automatically composing music through an algorithmic process based on combinatorics (see also [Zanzarella 2024, 117ff.]), the so-called *Musarithmica mirifica*. Syllabicated Latin poems could be used as an input. Depending on the meter and the level of complexity —*syntagma*— desired for the composition (simple or florid counterpoint), a *pinax* was chosen —a table containing cells (*musarithmi*) for the harmonization of each verse [Figure 8.1]. Each *musarithmus* encompassed as many columns as the syllables of a verse, and four rows corresponding to the four voices of the choir (Soprano, Alto, Tenor, Bass).

Menſa Tonographica

Qualitas Tonorum	Belli-cofus	Lætus vagus	Lachry-mofus	Hilaris	Amœnus	Plus religiof. fuavis	Triſtis querulus	Vlup-tuofus	Iucundus	Fiducia plenus	Mollis vanus	Magni ficus	Seue-rus va hemēs
Nomina antiqua	Hypo-lydius	Hypo-dorius	Phrygius	Lydius	Dorius	Dorius	Hypo-phrygius	Myxo-lydius	Hypo-myxo lydius	Ionius	Hypo-ionius	Iaftius	Hypo-iaftius
Signatio	Mollis	Mollis	Durus	Mollis	Durus	Durus	Durus	Durus	Mollis	Mollis	Durus	Durus	Mollis
Toni	VI.	II.	III.	V	VIII.	I.	IV.	VII.	VIII.	IX.	X.	XI.	VII.
F	8	7♯	6	5	4	3	2	7♯	4	3	6	4	8
E	7	♭6	5	4	3	2	8	6	3	2	5	3	7
D	6	5	4	3	2	8	7	5	2	8	4	2	6
C	5	4	3	2	8	7♯	6	4	8	7♯	3	8	5
B	4	3	2	8	7	♭6	♭5	3	7	♭6	2	7	4
A	3	2	8	7	6	5	4	2	6	5	8	6	3
G	2	1/8	7♯	6	5	4	3♯	8	5	4	7	5	2
F	1/8	7	6	5	4	3	2	7	4	3	6	4	1

Figure 8.2: The *Mensa Tonographica*, used to convert numerical figures into actual notes. As can be seen, the choice of musical mode is subordinated to a specific *qualitas tonorum*, i.e., the affect intended for the final piece —adapted from [Kircher 1650, VIII, 51].

The harmony was notated by the numbers of figured bass (1-8), recombined through suitable mathematical methods for yielding new *musarithmi* for the different verses and strophes of the text. In the same

way, rhythmic figures were then assigned to the harmonizations through the *notae temporis* table at the bottom of the *pinax* [Figure 8.1]. At this point, numbers could be converted into musical notes by the *mensa tonographica* [Figure 8.2], a table on which each figure was mapped to an actual pitch within a selected church mode, whose choice depended however on the *affective character* one desired for the final piece —a "*hilaris*" (hilarious) character corresponded to the Lydian mode, a "*tristis*" (sad) one to Hypofrigian, etc. This system could be operated through following the steps of the algorithm by anyone, even without previous musical knowledge —a clear advantage for Jesuit missionaries engaged in the evangelization of 17th-century South America— and was even "implemented" by Kircher as a kind of *ante litteram* "computer": The *Arca Musarithmica* was a wooden box that stored in it all *pinaces* (its "long-term memory"), and displayed the *mensa tonographica* (the "program") on its lid [Figure 8.3].

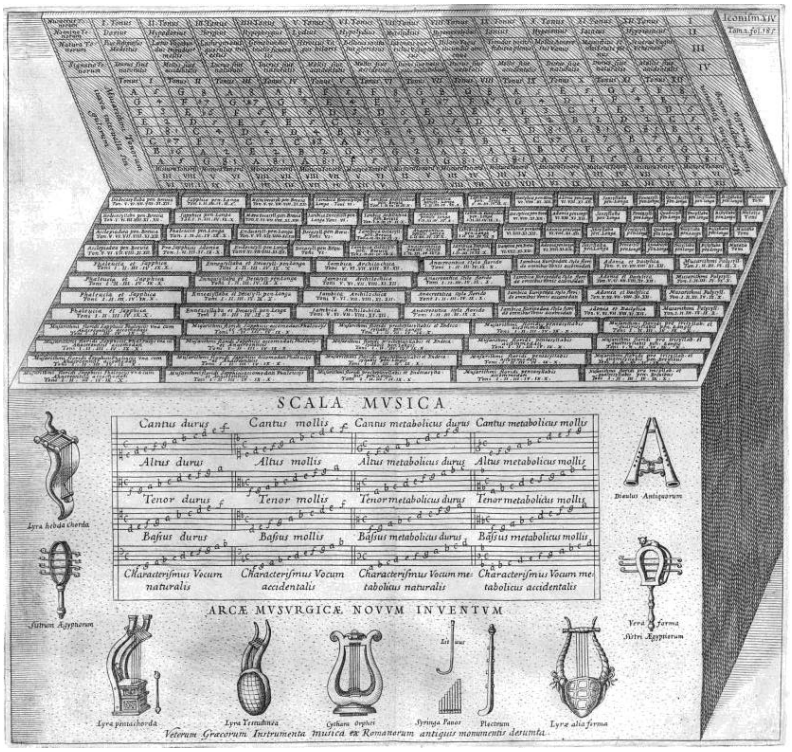


Figure 8.3: Kircher's *Arca Musarithmica* —from [Kircher 1650, VIII, 185-186].

Kircher's *Musarithmica* is (at least to my knowledge) the first algorithmic system to employ emotions as a variable in an automated process, linking them explicitly to the production of affectively meaningful musical outputs. In the same respect, another important historical example of musically enhanced affective automation has been represented, as seen already in [Section 5.1.1], by 18th-century musical automata. Crafted by Vaucanson, the Jaquet-Droz's and other skilled inventors, these early robots combined mechanical performance on real instruments with anthropomorphic gestures, aiming to mimic emotions and elicit affective responses in the public, thereby demonstrating how automation could employ music as a medium for emotional expression.

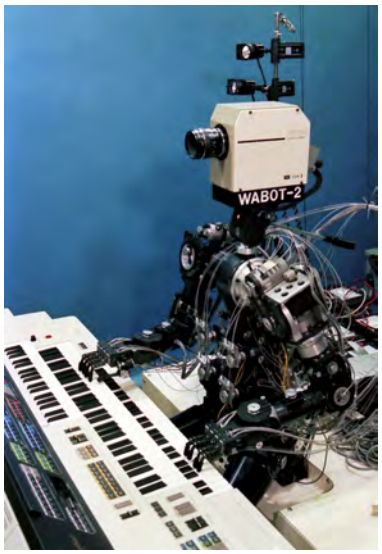


Figure 8.4: The WABOT-2, a musical robot able to play a real organ —from [Kato et al. 1987, 146].

The history of musical automation developed throughout the whole 19th century, with the construction of several prototypes of mechanical instruments for performing, or even composing music automatically [Zanzarella 2024, 137ff.] It is only later, however, that —also thanks to technological improvements in AI [Chapter 1] and robotics [Section 5.2])— a renewed interest arose in the use of music within automation as an explicit means of affective, social, and interactional enrichment. This trajectory emerged most notably in the musical robots developed by Ichiro Kato, Atsuo Takanishi, and their collaborators at Waseda University in Japan between the 1980s and the 2010s (see [Section 5.2.3],

[Footnote 92], [Solis et al. 2006; 2009; Petersen et al. 2010; Solis et al. 2010; Cosentino and Takanishi 2021]). In this context, at least three projects stand out as particularly significant.

The WABOT-2 (*Waseda Robot 2*) [Figure 8.4], completed in the early 1980s, represented the first modern humanoid robot explicitly designed to perform music on a real musical instrument. It integrated optical score-reading, symbolic processing of musical notation, and the actuation of anthropomorphic hands and feet capable of manipulating an electronic organ's manuals, pedals, and expression controls (see also [Kato et al. 1987]). The WF (*Waseda Flutist*) [Figure 8.5, above], developed in successive iterations from the late 1990s, reproduced the biomechanical and physiological principles of human flute performance. Its design employed a pneumatic-mechanical actuation system for controlled airflow into a transverse flute, complemented by artificial lips and tongue mechanisms for articulation and phrasing. Conceptually, this approach explicitly e-

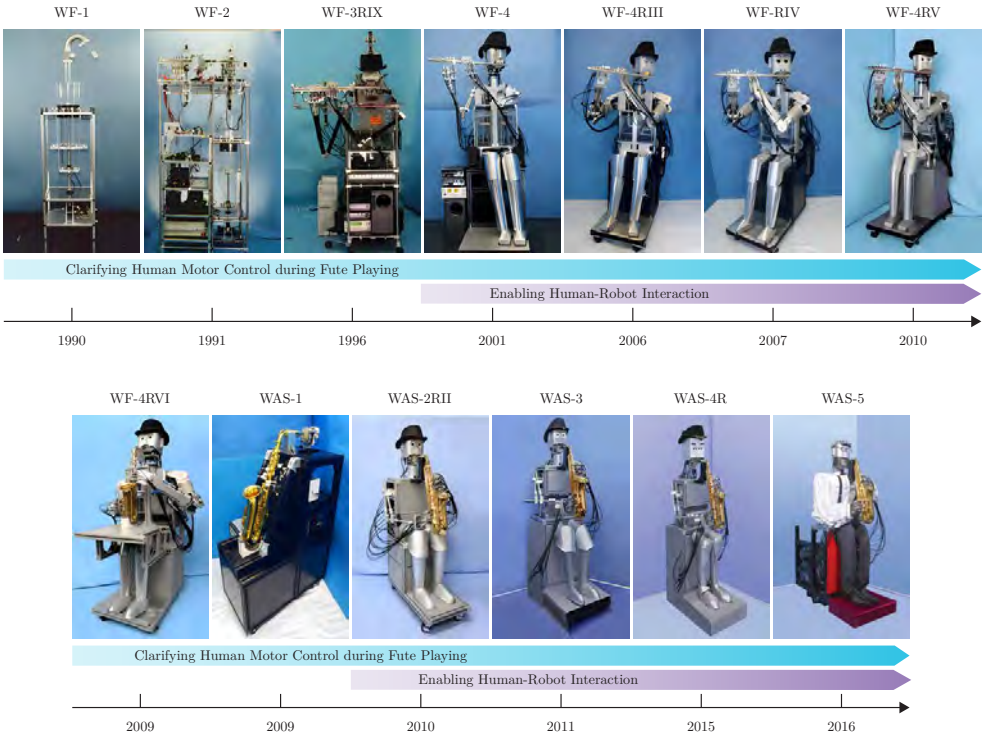


Figure 8.5: Development of Waseda Flutist (WF, above) and Waseda Saxophonist (WAS, below) from the 1990s onwards, with later prototypes explicitly oriented toward human-robot interaction —adapted from [Cosentino and Takanishi 2021, 802-803].

choed Jacques de Vaucanson's eighteenth-century flute-playing automaton, while extending it through modern robotics to enable not only the reproduction of melodies but also real-time ensemble performance with human musicians. Finally, the WAS (*Waseda Saxophonist*) [Figure 8.5, below], initiated in the 2000s, extended the same pneumatic-mechanical and articulatory principles to a single-reed instrument. Its design incorporated artificial lips, tongue, and hand mechanisms optimized for the alto saxophone, enabling precise control of reed vibration, embouchure pressure, and key actuation. Beyond sound production, the project emphasized expressive musical control —such as dynamic variation and phrasing— marking a shift from purely mechanical reproduction toward embodied and interactive musical expressivity.

Following the experiments at Waseda University, private companies too began to develop musical social robots in Japan. At the 2005 Aichi World EXPO, for example, the Toyota Motor Corporation —active in the field of industrial robotics already since the 1970s— unveiled the *Partner Robots*, a series of humanoid robots designed to assist people in factories and homes. Some of the prototypes were adapted to play real string, brass and drum instruments with the purpose of entertaining users (see e.g., [Toyota 2004; Toyota, Motor Corporation 2007; Toyota 2007; Kusuda 2008] and [Figure 8.6]). Or, still, Honda [Section 5.2.3]



Figure 8.6: Two musical robots of the Toyota's *Partner Robots* (the *Toyota Walking*) playing a trumpet and a violin —from [Toyota, Motor Corporation 2007; Toyota 2007].

as well developed musical robots, configuring ASIMO —one of its most famous humanoid platforms (see e.g., [Hirose and Ogawa 2007])— to

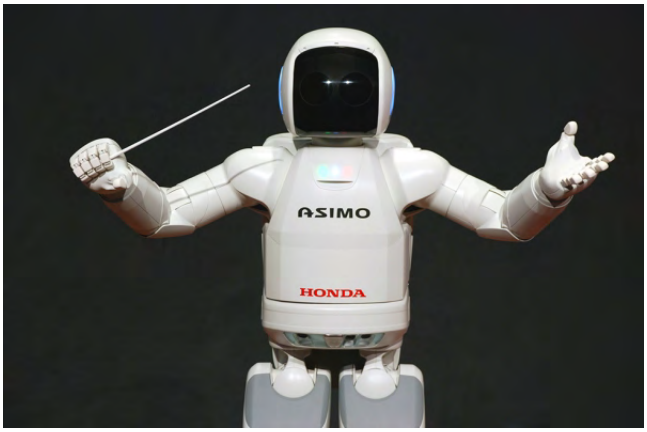


Figure 8.7: Honda's ASIMO conducting an orchestra —from Honda's offical website (link).

conduct live orchestral performances, beginning with a public concert in 2008 where it directed the Detroit Symphony Orchestra [Figure 8.7]. Unlike the Toyota prototypes, which focused on instrumental sound production, ASIMO's musical role relied on its motion control and timing precision, coordinating its arm movements with the musicians. These demonstrations were conceived primarily as a way of showcasing the robot's capabilities in smooth, human-like kinematics and real-time synchronization, thus positioning music as a communicative testbed for evaluating expressive, interactive, and socially oriented aspects of humanoid robotics.

All these developments have stimulated sustained interest in both research and applications within the domains of computational creativity, musical AI and musical robotics, giving rise to a wide range of outcomes that continue to yield striking results. These include advances in music education technologies, systems for automated musical performance, interactive frameworks for collaboration with human musicians, and approaches to music-movement coordination, among others (see e.g., [Savery 2023]), while also raising ethical concerns (see e.g., [Zanzarella 2025b]).

Another research trajectory has instead focused on sound in general, and its capacity to vehiculate, acoustically, emotional meanings. Beyond its (culturally, socially and historically mediated) *organization* into

a structural framework with an aesthetic value —i.e., what we namely call "*music*"— sound always carries expressive properties, modulated above all by its physical characteristics like pitch, intensity, timbre, duration, etc. This becomes particularly evident especially in the (purely) acoustic dimension of language, where *vocal prosody* —that is, the patterned variation of intonation, loudness, tempo, and rhythm in speech— functions as a primary, pre-linguistic, means through which emotions are conveyed and regulated within a social interaction (see e.g., [Fernald et al. 1989], [Table 8.1]).

	Fear	**Anger**	**Sorrow**	**Joy**	**Disgust**	**Surprise**
Speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
Pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
Pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
Intensity	normal	higher	lower	higher	lower	higher
Voice quality	irregular voicing	breathy chest tone	resonant	breathy blaring	grumbled chest tone	
Pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
Articulation	precise	tense	slurring	normal	normal	

Table 8.1: Common prosodic effects on humans of the six basic emotions —from [Breazeal 2001, 1389].

Hence, much robotic research in the last two decades has sought to devise mechanisms for endowing machines with prosodic competence, both in the recognition of affective cues from human speech and in the generation of synthetic vocalizations with appropriate expressive contours (see e.g., [Cahn 1990]). Breazeal's *Kismet* [Section 5.3.2], for example, was among the first robotic platforms to integrate both the analysis and synthesis of prosody within a unified affective architecture. The system coupled low-level detection of acoustic features in human speech —such as pitch contours, loudness dynamics, and temporal variations

— with the generation of synthetic vocalizations produced via a parametrically controlled DECTalk synthesizer. These vocalizations did not convey linguistic content, but rather affective states (e.g., happiness, sadness, anger, surprise), which were tightly coordinated with Kismet's expressive facial animations and bodily gestures. This multimodal integration allowed the robot to engage in proto-conversational exchanges reminiscent of infant-caregiver interactions, where prosodic cues served both as social signals and as reinforcements shaping the robot's behavior (see e.g., [Breazeal 2001; Breazeal and Aryananda 2002; Breazeal and Brooks 2005]).

Breazeal's studies laid the groundwork for later research on artificial affective prosody, enabling the transition from non-linguistic vocalizations to the incorporation of acoustic expressive cues in fully verbal robotic speech —a line of inquiry that still continues to shape contemporary efforts to endow social robots with more natural and emotionally resonant communicative abilities. In this respect, for example, recent studies have shown how text-to-speech systems can be employed in robots as boosters of emotional expressivity, by adapting intonation, rhythm, and loudness to the semantic content of utterances in order to convey affect more convincingly and improve social presence (see e.g., [De et al. 2025; Klüber et al. 2025]). Other studies have instead focused on the complementary problem of recognizing affect in naturalistic spoken interaction with robots, demonstrating that prosodic cues remain robust indicators of user emotion even in acoustically challenging environments (see e.g., [Grágeda et al. 2025]).

8.2.2 Examples from Contemporary Robotics

In continuity with these recent efforts, contemporary research has also merged prosody and music in exploring the affective potential of the acoustic medium in HRI. In this respect, the work of Gil Weinberg and his *Robotic Musicianship Lab* at Georgia Tech stands out as a particularly influential contribution, framing musical interaction as a privileged domain for investigating how robots can engage humans not only cognitively, but also emotionally and socially (see e.g., [Weinberg and Driscoll 2006b; Bretan and Weinberg 2016; Weinberg et al. 2020]).

Weinberg's group has introduced a number of robotic prototypes designed to participate in musical exchange through real-time listening, analysis, and sound generation. Among the earliest of these experiments is *Haile* [Figure 8.8, left], a robotic percussionist created to capture rhythmic patterns from human drummers and respond with improvisations,



Figure 8.8: Gil Weinberg's musical robots: *Haile* (left), *Shimon* (above right), *Shimi* (bottom right) —from Gil Weinberg's personal website (link).

thus turning percussive performance into a reciprocal and socially charged dialogue (see e.g., [Weinberg et al. 2005; Weinberg and Driscoll 2006a]). The project has been later expanded with *Shimon* [Figure 8.8, above right], a marimba-playing robot whose four arms allow for complex polyphonic textures (see e.g., [Weinberg et al. 2009; Hoffman and Weinberg 2010]). Here, algorithmic models of improvisation are combined with bodily cues such as head orientation and gaze, designed to synchronize with fellow musicians and to produce an impression of attentiveness and social involvement. A further step was taken with the creation of a robotic drumming prosthesis, in which one stick is guided by the wearer while the other follows an autonomous control system. This hybrid configuration have opened new forms of shared agency between human and

machine, expanding both the expressive range and the affective intensity of performance. More recently, *Shimi* [Figure 8.8, bottom right] has brought the focus into everyday settings (see e.g., [Savery et al. 2019a,b]): Conceived as a companion robot, it makes use of music analysis not only to choose and recommend songs, but also to animate its movements in dance-like ways (tapping feet, oscillating head, etc.), thereby investigating how musical responsiveness can foster enjoyment, trust, and attachment in domestic or leisure environments.

Taken together, these projects articulate a coherent methodology that integrates computational techniques for pattern recognition and predictive improvisation with the tight coupling of auditory and motor control, as well as the alignment of prosodic variation with musical structure, in order to generate complex forms of musical behavior capable of sustaining engagement and affect in human-robot encounters. As even empirically demonstrated (see e.g., [Hoffman and Weinberg 2011; Savery et al. 2021a,b]), such an approach enables robots to participate in emotionally resonant and socially meaningful interactions, extending their role from automated performers to genuine partners in co-creative exchange.

8.3 The EEVAN Project

In the wake of the experimental tradition outlined in the preceding [Section 8.2], and building on the historical and philosophical considerations of the previous chapters as an attempt to move beyond the constraints of PBR, I here propose EEVAN, a project that further explores the potential of music and sound in HRI and in the design of affective and social robots.

8.3.1 Context

The project EEVAN (*Enhancing Emotions in Vision-Acoustic Neurorobotics*) has been developed at Cube Ltd. under the guidance of Dr. Michele Facecchia, computer engineer and developer, within the Italian subsidization program PNRR.

8.3.1.1 The PNRR

The PNRR (*Piano Nazionale di Ripresa e Resilienza*, in English: *National Recovery and Resilience Plan*) is a national investment and reform program that has been activated in Italy in 2021 with Italian Decree-Law No. 59 of 6 May 2021 [Repubblica Italiana 2021], during the Government of Mario Draghi, with a time frame extending until August 2026. It has been designed within the European Union's *NextGenerationEU* (NGEU), a €750 billion recovery instrument approved by the European Council in July 2020, whose central component is the *Recovery and Resilience Facility* (RRF) [European Union 2021]. The RRF provides financial support to all Member States to stimulate economic recovery after the COVID-19 pandemic and to foster medium- and long-term structural transformation across a broad range of key economic sectors. Italy, in particular, has been allocated €191.5 billion, of which €70 billion (36.5%) in non-repayable grants and €121 billion (63.5%) in loans.

The Italian PNRR has been articulated into six "missions" each addressing a key strategic area of economy and society [Figure 8.9]:

1. *Digitalization, Innovation, Competitiveness, Culture and Tourism*: focused on the digital transformation of public administration and justice, the adoption of cloud and AI technologies and ultra-broadband connectivity, the strengthening of cybersecurity and worker's advanced digital skills, as well as the promotion of innovation in cultural and touristic sectors;

2. *Green Revolution and Ecological Transition*: aimed at supporting decarbonization, renewable energy, energy efficiency and circular economy practices, in line with EU climate neutrality goals;

3. *Infrastructure for Sustainable Mobility*: centered on the modernization of transport networks, particularly high-speed rail and intermodal logistics, to reduce territorial gaps and foster sustainable mobility;

4. *Education and Research*: designed to improve education outcomes, strengthen research and innovation, support doctoral programs in both STEAM and humanities, and foster knowledge transfer to the productive system;

5. *Cohesion and Inclusion*: targeting social and territorial disparities, enhancing labour market participation, and promoting inclusion of disadvantaged groups;

6. *Health*: directed at reinforcing the resilience of the national healthcare system through new infrastructures, digital health solutions, and preventive care.

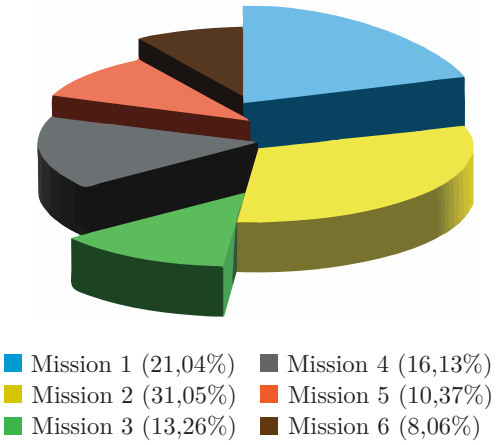


Figure 8.9: Allocation of Italian PNRR funds across the six missions.

Within Mission 4, the Italian Ministry of University and Research (MUR) has introduced extraordinary funding lines for doctoral programs in all academic disciplines, including humanities. In particular, Ministerial Decree No. 351 of 9 April 2022 promoted 2,500 doctoral positions in partnership with enterprises and public administrations, with the aim of reinforcing applied research and fostering innovation-oriented collaboration between industry and academy [Repubblica Italiana 2022].

My doctoral position has been established and funded under this framework. The project has been conceived as a scientific (historical, philosophical and epistemological) investigation into affective robotics. Collaboration with an industrial partner has provided the context and resources to develop a technological prototype, through which the practical implications of this investigation can be explored and tested, and their potential applications in public administration —particularly in education and healthcare— critically assessed for their innovative value, in accordance with the strategic goals of PNRR's *Mission 4*.

8.3.1.2 Cube

Cube is a company based in Bari (Apulia, Southern Italy) operating in the wider domain of digital communication and innovation. While originally established as an agency for integrated communication, it has progressively expanded its activities toward the design and development of technologically enhanced formats, including interactive systems, audiovisual installations, and immersive digital experiences. This trajectory, combining its background in communication with a growing orientation toward technological innovation, has fostered a work environment where creativity and experimentation converge, providing fertile ground for projects that explore the interaction between humans and advanced digital systems.

Within this context, the collaboration for the EEVAN project has been carried out under the supervision of Dr. Michele Facecchia, computer engineer, developer, and head of the *Digital Lab* at Cube, whose professional expertise lies in software development and system integration applied to innovative digital solutions. His guidance has ensured that the project benefited from the company's technological competencies and infrastructure, supporting the adaptation of digital platforms for exploratory research in affective and social robotics.

My collaboration with Cube has taken place in the form of an internship, distributed over three years (2023-2025) and lasting for a total of six months. Each year, between June and September, I have spent two months working on site at the company's headquarters in Bari. This arrangement has allowed me to combine the academic framework of the doctoral program with hands-on experience in an applied environment, ensuring continuity between theoretical research and practical prototyping.

The internship format has facilitated a gradual progression: every summer session has built upon the outcomes of the previous one, enabling incremental refinement of the project. This rhythm of alternating academic study and applied work has created a productive balance, where philosophical and historical investigation has informed design choices, while practical challenges encountered during implementation have fed back into the conceptual development of the thesis. In this respect,

particularly fruitful has been the direct engagement with the building process and the materiality of the robot for developing some of the central research ideas of this thesis, ranging from the uncanny dimension of human-robot interaction and the role of propositionality in modeling affective states to the specific nexus between robotics and music as a medium for affective expression. In this sense, the internship has not only enhanced my technical and practical skills in design thinking, engineering, and programming, but also served as a conceptual laboratory, where theoretical and philosophical insights of the thesis have been tested, reoriented, and expanded through concrete and direct experimentation.

The three stages of the internship have reflected a progressive trajectory in the development of the project. The first year has been dedicated mainly to the hardware configuration of the robot, focusing on the assembly of all mechanical parts, integration of components, and initial calibration. The second year has concentrated on electronics and circuitry, with mobility tests addressing issues of actuation, control, and refinement of embodied movement. The third year has been devoted to programming the robot's software, in particular, with the development of an interactive module exploring the relation between musical stimuli, affective recognition, and expressive feedback.

8.3.2 The Prototype

8.3.2.1 InMoov

EEVAN is a prototype based on the open-source platform *InMoov* [Figure 8.10], created by the French sculptor and designer Gaël Langevin [2012] as the first full-scale humanoid robot entirely reproducible through 3D printing. Originally conceived as an artistic *do-it-yourself* project to be built with relatively inexpensive equipment, InMoov has rapidly gained attention as a low-cost experimental platform, accessible not only to the community of hobbyists and makers, but also to schools, universities, and research groups interested in exploring robotics outside proprietary environments. Moreover, it is increasingly inspiring the field of DIY robotic prosthetics, enabling low-cost and customizable assistive devices.



Figure 8.10: The humanoid robot InMoov —from [Langevin 2012].

The architecture of InMoov is based on a *modular framework* composed of mechanical components such as gears and levers, electronic boards, servo motors, as well as software libraries —available particularly through the *MyRobotLab* system— for managing movement, sensors, and facial expressions. Its design includes articulated hands, arms, torso —optionally non-motorized legs— as well as a head with movable eyes and jaw. This configuration enables the reproduction of basic gestures and elementary forms of speech synchronization. Moreover, the modularity of the hardware allows the customization of the robot's parts according to specific design requirements or experimental constraints, including modifications to accommodate additional sensory modules or video and audio systems.

This combination of openness, transparency, and modularity makes InMoov a distinctive platform within the field of robotics. The entire set of 3D-printable STL models, wiring diagrams, and assembly guides is distributed online, ensuring replicability and technical verification, while a large and active community of developers and makers continuously shares improvements, tutorials, and detailed documentation about the project.

8.3.2.2 From InMoov to EEVAN: An Overview

Within this landscape, the choice of InMoov as the basis for EEVAN was dictated less by its technical sophistication than by such qualities, deemed as particularly valuable for studying and prototyping new forms of human-robot interaction.

In the prototype of EEVAN, InMoov was extended to integrate musical input and affective recognition modules, with expressive outputs including bodily movement, gesture, and auditory or visual responses. The underlying hardware and software were also adapted to support the reception and processing of sound, as well as the synchronization between musical stimuli and corporeal reactions, thereby enabling the generation of expressive feedback during interaction.

The central task of EEVAN was conceived as an exploratory investigation into the affective dynamics elicited by musical stimuli in the context of human-robot interaction. In practice, the system engages a user in a short interactive session, during which the robot initiates basic verbal exchanges accompanied by elementary gestures and movements of the jaw. The user is then invited to listen to a brief musical excerpt, while the robot monitors facial expressions through a computer-vision module. At regular intervals, the system classifies these expressions in affective terms and provides immediate feedback to the user, before producing at the end of the session a synthetic profile of the overall emotional response to the listening experience.

Although modest in scope and limited in technical implementation, this setup illustrates how a humanoid platform can serve as a testing ground for examining the integration of music, embodied interaction, and affective recognition. The results remain preliminary and primarily exploratory, yet they highlight the potential of such low-level prototypes for informing the design of more refined affective architectures in social robotics.

In the following, the main stages of construction and implementation of the project will be outlined, highlighting the progressive development of the prototype across its different phases.

File STL	Tipo di filamento	Colore	Filamento necessario (g)	Tempo di stampa
EyeglassV4.stl	PLA	White	33,12	6:55:00
JawHingeV1.stl	PLA	White	9,73	1:14:00
JawHingeV3.stl	PLA	White	10,11	1:25:00
JawPistonV1.stl	PLA	White	3,20	1:01:00
JawSupportV1.stl	PLA	White	19,07	2:41:00
JawV4.stl	PLA	White	33,24	7:22:00
LowbackV5.stl	PLA	White	46,10	7:29:00
TopMouthV4.stl	PLA	White	34,13	7:17:00
EarLeftV0.stl	PLA	White	44,81	5:40:00
EarLeftV1.stl	PLA	White	49,52	6:21:00
EarRightV0.stl	PLA	White	44,79	5:44:00
EarRightV1.stl	PLA	White	49,50	6:20:00
EarSpeakerRightV1.stl	PLA	White	39,93	6:39:00
EarSpeakerLeftV1.stl	PLA	White	39,92	6:28:00
Speaker1.stl	PLA	White	29,76	3:08:00
SpeakerSpacer1.stl	PLA	White	1,85	0:28:00
TopBackskullV4.stl	PLA	White	63,13	10:31:00
TopskullLeftV4.stl	PLA	White	66,55	11:37:00
TopskullRightV4.stl	PLA	White	66,53	11:47:00
EyeBallFullV2.stl	PLA	White	14,59	3:33:00
EyeBallSupportLifeCamHDLeftV1.stl	PLA	White	6,86	1:43:00
EyeBallSupportLifeCamHDRightV1.stl	PLA	White	6,84	1:41:00
EyeBallSupportHerculesHdLeftV1	PLA	White	8,03	1:44:00

Figure 8.11: Example of a 3D printing log sheet (head components) used during the first year of the internship, indicating the STL files required for the assembly, together with filament type, color, estimated material consumption, and printing time for each component. The printing time has been calculated through the dedicated slicing software in order to optimize the overall workflow.

8.3.2.3 Hardware and Mechanics

The first year of the internship has been devoted primarily to the mechanical completion of the InMoov prototype on which EEVAN is based. At the beginning of the collaboration, the robot already included several (partly) assembled modules [Figure 8.14, above left] —namely the head, torso, right arm, and right hand— since its construction had initially been undertaken within the company by Dr. Facecchia and later suspended. The first step has therefore been a general survey of the project, which has involved verifying the existing parts, compiling an inventory of the missing or damaged components, and organizing the subsequent workflow. Dedicated Google spreadsheets have been created for this purpose, allowing a systematic overview of the pieces to be printed and assembled [Figure 8.11].

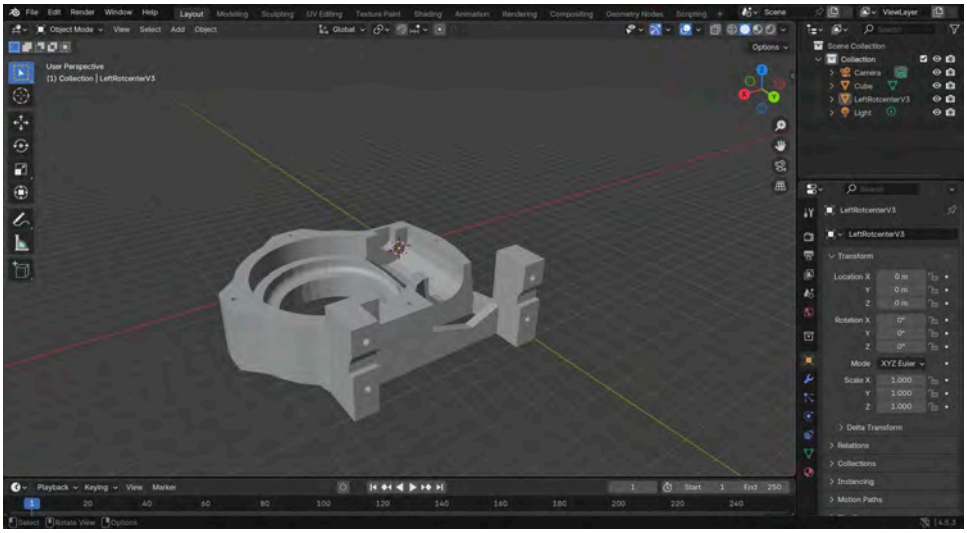
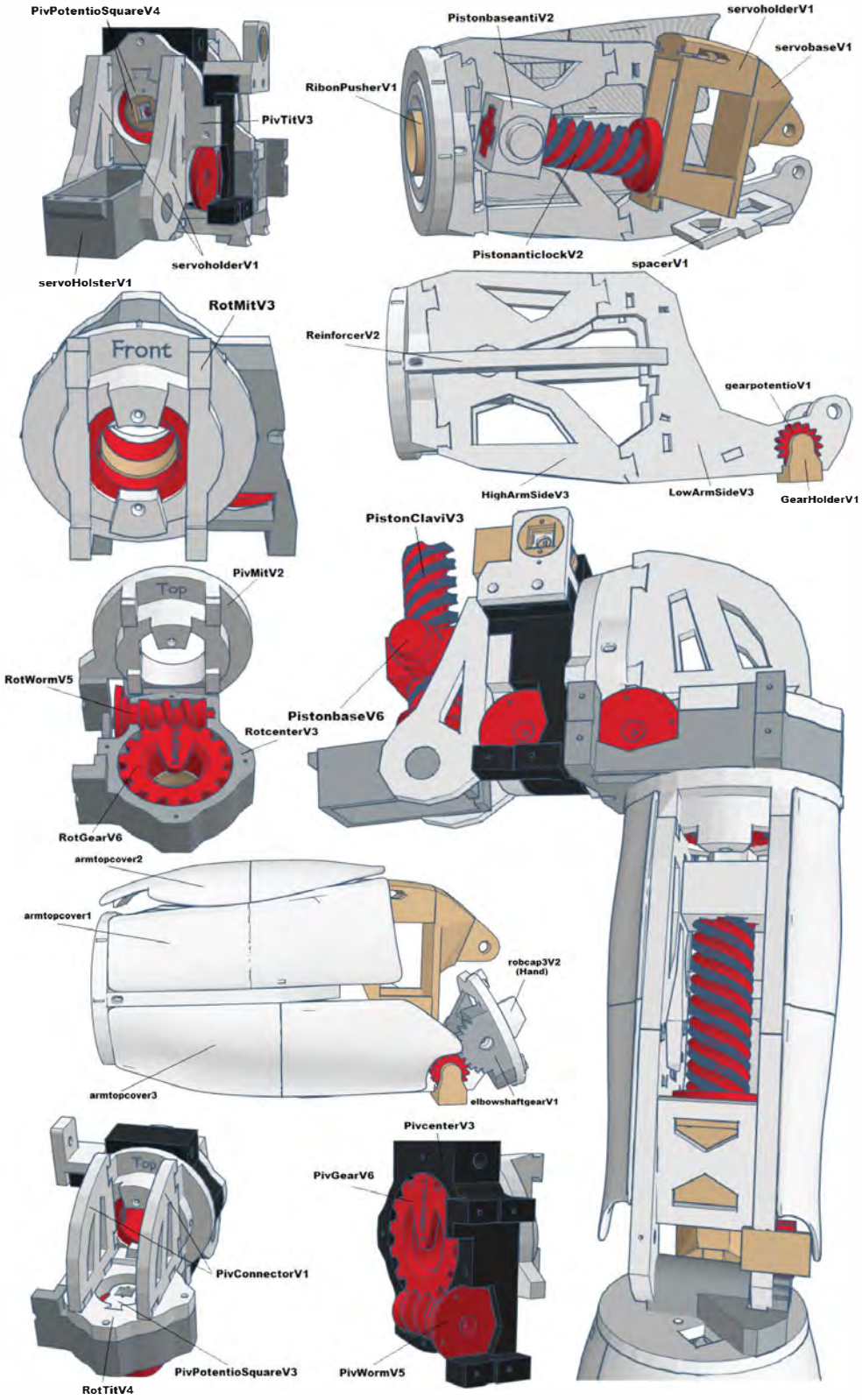


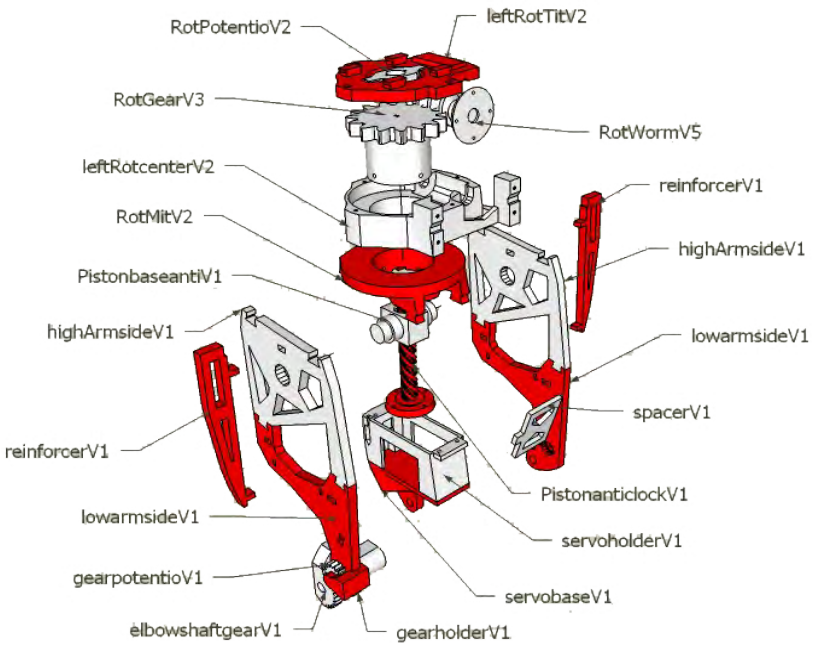
Figure 8.12: Example of 3D of STL components on Blender.

The absent elements have been reproduced with an Anycubic i3 Mega S printer [Figure 8.14, above], using both the proprietary software of the printer and, when required, Blender for minor adjustments of the STL files [Figure 8.12]. The printing process has required continuous calibration, particularly of bed leveling, extrusion temperature, and filament feed rate, as the quality of the mechanical joints depends directly on the precision of these parameters. In several cases, reprinting has been necessary in order to correct issues such as warping or insufficient infill density. Over the course of the summer session, the timeline of the work has alternated between periods of continuous printing —sometimes lasting several days for larger modules— and phases of assembly, in which the new pieces have been integrated into the existing structure.

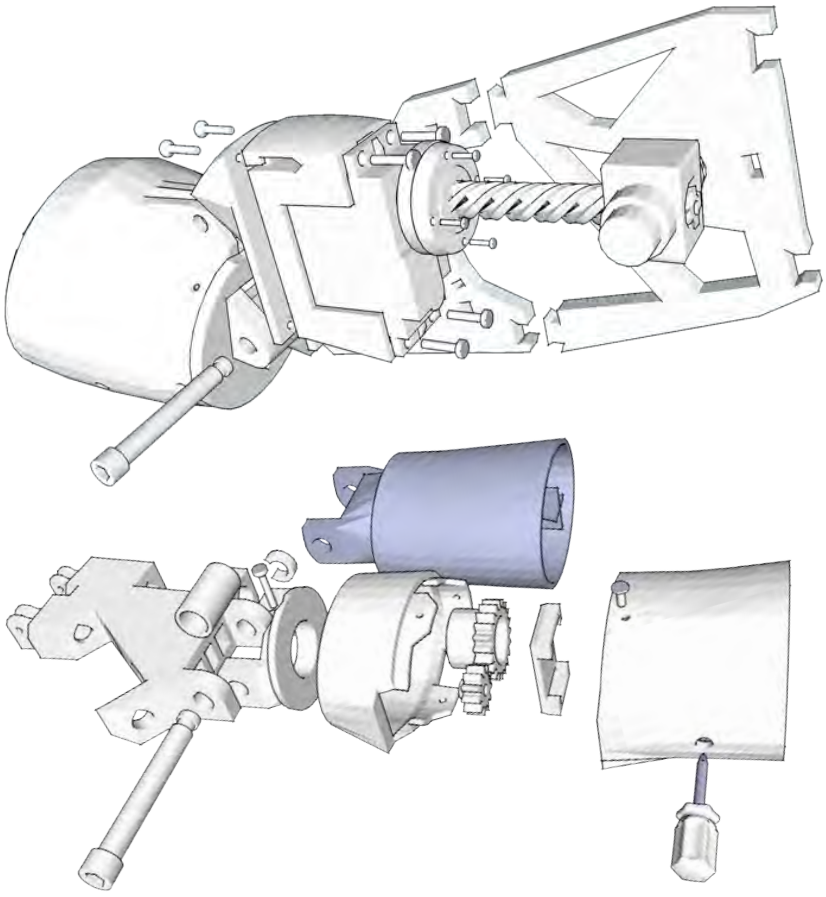
The assembly has concerned above all the left arm and hand (see [Figure 8.14, middle] and [Figure 8.14, above right]), as well as the stomach section with its internal rotation mechanism [Figure 8.14, below]. The hand has required particular care, due to the delicate integration of servomotors and tendons for finger articulation. Damaged or defective components identified during the survey have also been replaced at this stage, ensuring greater consistency and structural stability. Although the legs of InMoov are conceived as optional and non-motorized elements, a decision has also been made to include their construction for reasons of completeness and aesthetic consistency of the prototype [Figure 8.14, below]).



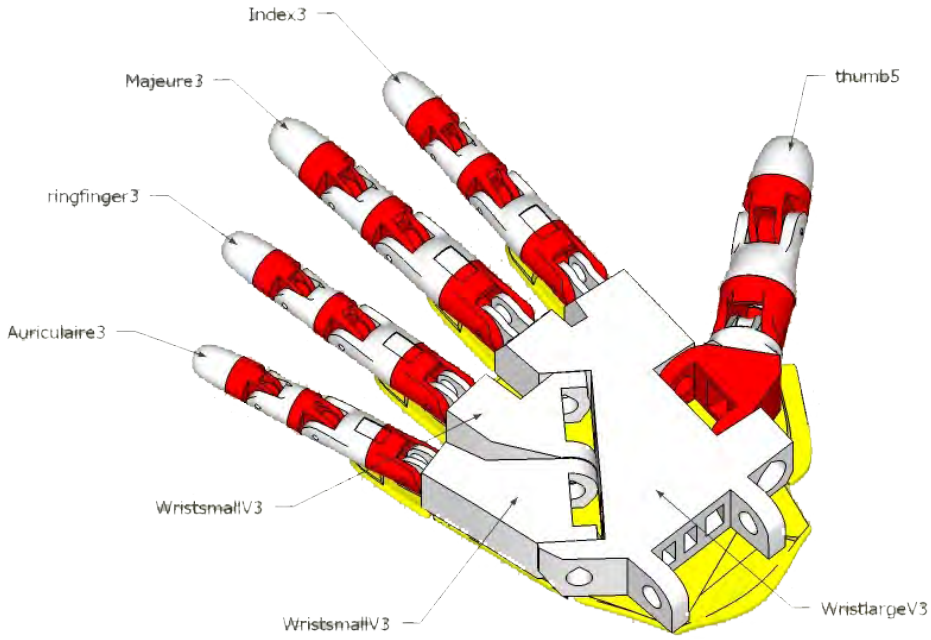
(a) *Outline of the mechanical components of the arm, including servo housings, gear supports, and rotational-tilting joints.*



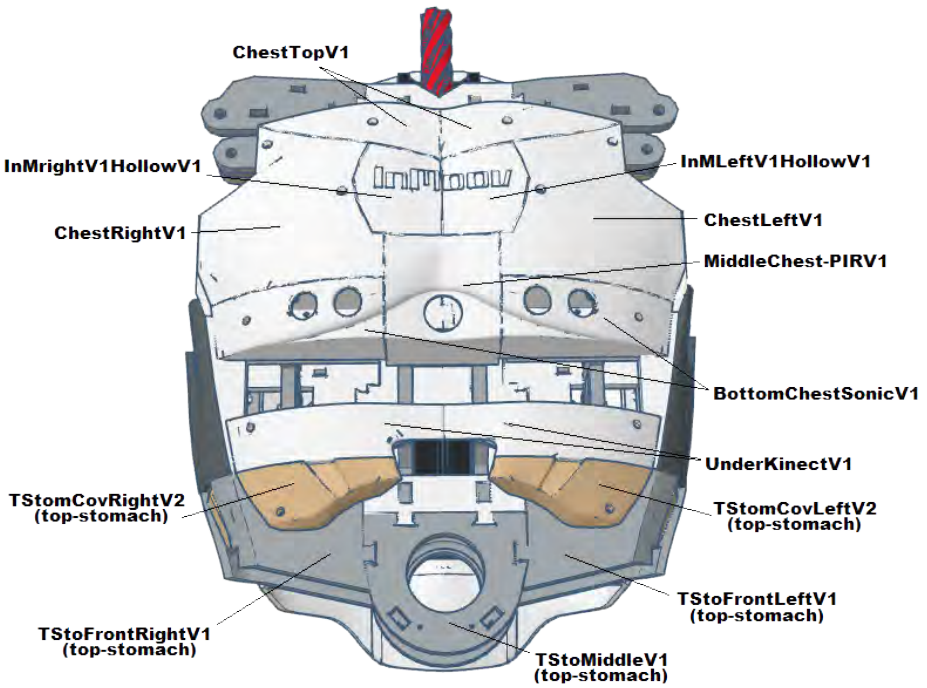
(b) *Exploded view of the upper arm assembly, with the piston mechanism for arm raising and lowering, and the gear-worm for horizontal rotation.*

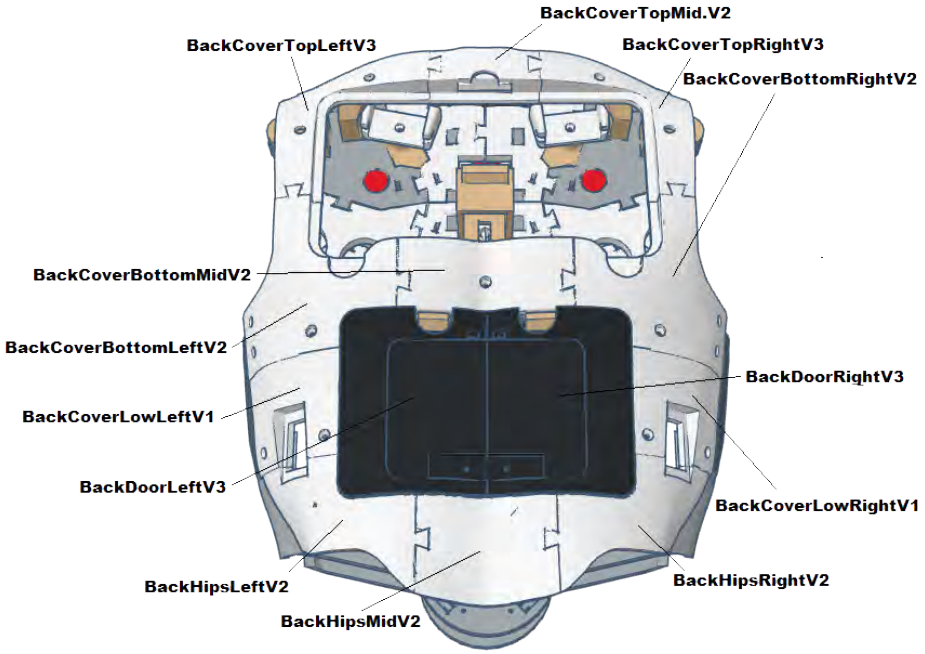


(c) *Connection of upper arm and forearm (above) and rotational mechanism of the wrist (below).*

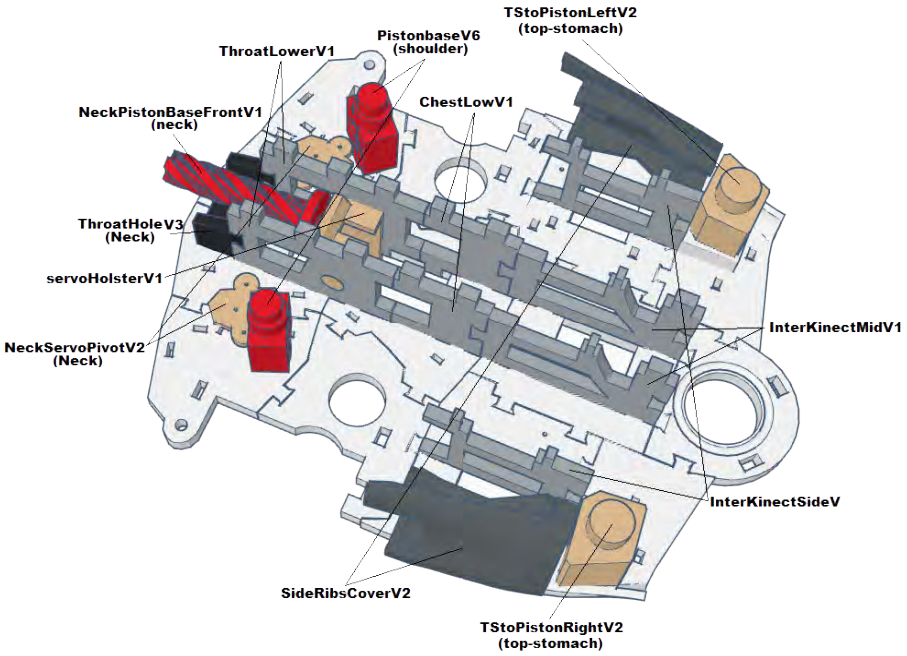


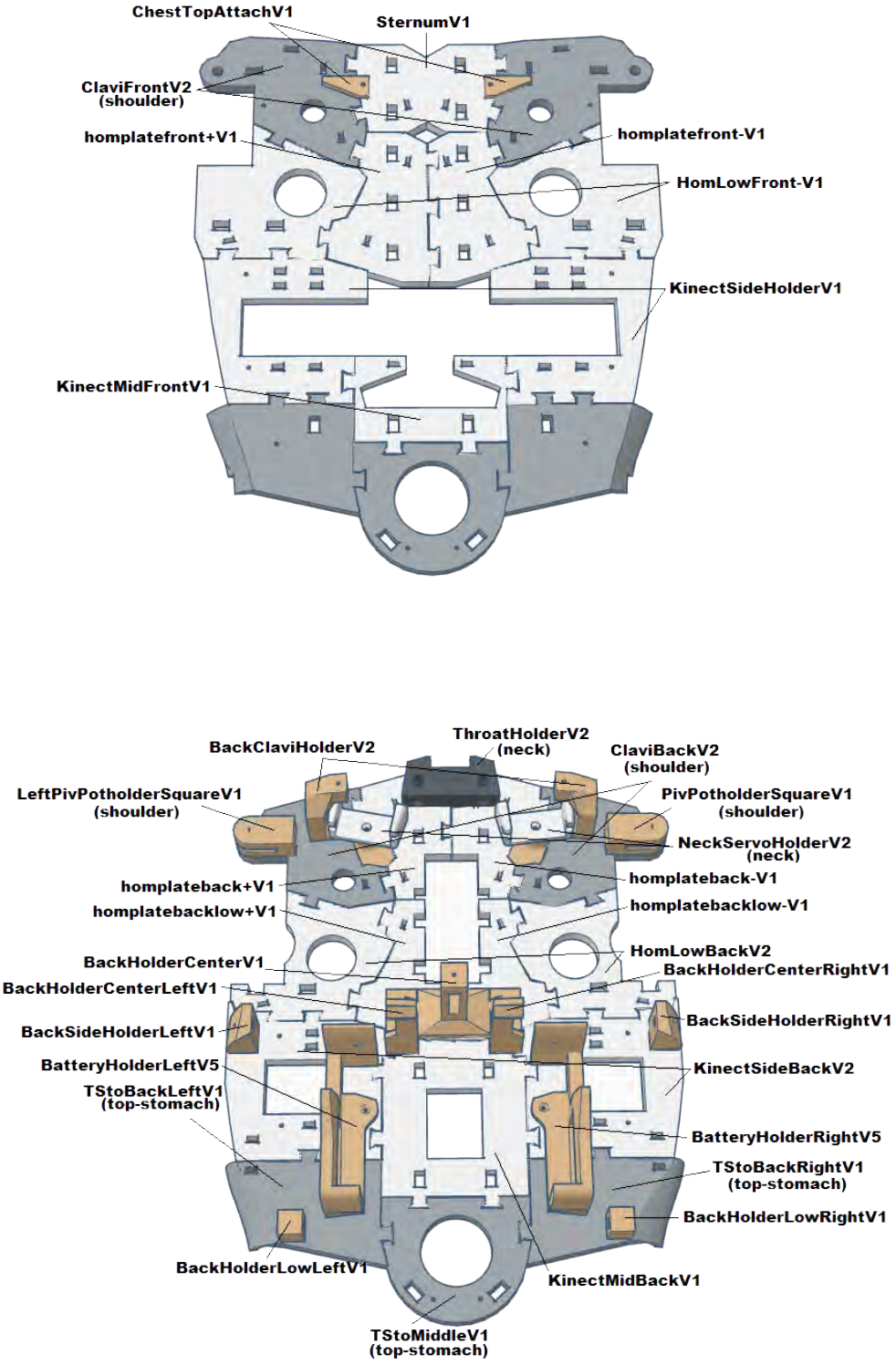
(d) *Components of the hand.*



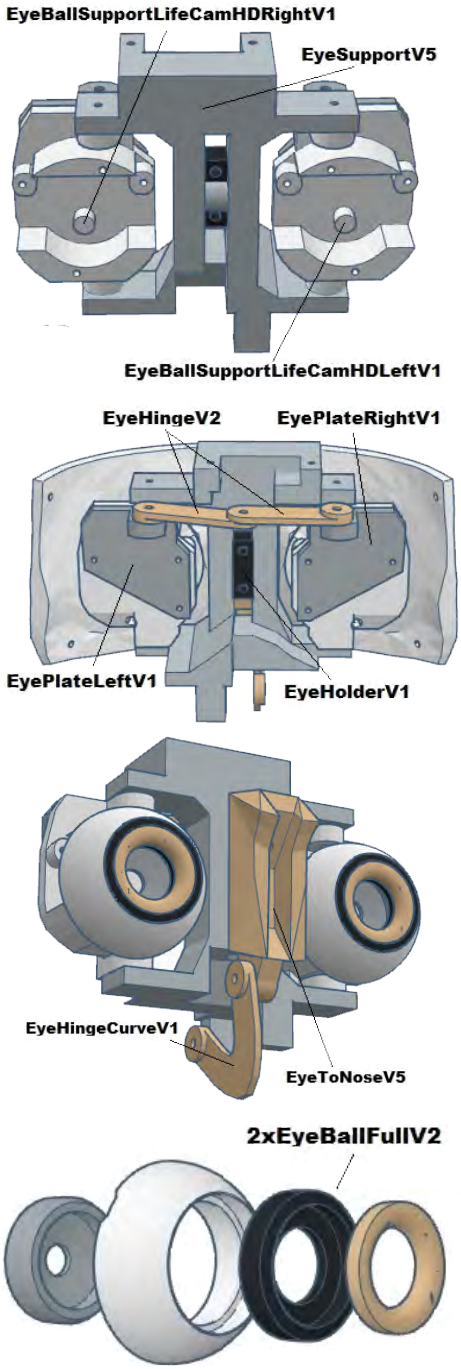


(e) *Front and rear views of the chest assembly, including the enclosing panels and stomach covers that protect and support the internal mechanisms and ciruitry.*

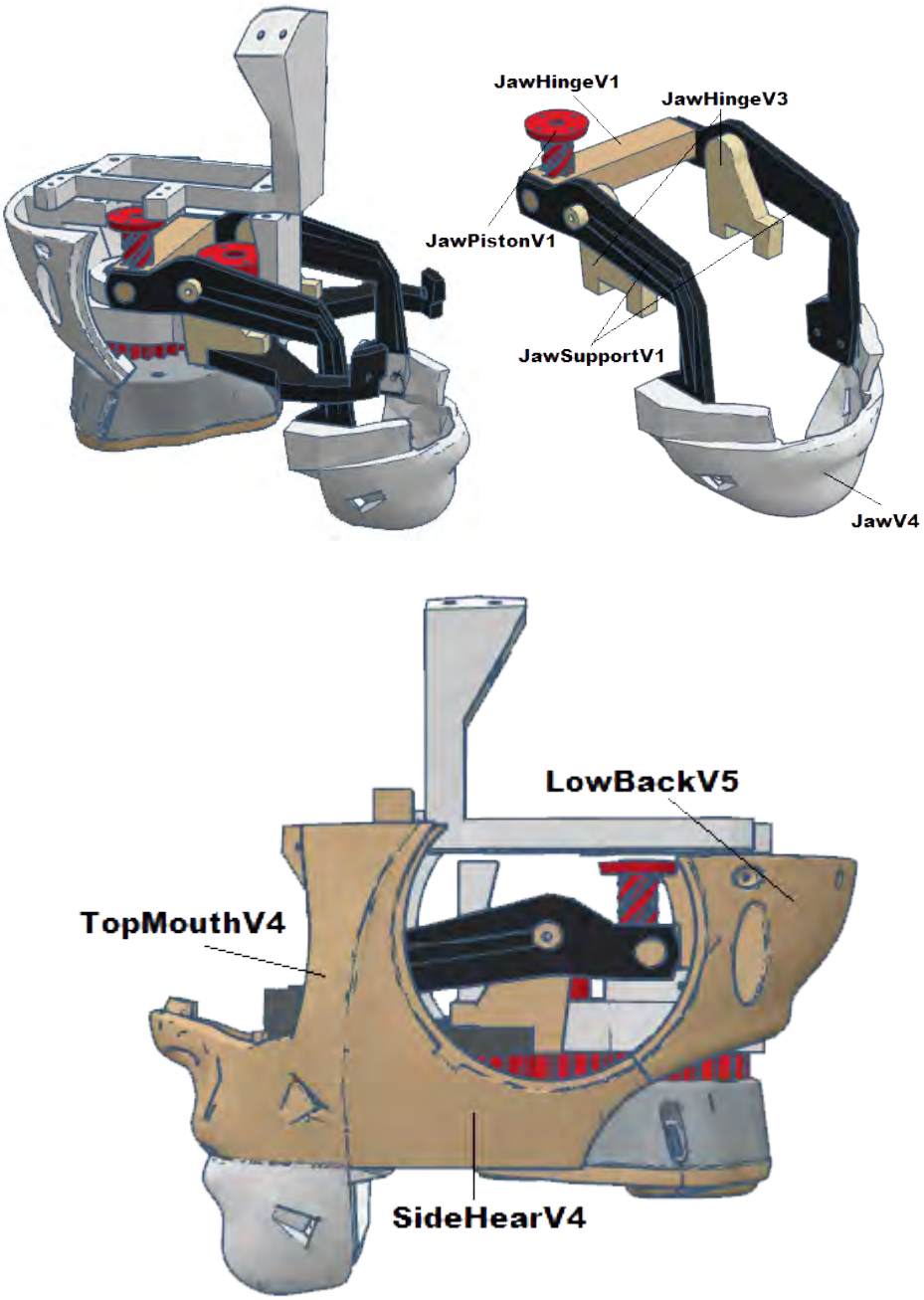




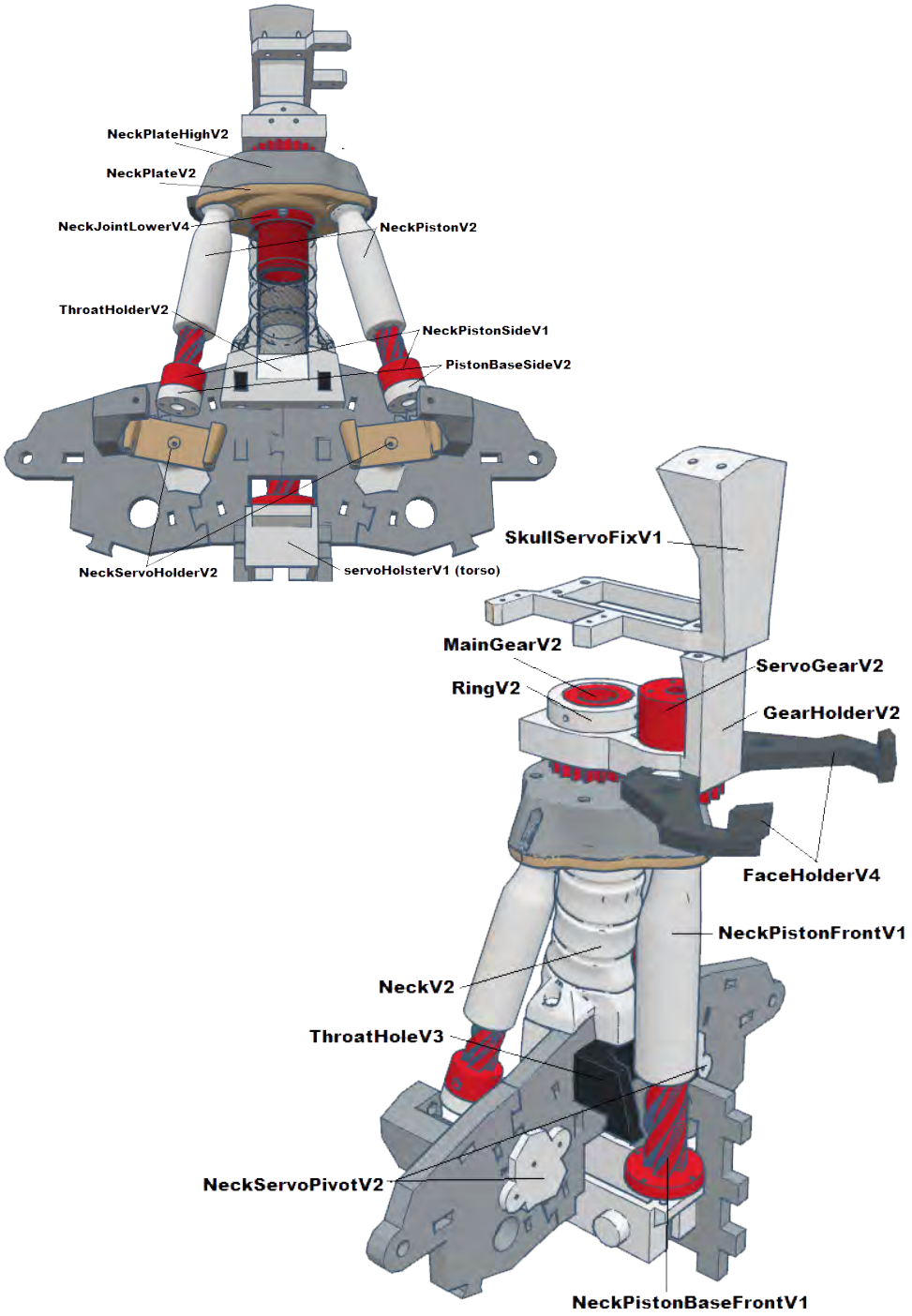
(f) *Internal structure of the chest and shoulder assembly, illustrating the arrangement of piston bases, servo mounts, and structural plates that support the neck and shoulder joints together with holders for sensors and internal electronics.*



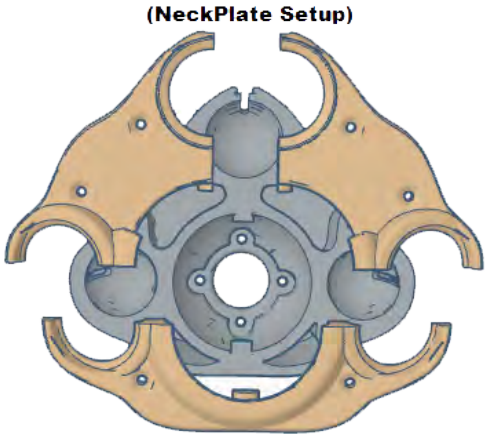
(g) *Exploded schematics of the ocular assembly, showing the support frames, hinge structures, and complete eyeball components designed to host the cameras and enable eye movement.*

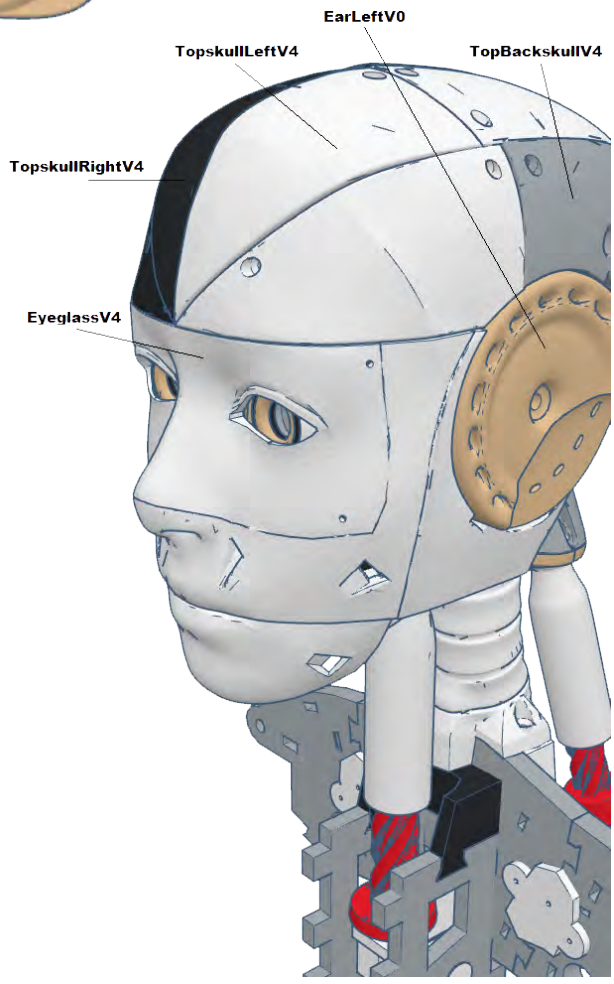


(h) *Exploded schematics of the jaw assembly, illustrating the piston and hinge mechanisms together with the mouth and side structures supporting mandibular movement.*



(i) *Schematics of the neck assembly, showing the piston-based actuation system, servo gears, and structural holders that support head articulation.*





(j) *Assembly of the head external structure, including skull plates, ear covers, and frontal elements completing the cranial enclosure.*

Figure 8.13: Graphic outline of the mechanical components of the upper body, including the arm (a), hand (d), chest (f), ocular system (g), jaw (h), neck and head external structure (i-j), shown through exploded schematics of their main subassemblies —illustrations by Shido Xen.



(a) *Printing phase (above); assembly of the forearm and hand with testing of the finger servomotors (middle); modification and reversed electrical connection of the lower abdominal servomotors for lateral torso bending; tests with Arduino (below).*



(b) *EEVAN's initial assembly state (top left); adjustment of the wire mechanisms for finger movement (top right); assembly and bonding of the leg components (bottom left); completed assembly with non-motorized legs (bottom right).*

Figure 8.14: Photographic record of the EEVAN's hardware assembly process.



(a) *Testing gears and potentiometer of Maxi Servo HS-805BB via Arduino* (link).



(b) *Testing hand and finger movement* (link).



(c) *Another highly recommended electrical test* (link).

Figure 8.15: Video record of the first tests on EEVAN's hardware and servomotors.

Parallel to the mechanical assembly, initial work has been carried out on the electrical setup. Several servomotors have been mounted and tested [Figure 8.15]; non-functioning units have been disassembled and

repaired, with specific attention to the potentiometers that regulate position feedback. In some cases, servomotors have been modified with the removal of the mechanical stop on their internal gears in order to permit continuous rotation. Also, the extraction of the potentiometer from the case has been necessary, to permit its external insertion into the robot's gear system, allowing closed-loop feedback that automatically regulates the speed and direction of the servomotor and, consequently, the movement of the corresponding robotic part. In general, the movement systems of InMoov are based on spiral mechanisms in which rotating cams gradually raise or lower the internal linkages that actuate the robot's parts —see e.g., [Figure 8.13, a-c]. This principle, replicated across several modules of the platform, allows relatively simple servomotor rotations to be translated into more complex motions, by converting rotary input into linear displacement. Consequently, the first tests of movement for the hand, arm, and torso have been carried out through Arduino boards programmed with basic control scripts [Figure 8.15]. These initial trials have confirmed both the feasibility of the assembly and the need for further refinement in later stages of the project.

This initial year of work has thus transformed a partially assembled prototype into a functioning structure, establishing the mechanical and electrical basis on which the subsequent phases of development have been built. At the same time, it has provided a concrete engagement with the challenges of constructing a humanoid robot through low-cost additive manufacturing, where durability, alignment, and reproducibility have to be constantly balanced against the constraints of the available tools and materials.

8.3.2.4 Electronics and Circuitry

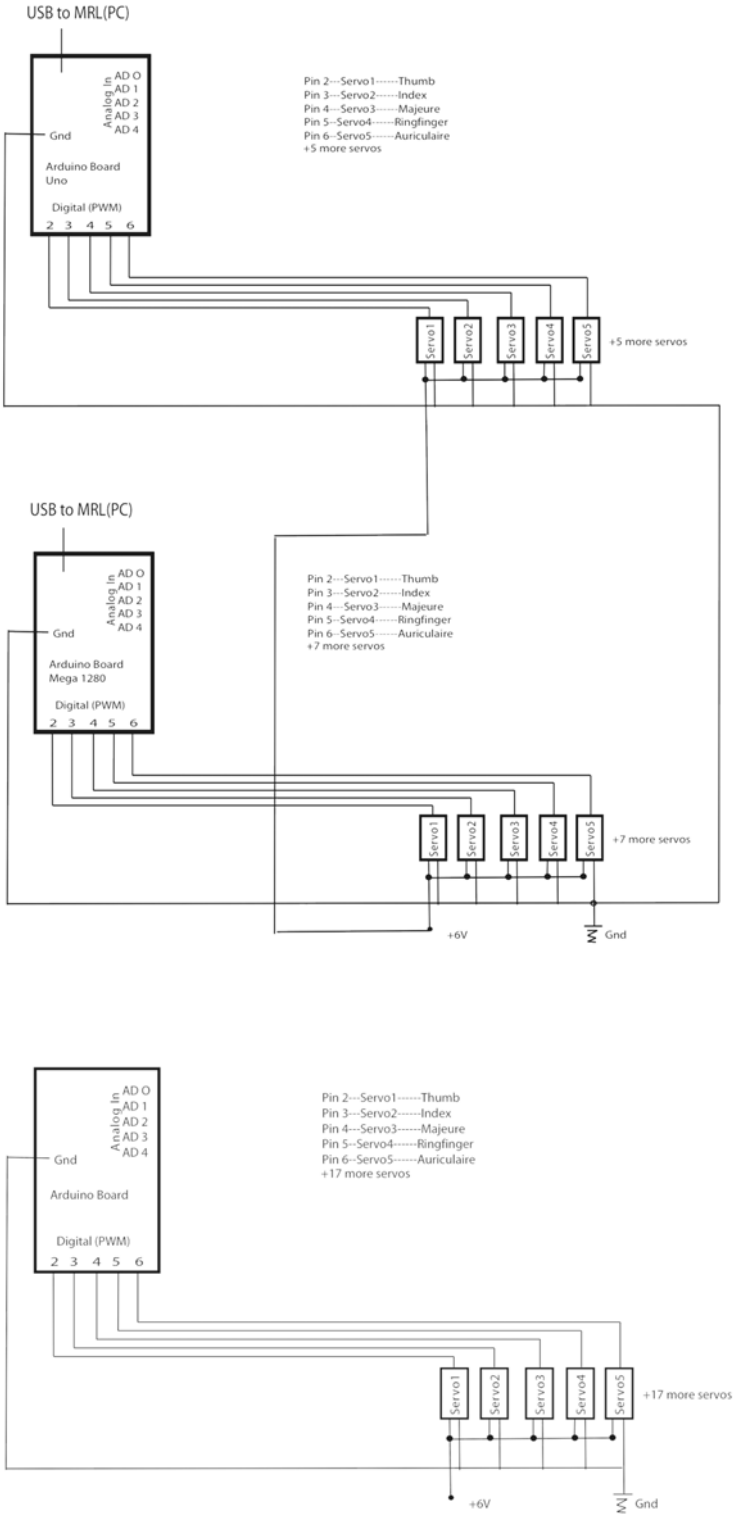
The second year of the internship has been dedicated to the implementation of the electronic and circuitry systems of the EEVAN prototype. Although additional mechanical work has been carried out —including the reprinting of damaged torso and head parts, as well as the optional construction of the legs for aesthetic completeness— the main focus has concerned the setup, testing, and organization of the electrical infrastructure required for basic robot operation.

The work has begun with a reconnaissance of the available electronic material and laboratory equipment. This has included the collection and classification of pre-assembled circuits, Arduino boards, wiring components, and soldering materials, together with the tools necessary for their integration, such as soldering stations, multimeters, and laboratory power supplies. This phase has been essential to establish a systematic overview of the resources available and to plan the workflow for the subsequent assembly and testing activities.

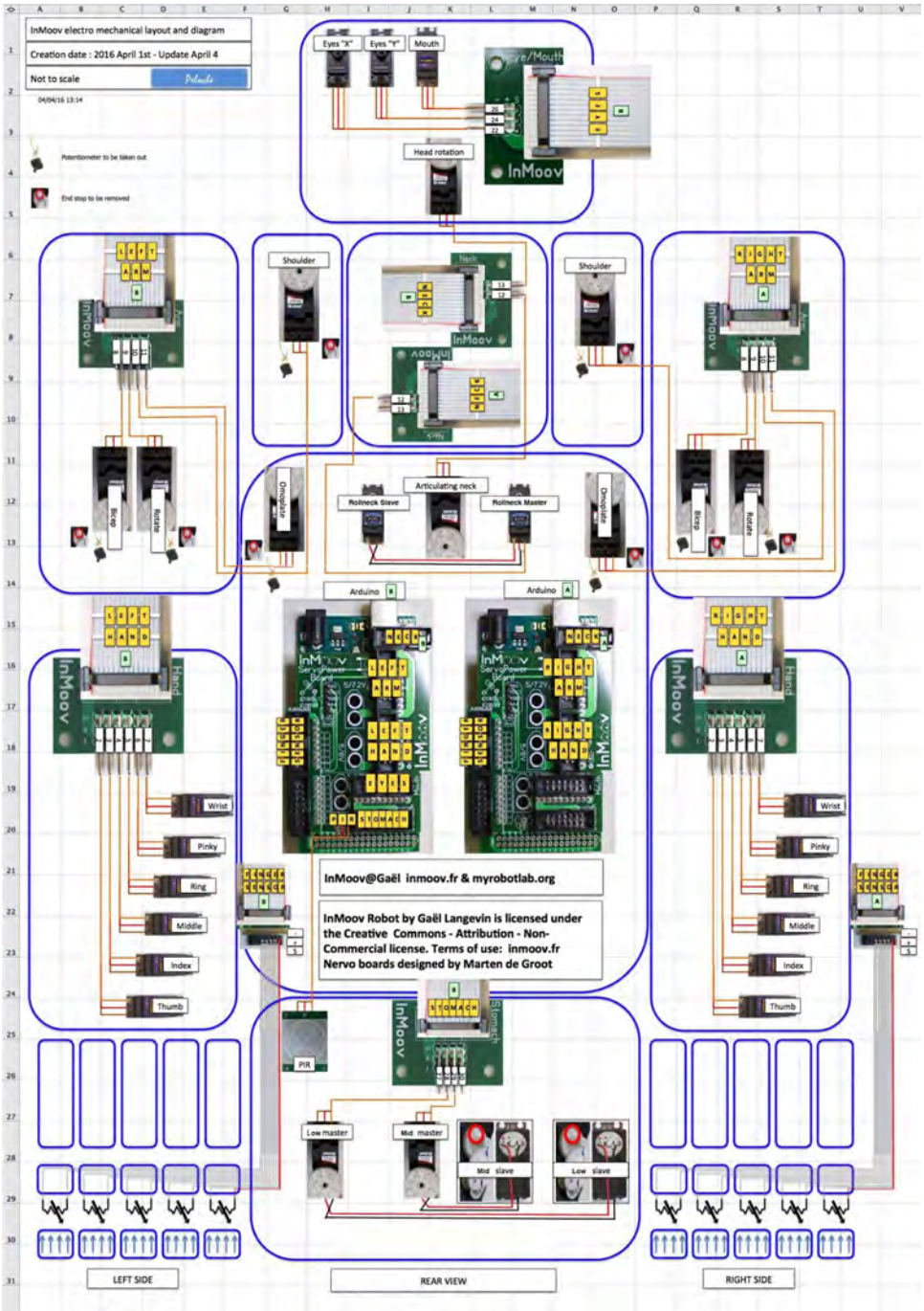
The following stage has involved the soldering of individual components, the routing and bundling of wiring within the robot's internal structure, and the connection of peripheral circuits to the mainboards and the controlling computer [Figure 8.17, a–c]. Standard XT90 cabling has been used for the connection to the 220V power supply, while Arduino boards have provided the primary interface for sensor input and actuator control. Careful attention has been given to the spatial organization of cables and boards inside the torso compartment, in order to reduce electromagnetic interference, avoid overheating, and facilitate future maintenance.

As recommended in the InMoov documentation, the electronic system has been structured around the *Nervo Board* (NB), a custom shield for *Arduino Mega* designed to simplify cabling and to centralize the control of servomotors and sensors. Each NB can manage up to nineteen servomotors and ten sensor inputs, together with additional modules such as a PIR sensor, NeoPixel rings, and i2C devices. In the standard InMoov architecture, two NBs combined with two Arduino Mega units are sufficient to operate a full robot [Figure 8.16, a]. In the case of EEVAN, this setup has been implemented with exception of (optional) additional modules (PIR, i2C, NeoPixel rings), providing a coherent framework for the distribution of power and signals across the arm, hand, and upper body assemblies.

The assembly of the NB has required the soldering of pin headers, jumpers for voltage selection, and breakout mini-boards, followed by the connection to the Arduino Mega through flat ribbon cables [Figure 8.16, b]. This configuration has significantly simplified the wiring process, reducing the likelihood of misconnection and facilitating the organization of the internal circuitry.

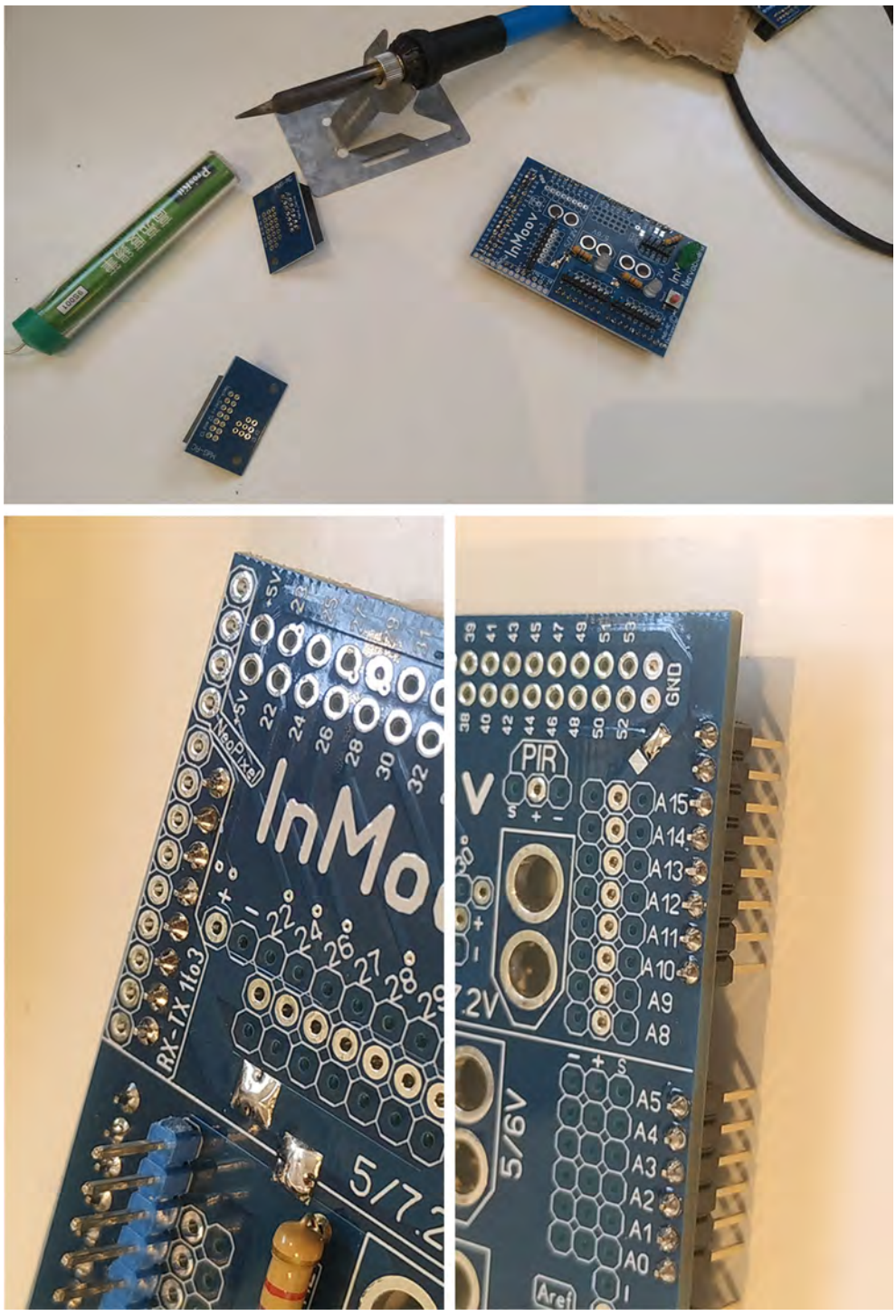


(a) *Default hardware map of the InMoov system, showing the Arduino-based control architecture with USB interfaces to the host PC, distributed power lines, and PWM pin assignments for servomotors, including dedicated channels for finger actuation and additional actuators.*

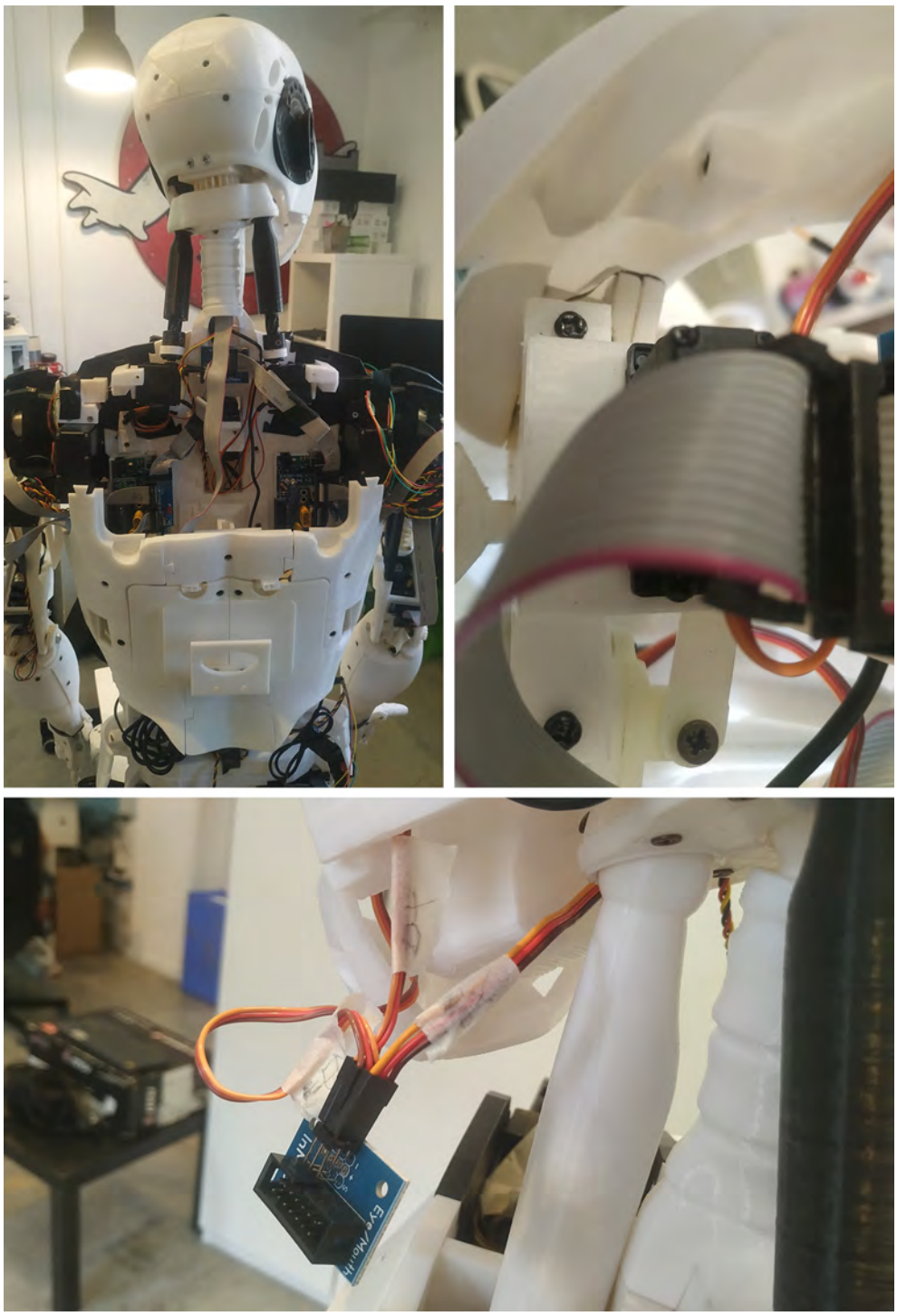


(b) *Electro-mechanical layout of the InMoov system, showing the distribution of Nervo Boards, Arduino controllers, and peripheral connections for servomotors, sensors, and actuators across the head, torso, arms, and hands —by Gaël Langevin, and Marten de Groot.*

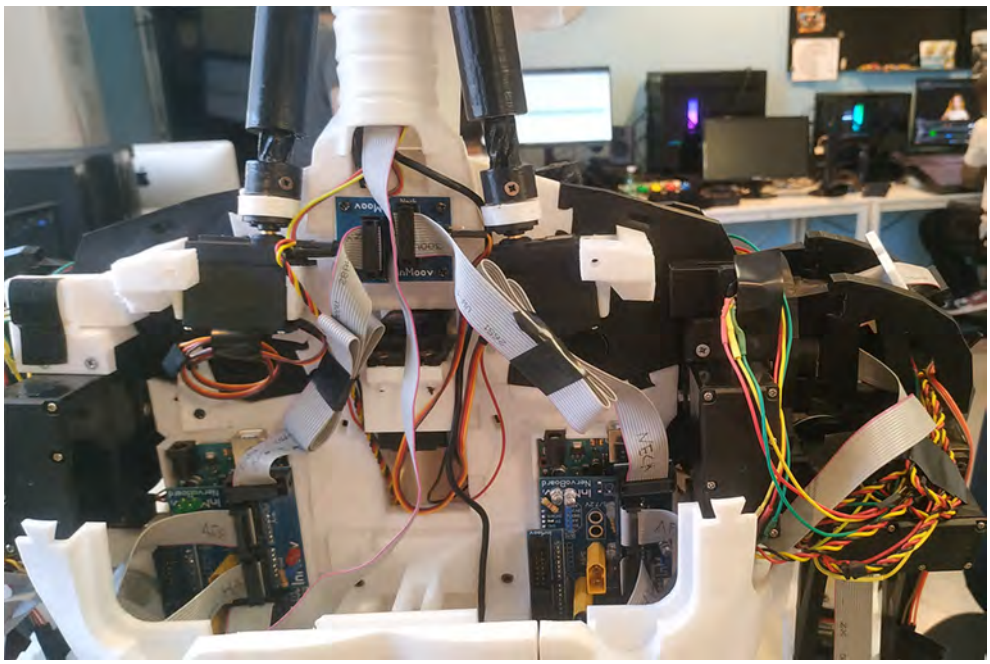
Figure 8.16: Default hardware schematics (a) and electro-mechanical layout (b) of the InMoov system, illustrating the Arduino-based control architecture with NBs, distributed power supply, and wiring connections for servomotors, sensors, and actuators across the main modules of the robot



(a) *Assembly of the NB, with soldering of headers and breakout modules for the integration of servomotors and sensors.*



(b) *Integration of the NB within the torso of the robot, showing the ribbon cable connections and sensor wiring routed through the internal structure.*



(c) *Detailed view of the torso interior with dual Arduino boards, NBs, and ribbon cabling for the distribution of signals and power to the upper body modules.*

Figure 8.17: Photographic report of the electronic assembly of EEVAN.

(a) *Testing arm mobility after complete assembly* (link).



(b) *Testing jaw mobility and speech synchronization* (link).

Figure 8.18: Video record of mobility tests on the completed prototype.

Particular attention has then been devoted to the integration of the sensing apparatus. A standard USB camera has been mounted inside the ocular assembly, providing the robot with a functional visual channel directly aligned with its anthropomorphic design. In addition, a Microsoft Kinect sensor has been embedded in the abdominal section, where it has served both as a motion detection device and as a platform for initial tests in human-robot mirroring of movements. This configuration has made it possible to experiment with the reproduction of user gestures through the robot's actuators. As for the auditory system, the microphone and speakers have so far been implemented externally, through the connected computer, rather than embedded into the robot's structure. While this solution is temporary, it has ensured stable performance for testing without requiring additional modifications to the head assembly.

Trials with OpenCV libraries have allowed basic implementations of facial, vocal, and movement recognition, demonstrating the feasibility of connecting the visual and auditory inputs with the mechanical responses of the robot. Although exploratory in scope, these experiments have established a foundation for subsequent developments in multimodal interaction and user monitoring.

After the completion of the wiring layout and sensor integration, a series of systematic electrical tests has been carried out, measuring voltage, current, and resistance across the main circuits. These tests have ensured the stability of the power distribution system and the reliability of the connections, reducing the risk of component damage during operation. On this basis, movement tests of the individual modules have been

conducted, beginning with the hands, arms, and head, and extending to the ocular system with preliminary implementations of eye tracking. The tests were executed through custom Arduino scripts, complemented by Python programs and the MyRobotLab software. By integrating libraries for face and shape recognition, this framework enabled the coordination of multiple actuators simultaneously while also supporting the first experiments in gaze tracking [Figure 8.18, a–b].

Overall, the work on the electronic and circuitry setup has provided the prototype with a functional, though still provisional, infrastructure for actuation and sensing. Although limited in terms of robustness and performance, this stage has represented an essential step towards transforming EEVAN from a static mechanical structure into a system capable of basic interactive behavior.

8.3.2.5 Software

The software development of the EEVAN prototype has constituted the final phase of the project, where the mechanical and electronic components have been integrated into a provisional yet functional architecture for perception and interaction. The goal has not been to achieve an advanced robotic platform, but to implement a minimal system capable of supporting exploratory tests on the affective dynamics elicited by musical stimuli within human-robot interaction.

The first trials have been carried out using *MyRobotLab* (MRL), an open-source software environment that represents the default operating system for the InMoov platform. MRL is a modular framework that allows the management of servomotors, sensors, and speech synthesis through a graphical interface and a wide library of predefined services [Figure 8.19]. Its architecture is based on the concept of *services* that can be activated and interconnected in order to perform specific tasks, ranging from motor control and facial tracking to voice recognition. For the purposes of the prototype, the software has provided a convenient starting point, as it allowed rapid testing of the initial mechanical assembly and synchronization between speech and jaw movements without the need for low-level programming.

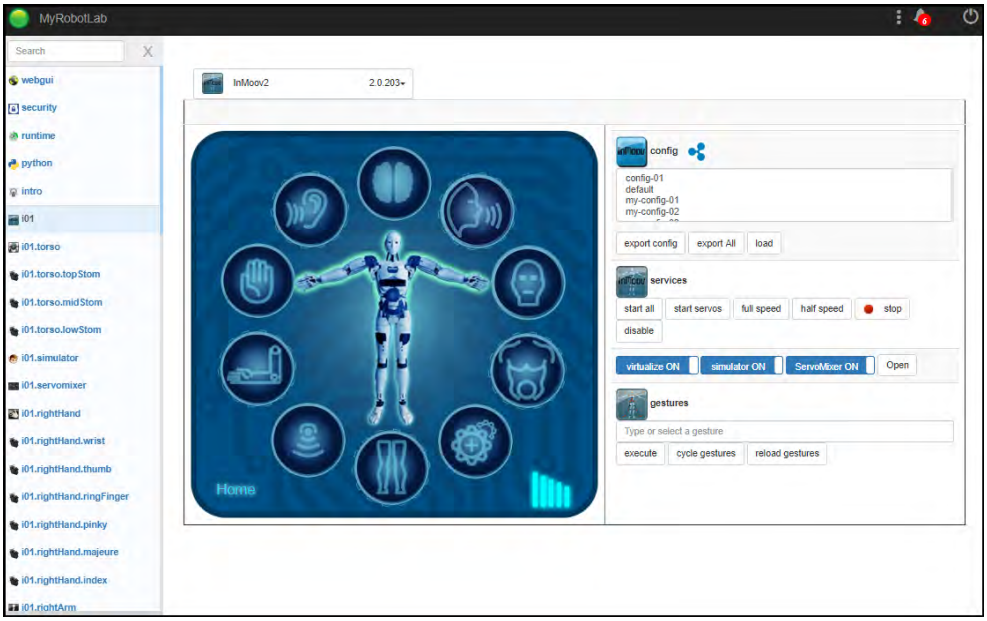


Figure 8.19: Example of MyRobotLab's interface, showing the default control dashboard of the InMoov platform. The graphical layout provides access to different body modules (arms, hands, torso, head), together with configuration options for motor control, gesture execution, and sensor management.

However, two main limitations have emerged during the testing phase. First, the electromechanical management of the potentiometers previously embedded in InMoov's joints has proven inadequate and difficult to properly configure, resulting in imprecise control of specific gestures and complications in calibrating the range of motion. Second, the integration of additional functionalities required by the project —particularly those related to affective recognition and musical interaction— has encountered significant obstacles, as MRL offered limited flexibility for implementing custom routines within its environment.

For these reasons, the development process has shifted toward an alternative software architecture, conceived to overcome these limitations and to provide a more adaptable framework for the purposes of the project.

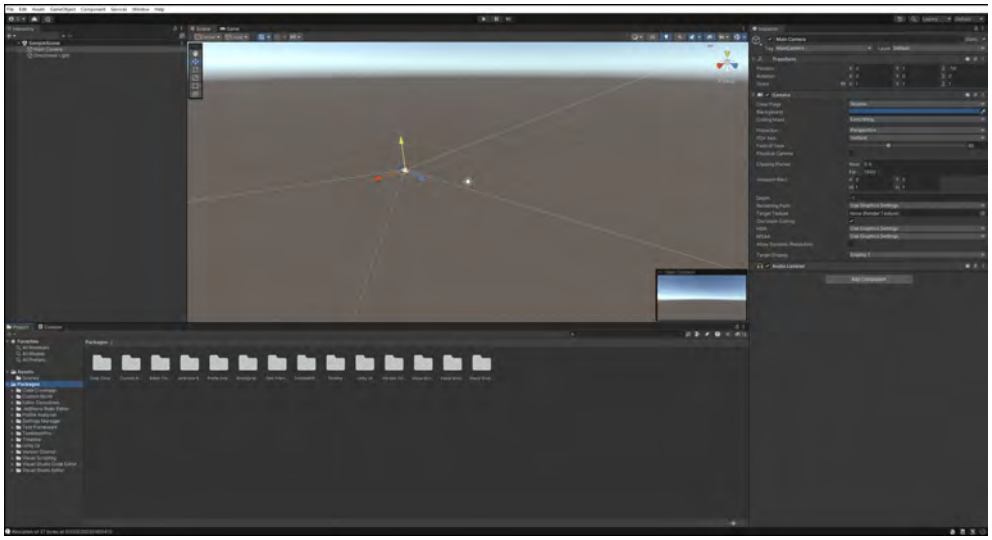


Figure 8.20: Example of Unity working environment of the EEVAN prototype, employed for integrating motor control with perception modules and for monitoring the execution of the interaction pipeline in real time.

At the most basic level, custom *sketches* written in the Arduino IDE have been employed to ensure direct control of single servomotors [Listing 8.1, a]. These programs, running permanently on the microcontrollers, have been deployed so as to receive and interpret serial inputs and to translate them into low-level routines governing the actuation of servomotors. Within the EEVAN prototype, the sketches have been implemented to encapsulate hardware-specific responsibilities —pin mapping, motion limits, timing constraints and basic safety checks— thereby decoupling higher-level decision logic on the host computer from the particulars of electrical actuation.

On top of this layer, C# scripts have been employed on the host side to regulate the communication with the Arduino board [Listing 8.1, b]. Through these routines, the host has been able to transmit control commands to the microcontrollers, invoking, once interpreted, the corresponding low-level instructions embedded in the sketches. This intermediate level has thus enabled the definition of simple gestures and the refinement of motor coordination, while preparing the ground for the integration of additional perceptual and interactive modules.

At a higher level, all these components have been integrated into Unity [Figure 8.20]. As a widely adopted cross-platform engine origi-

nally conceived for interactive 3D applications and now increasingly used in simulation, prototyping, and robotics, Unity provides a modular environment in which scripts, external libraries, and hardware interfaces can be combined into a coherent workflow. Its architecture is structured around the concepts of *game objects* and *components*, extended through C# scripts to define behaviors, interactions, and communication with external systems.

Within the EEVAN prototype, Unity has operated as the orchestration layer of the overall software architecture. In this role, it has managed the transmission of serial commands to the Arduino microcontrollers, while simultaneously coordinating the perceptual modules implemented through OpenCV and handling the playback of musical excerpts. This integration has enabled a multimodal synchronization between affective recognition, gesture execution, and speech output. Moreover, the graphical environment of Unity has facilitated the visualization of the interaction pipeline, the monitoring of internal processes, and the debugging of system performance during testing. In this way, Unity has provided the prototype with a flexible control center, supporting both the experimental goals of the project and the possibility of scaling the system toward more advanced implementations.

```
#include <Servo.h>

Servo arm;

void setup() {
  arm.attach(9);         // pin association
  Serial.begin(9600);    // initialize serial communication
}

void loop() {
  if (Serial.available()) {
    char cmd = Serial.read();
    executeCommand(cmd); // delegate to routine
  }
}

void executeCommand(char cmd) {
  switch (cmd) {
```

```
    case 'L':
      arm.write(60);      // gesture: move left
      break;
    case 'R':
      arm.write(120);     // gesture: move right
      break;
    default:
      break;              // ignore undefined inputs
  }
}
```

(a) *Exemplary* Arduino *sketch for motor control, adapted in the actual implementation to the specific hardware configuration.*

```
using UnityEngine;
using System.IO.Ports;

public class ArmController : MonoBehaviour
{
    private SerialPort serial;

    void Start() {
        serial = new SerialPort("COM3", 9600);
        serial.Open();
    }

    void Update() {
        if (Input.GetKeyDown(KeyCode.LeftArrow)) {
            SendCommand("L");   // request left movement
        }
        if (Input.GetKeyDown(KeyCode.RightArrow)) {
            SendCommand("R");   // request right movement
        }
    }

    private void SendCommand(string command) {
        if (serial != null && serial.IsOpen) {
            serial.Write(command);
        }
    }

    void OnApplicationQuit() {
        if (serial != null && serial.IsOpen) {
```

```
30                 serial.Close();
31             }
32         }
33 }
```

(b) *Exemplary* Unity C# *script for serial communication, extended in the prototype to integrate perception and interaction modules.*

Listing 8.1: Exemplary Arduino (a) and Unity C# (b) routines illustrating the basic communication scheme adopted in the prototype, where serial inputs are transmitted from the host and interpreted by the microcontroller to actuate motor commands. The actual code implementation —not directly reported in this chapter due to length, context-specific adjustments, and confidentiality constraints requested by the company— has extended this structure to integrate additional perceptual and interaction modules.

8.3.2.6 Program

The behavioral pipeline of the EEVAN prototype has been implemented as a sequential loop, designed to orchestrate perception, motor control, and feedback within a minimal yet coherent architecture [Figure 8.21]. The interaction begins when a person stands in front of the robot: If presence is detected for more than five seconds, the robot initiates a short presentation, accompanied by simple head and jaw movements. At the end of this introduction, the user is invited verbally to interact with the robot.

This is possible through an additional speech recognition module implemented into the prototype beyond visual recognition. This module is capable of detecting simple keywords commands (e.g., “yes”, “no”, “information”, etc.) that enable the user to select and trigger predefined

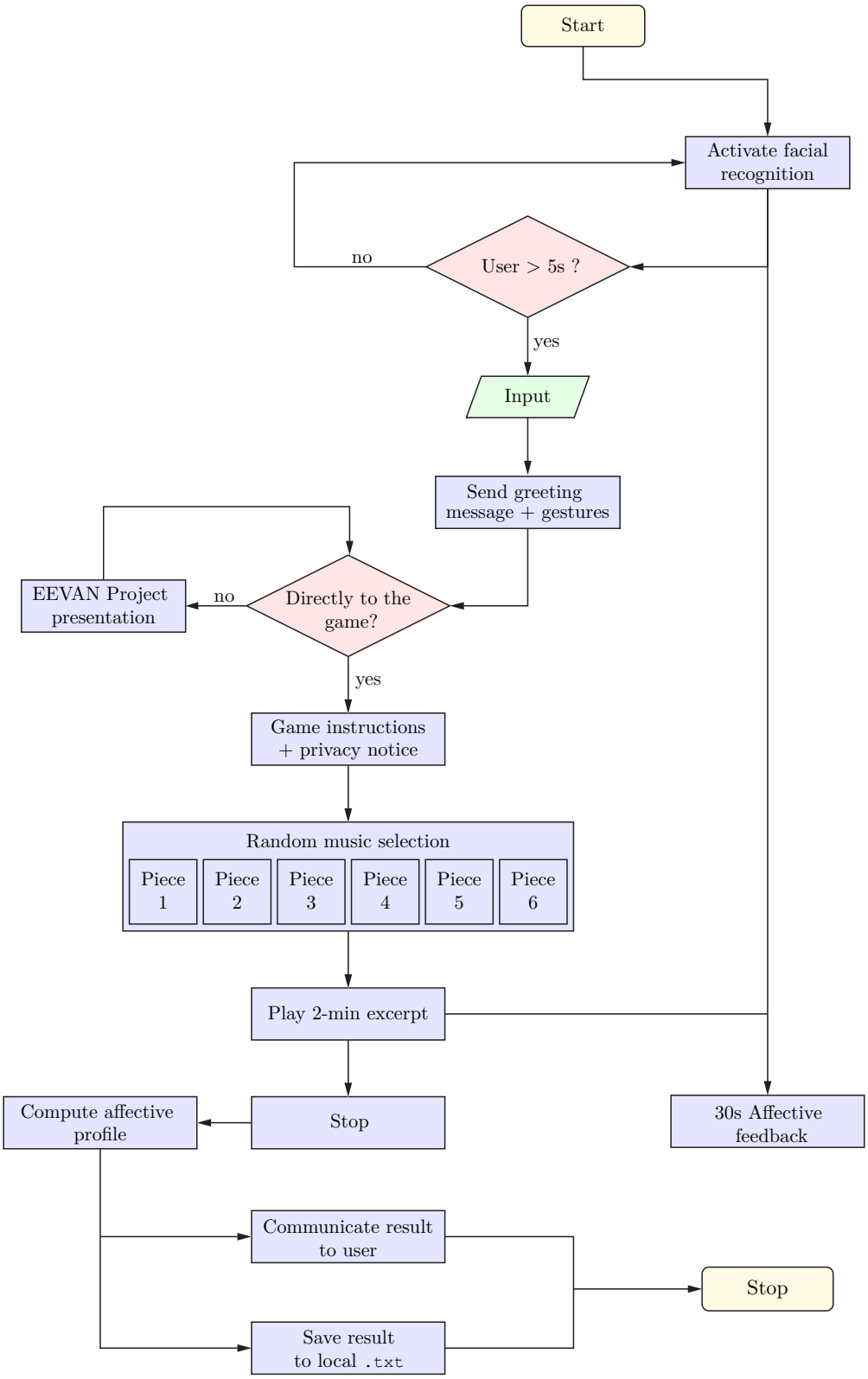


Figure 8.21: Flow chart of EEVAN's currently implemented program.

subroutines[126]. In particular, the user is asked to choose between two options: receiving information about the project or engaging in a musical interaction game. In the first case, the robot delivers an extended presentation on the project, its context, origins, and integration within the research framework of this thesis. In the second case, the interaction follows a different path. After providing the user with instructions, a short description of the activity, and basic information regarding privacy and data use, the system initiates the game.

At this stage, the robot randomly selects one sound excerpt from a predefined set of six short compositions [Figure 8.22], originally produced for the prototype with Logic Pro from open-source sound libraries (in particular, "free-sound.org"). These excerpts, while musical in form, have been conceived as affective sonic events, designed to elicit variations of emotional response rather than aesthetic appreciation (see [Section 8.1] and [Section 8.2]). The listening session lasts for two minutes, during which the user remains seated in front of the robot.



Figure 8.22: EEVAN's music and voice (link).

Throughout playback, the computer-vision module continuously monitors facial expressions. Recognition is implemented through OpenCV libraries, combining Haar cascades for face detection with local feature descriptors and supervised models trained on open datasets for expression classification. This process enables the categorization of basic affective

[126] At this stage, EEVAN has not yet been integrated with advanced AI-driven speech recognition and natural language generation systems (e.g., ChatGPT) for real-time conversational interaction, as the musical task under consideration did not require such functionality. Nevertheless, integrating such capabilities in future implementations would be technically feasible and could significantly expand the system's interactive potential.

states in real time, within the computational constraints of the prototype. At fixed intervals of thirty seconds, the system verbalizes short feedback messages —brief comments on the detected emotional state— accompanied by synchronized jaw movements. All the speeches produced by EEVAN have been generated using ElevenLabs, with its API adapted through custom Python scripts to handle longer or numerous speech excerpts.

At the conclusion of the two-minutes musical excerpt, the system computes an aggregate profile of the monitored responses, synthesizing them into a verbal summary and saving the recorded data together with the affective average into a local `.txt` file. The robot then closes the session with a short thank-you message, thus terminating the interaction.

While limited in robustness and scope, this implementation illustrates how a humanoid platform can integrate sequential modules of perception, expression, and affective feedback into a minimal yet complete pipeline. It provides a proof-of-concept for coupling sonic stimuli with real-time monitoring and embodied communication, while also offering a testbed for future refinements in multimodal interaction design.

8.3.3 Experiment

Upon the implementation of the EEVAN prototype, an exploratory experiment has been conceived as a preliminary examination and validation, in a controlled setting, of the idea that music and sound can enhance affective human-robot interaction. In the following, the structure, organization and results of the experiment are described.

8.3.3.1 Background

The particular aim of the experiment has been to assess whether the affective impact of musical interaction is modulated by the degree of embodiment through which it is delivered. Since the prototype has been designed around the integration of sound as an expressive and relational medium, the experiment has sought to investigate whether the same musical routine, when mediated by a humanoid robot rather than by a disembodied computer interface, would elicit stronger emotional engagement in users. In this way, embodiment has been not treated as an alter-

native to the musical component but rather as a condition for testing its affective efficacy: The question has been whether music, already present as the stimulus of interaction, can gain in intensity and resonance when coupled with a physical agent endowed with anthropomorphic form and behavior. This design has eventually allowed for a first empirical testing of the central hypothesis of the project —that music can operate as a privileged channel for affective communication in human-robot interaction— while pragmatically adapting to the constraints of the available prototype and experimental resources.

8.3.3.2 Conditions and Hypotheses

In order to examine the role of embodiment in mediating the affective impact of music, two experimental conditions have been defined:

> *Embodied condition* (C_1): The musical routine has been presented by the humanoid robot;
>
> *Non-embodied condition* (C_2): The musical routine has been presented by a computer interface.

Building on this design, the experiment has been articulated around two falsifiable hypotheses:

> *Null hypothesis* (H_0): There is no significant difference in the level of emotional engagement observed between C_1 and C_2;
>
> *Alternative hypothesis* (H_1): Emotional engagement is greater in C_1 than in C_2.

8.3.3.3 Participants

Participants have been recruited among voluntary employees of the hosting company, resulting in a total sample of $N = 10$. The reduced size of the sample has reflected both the exploratory character of the study and the limited availability of personnel. In order to implement a *within-subject* design while controlling for order effects, participants have been divided into two equivalent groups:

> *Group A*: P1–P5;

Group B: P6–P10.

Each participant has completed both conditions, with the order counterbalanced across groups: *Group A* has first experienced the embodied condition (C_1) and subsequently the non-embodied condition (C_2), whereas *Group B* has followed the reverse sequence [Table 8.2]. This arrangement has been chosen to mitigate possible novelty, learning, or fatigue effects, thereby ensuring that any observed differences can be more reliably attributed to the experimental variable under study.

Group	**Session 1**	**Session 2**
A (P1–P5)	C_1 (Robot)	C_2 (Computer)
B (P6–P10)	C_2 (Computer)	C_1 (Robot)

Table 8.2: Experimental design with counterbalanced order of conditions across the two participant groups. Group A has first completed the embodied condition (C_1) followed by the non-embodied condition (C_2), whereas Group B has performed the reverse sequence.

8.3.3.4 Experimental Environment

The experimental sessions have taken place in a controlled environment, specifically a quiet room with constant artificial lighting and minimal external distractions. A dedicated computer has been used for both audio playback and facial expression recording. In order to preserve comparability across conditions, the audio volume has been kept constant throughout all sessions.

In the embodied condition C_1, the humanoid robot has been positioned in front of the participant, executing its standard interactive program. The musical excerpts have been played through speakers co-located with the robot to ensure that the sound source has been spatially attributed to the embodied agent.

In the non-embodied condition C_2, the robot has been removed from the room, and the same musical routine has been reproduced via the computer, using speakers placed in the exact physical position previously occupied by the robot. This arrangement has guaranteed that the

only experimental variable manipulated between conditions has been the presence or absence of embodiment.

In both cases, participants have been alone in the room during the session, in order to minimize social influences and external distractions.

8.3.3.5 Experimental Procedure

The experimental procedure has followed a standardized sequence of steps, with an overall duration of the protocol of approximately 90 minutes.

At the beginning of each session, participants have received an information sheet [Figure 8.23] explaining the purpose of the study, the nature of the tasks, and the available options to choose during the interaction, including the musical game described in [Section 8.3.2.6]. The document has also clarified that participation has been voluntary, that all data would remain anonymous, and that responses would be used exclusively for research purposes in full compliance with privacy regulations.



Figure 8.23: Participant's information sheet (in Italian) for the EEVAN experiment (link). The sheet has been adapted for practical and time constraints to both A and B group members.

Participants have then completed the two conditions in counterbalanced order according to group assignment. Group A has first completed the embodied condition C_1 and then the non-embodied condition C_2, whereas Group B has performed the reverse sequence. Each session has been conducted individually, and at the conclusion of both conditions participants have been asked to complete a short online post-experiment questionnaire (Google Form) [Figure 8.24], with questions about their participant code (P1-P10), previous experience with the robot, perceived

novelty of the encounter with the robot, boredom during interaction with the robot, boredom during interaction with the computer, overall enjoyment of the experience, comprehension of the experiment's purpose, and optional general comments.



Figure 8.24: Subjective questionnaire for the EEVAN experiment (link). See [Figure 8.25] for answers given by participants within the experiment.

8.3.3.6 Measures

Two complementary measures, objective and subjective, have been employed to assess the effect of embodiment on the affective impact of music.

The objective measure has consisted of the automatic recognition of facial expressions captured via webcam. In C_1, recordings have been obtained from the robot's eye-mounted camera, exploiting the robot's integrated face-tracking module; in C_2, recordings have been obtained from the host computer's webcam. The integrated software has analyzed each recording frame by frame, assigning an emotional label to each frame from a set of seven possible expressions (angry, disgust, fearful, happy, neutral sad, surprised). For the purposes of the study, the primary metric has been the total duration or percentage of frames classified as non-neutral expressions. This continuous measure has allowed for a direct within-subject comparison between C_1 and C_2.

The subjective measures have been collected through a short post-experiment questionnaire. The questionnaire has been conceived to gather evaluations along three dimensions: (i) which condition has been perceived as more engaging (robot, computer, or no difference); (ii) how

novel the first experience has felt, rated on a 7-point Likert scale; and (iii) how boring the second experience has been perceived, also rated on a 7-point Likert scale. Additional open comments have been allowed, providing qualitative insights to complement the objective measures.

8.3.3.7 Data Analysis

The data collected from the two experimental conditions have been analyzed through a combination of quantitative and qualitative methods.

For the objective component, each experimental session has generated a raw log file produced by the facial expression recognition software [Listing 8.2], containing a frame-by-frame classification of the participant's affective state throughout the 120-second interaction. Each file has been labeled according to participant code and condition (e.g., `P1_C1.txt`, `P1_C2.txt`) and subsequently exported into Microsoft Excel for further processing.

```
Data e ora: 2025-10-09 10:01:46
Periodo: Periodo_1
Angry: 0,00
Disgust: 0,67
Fearful: 0,00
Happy: 14,63
Neutral: 8,23
Sad: 6,47
Surprised: 0,00
------------------------------
Data e ora: 2025-10-09 10:01:46
Periodo: Periodo_2
Angry: 0,00
Disgust: 2,71
Fearful: 0,00
Happy: 15,40
Neutral: 5,32
Sad: 6,51
Surprised: 0,04
------------------------------
Data e ora: 2025-10-09 10:01:46
Periodo: Periodo_3
Angry: 0,05
Disgust: 13,54
Fearful: 0,00
Happy: 9,73
Neutral: 6,17
Sad: 0,52
Surprised: 0,00
------------------------------
```

```
Data e ora: 2025-10-09 10:01:46
Periodo: Periodo_4
Angry: 0,00
Disgust: 4,05
Fearful: 0,00
Happy: 0,36
Neutral: 23,08
Sad: 2,45
Surprised: 0,06
```

Listing 8.2: Example of the content of a raw log `.txt` file for a single experimental session (`P1_C1`).

Each log file is structured into four consecutive segments, corresponding to the four feedback moments provided by the system every 30 seconds during the listening phase [Section 8.3.2.6]. For each feedback interval, the software has recorded the total number of seconds in which the participant's facial expression has been classified under each of the seven emotion categories recognized by the program (Angry, Disgusted, Fearful, Happy, Neutral, Sad, and Surprised). This temporal subdivision has allowed for a detailed analysis of emotional dynamics over time, while ensuring that each participant contributed an equal total duration of 120 seconds per condition (4×30s).

For each participant and condition, the values of the four feedback segments have been aggregated to compute the total duration associated with each affective state. The overall duration has been then checked against the nominal 120-second reference, and a normalization factor has been calculated to correct any minor variations in total recorded time, according to the expression

$$\text{Scaling factor} = \frac{120}{\text{Total time reported (s)}} \tag{8.1}$$

This factor has been applied to all emotion values to obtain normalized data, ensuring comparability across participants and conditions.

From these normalized data, the main measure of emotional engagement has been derived as the percentage of non-neutral expressions, de-

fined as the proportion of frames classified as any emotion other than "Neutral". This measure has been computed using the formula

$$\%_{\text{non-neutral}} = \frac{(\text{Angry} + \text{Disgusted} + \text{Fearful} + \text{Happy} + \text{Sad} + \text{Surprised})}{120} \times 100 \tag{8.2}$$

This percentage has been obtained for each participant in both conditions. The individual difference between the embodied (C_1) and non-embodied (C_2) settings has been then computed as

$$\Delta = C_1 - C_2 \tag{8.3}$$

where positive values of Δ indicated a higher proportion of non-neutral expressions during the interaction with the robot.

Descriptive statistics (mean, median, standard deviation, minimum, and maximum) have been computed to summarize the distribution of C_1, C_2, and Δ across participants, and the normality of the Δ distribution has been assessed using the Shapiro–Wilk test. Based on its outcome, the appropriate inferential procedure has been selected: a paired-samples t-test in the case of compliance with the normality assumption, or a Wilcoxon signed-rank test as a non-parametric alternative to determine whether the observed difference between C_1 and C_2 has been statistically significant. Effect sizes have been then computed as Cohen's d for the t-test and as $r = Z/\sqrt{N}$ for the Wilcoxon test. In addition to the global measure of non-neutral expressions, an exploratory within-subject comparison has been conducted for each individual emotion category across conditions, using Wilcoxon tests with Bonferroni correction for multiple comparisons.

For the subjective component, participants' responses to the post-experiment questionnaire have been aggregated and analyzed in Excel. The categorical preference question (Robot / Computer / No difference) has been summarized in terms of frequency counts and analyzed using a chi-square (χ^2) test to evaluate whether the observed distribution differed significantly from a uniform one. The Likert-scale items assessing perceived novelty, boredom, appreciation, and comprehension have been analyzed using median, interquartile range, mean, and standard deviation, providing a descriptive overview of self-reported experience.

Finally, correlations between the objective engagement measure (Δ) and the subjective variables have been explored using Spearman's rank correlation coefficient (ρ), in order to examine potential associations between expressive behavior and self-reported affective engagement. In addition, open-ended responses have been reviewed qualitatively to identify recurring themes related to perceived naturalness, emotional involvement, and overall preference between the two conditions. These qualitative insights have been integrated with the quantitative findings to construct a comprehensive interpretation of the results.

By combining objective, subjective, and qualitative measures, the analysis aimed to provide a coherent understanding of the affective effects of embodiment, ultimately testing whether the embodied condition (C_1) elicited a significantly higher level of emotional engagement than the non-embodied condition (C_2).

All numerical computations and inferential tests in the analysis have been performed using Python-based statistical libraries (NumPy, Pandas, and SciPy), integrated within the ChatGPT analysis environment. Graphical visualizations (individual and aggregate bar plots, emotion-specific comparisons, and Likert-scale charts) have been generated using Matplotlib. Descriptive and summary tables were subsequently verified and formatted in Microsoft Excel.



Figure 8.25: Complete set of objective and subjective data retrieved from the experiment with EEVAN (link).

8.3.3.8 Results

The statistical analysis produced a coherent set of quantitative results concerning both the objective and subjective measures collected during the experiment (for the complete dataset, see [Figure 25]).

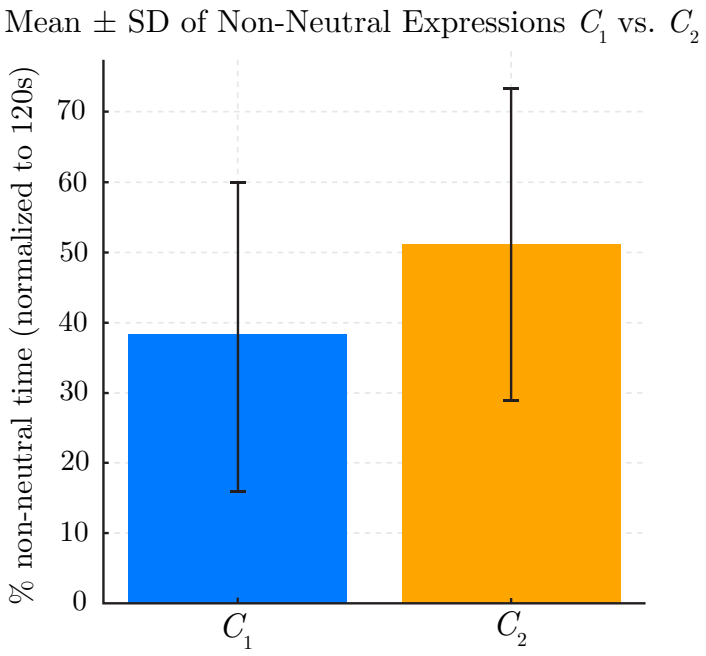


Figure 8.26: Mean percentage (±SD) of non-neutral facial expressions during the 120-second interaction in the embodied (■ = C_1) and non-embodied (■ = C_2) conditions. Error bars represent standard deviations.

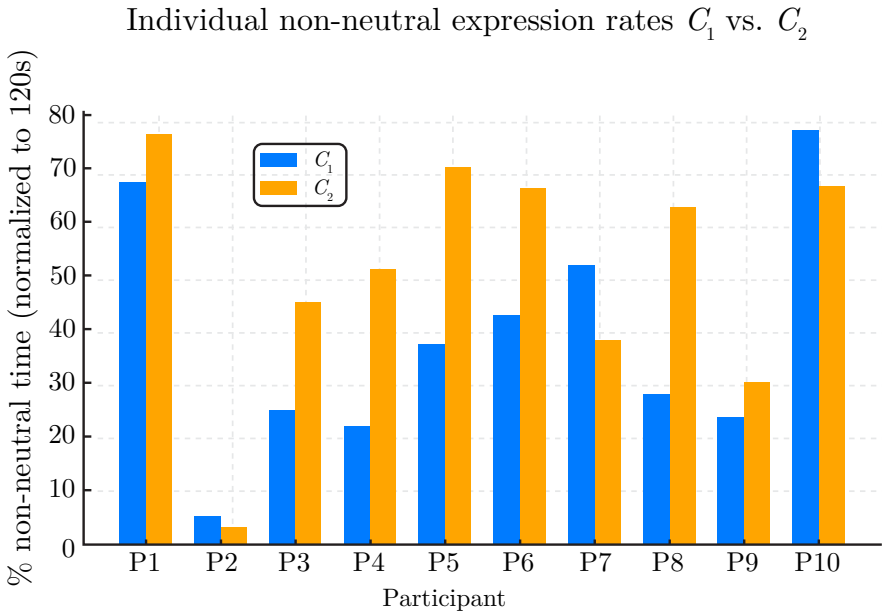


Figure 8.27: Normalized percentage of non-neutral facial expressions for each participant in the embodied (■ = C_1) and non-embodied (■ = C_2) conditions. Each pair of bars represents the within-subject comparison.

For the main quantitative variable —the normalized percentage of non-neutral expressions over the 120-second interaction period— descriptive statistics showed a mean of 38.02% ($SD = 22.19$) in the embodied condition (C_1) and a mean of 51.05% ($SD = 22.55$) in the non-embodied condition (C_2) [Figure 8.26]. The minimum and maximum values for C_1 ranged from 5.08% to 77.03%, whereas for C_2 they ranged from 2.90% to 76.37% [Figure 8.27]. The difference between the two conditions ($\Delta = C_1 - C_2$) exhibited a mean of -13.02% with a standard deviation of 17.87, a minimum of -34.98%, and a maximum of 14.01%.

The Shapiro–Wilk test for normality applied to the distribution of Δ yielded $W = 0.928$ and $p = 0.428$, indicating that the data did not significantly deviate from normality and therefore satisfied the assumption required for parametric testing. Consequently, a paired-samples t-test was performed to compare the two conditions. The test produced a statistic of $t = -2.304$ with a corresponding $p = 0.0467$, thus revealing a statistically significant difference between C_1 and C_2. The effect size, calculated as Cohen's $d = -0.729$, indicated a medium-to-large magnitude of difference in the proportion of non-neutral expressions between the two conditions.

To further explore the specific emotional dynamics underlying this global effect, a within-subject analysis was conducted for each individual emotion category using the Wilcoxon signed-rank test, with Bonferroni correction applied for multiple comparisons. The results indicated that none of the individual emotion differences survived the corrected significance threshold. The raw, uncorrected p-values ranged from 0.128 to 1.0, with the smallest observed for *Sad* ($W = 5.0$, $p = 0.1282$, $p_{bonf} = 0.7691$, $z = -2.2934$, $r = -0.7252$) and *Surprised* ($W = 8.0$, $p = 0.1614$, $p_{bonf} = 0.9686$, $z = -1.9876$, $r = -0.6285$). Other categories, including *Angry*, *Disgusted*, *Fearful*, and *Happy*, showed no statistically significant changes across conditions after correction, with corresponding effect sizes ranging between $r = -0.274$ and $r = -0.886$.

The subjective data obtained from the post-experiment questionnaire provided complementary information to the objective measures. Half of the participants (P1, P2, P4, P8, P10) had already interacted with the robot in previous sessions of the same experimental game, whereas the remaining half (P3, P5, P6, P7, P9) had seen the robot (e.g., in its

construction phase) but not actively interacted with it within the musical game [Figure 8.28].

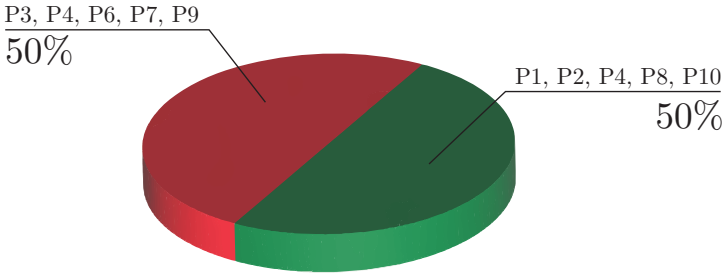


Figure 8.28: Distribution of participants according to their previous experience with the robot. Half of them (■ = P1, P2, P4, P8, P10) had already interacted with the robot in earlier sessions of the same experimental game, while the remaining half (■ = P3, P5, P6, P7, P9) were first-time users.

Participants' self-reports indicated a high perceived novelty of the interaction with the robot ($M = 5.9$, $SD = 0.88$) [Figure 8.30]. The perceived comprehension of the task and of the overall purpose of the experiment was also rated very positively ($M = 6.1$, $SD = 0.99$). Likewise, the general appreciation of the experience as a whole reached high values

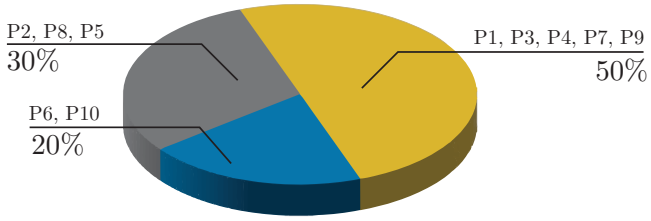


Figure 8.29: Descriptive preference of participants for "Robot" (■), "Computer" (■) or neither (■).

($M = 5.9$, $SD = 0.74$). Perceived boredom was relatively low in both modalities, with mean ratings of 2.8 ($SD = 1.93$) for the interaction with the robot (C_1) and 3.4 ($SD = 1.96$) for the interaction with the computer (C_2). The categorical preference data revealed no significant predominance for either interaction modality. The chi-square test yielded $\chi^2 = 0.6842$ with $p = 0.7103$, suggesting that the distribution of preferences (inferred from the boredom levels) among "Robot", "Computer",

and "No difference" did not significantly deviate from uniformity. Nevertheless, at a descriptive level, the majority of participants showed an implicit preference for the robot condition (C_1), while fewer favored the computer (C_2) or expressed no clear difference between the two [Figure 8.29].

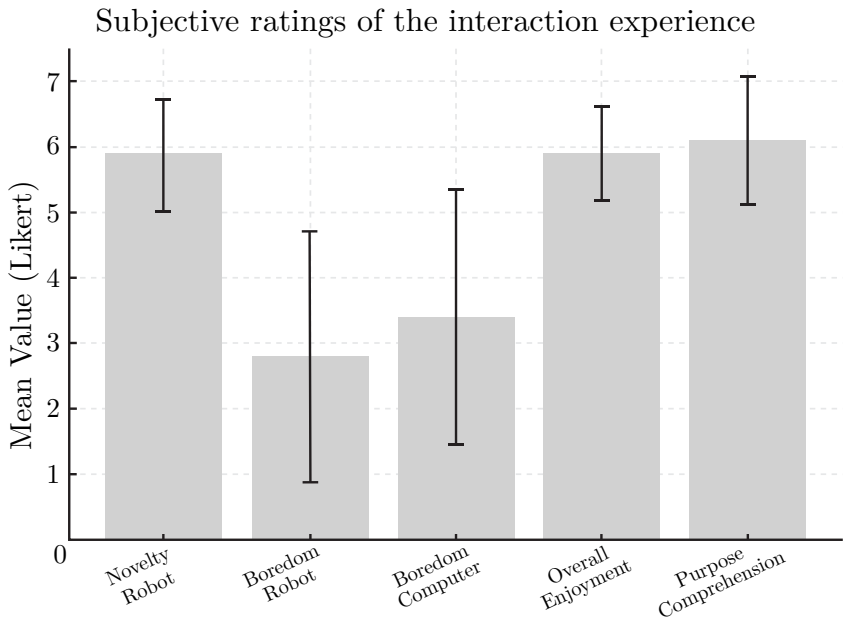


Figure 8.30: Mean (±SD) self-reported ratings on a 7-point Likert scale for novelty, comprehension, enjoyment, and boredom in both interaction modalities (robot and computer).

Subsequently, correlations between the objective engagement difference (Δ) and subjective ratings were examined using Spearman's rank correlation coefficient (ρ). None of the associations reached statistical significance. The strongest, though non-significant, correlation was observed between Δ and perceived boredom in the computer condition ($\rho = 0.539$, $p = 0.108$), followed by novelty ($\rho = 0.360$, $p = 0.308$). All other variables —boredom in the robot condition C_1 ($\rho = 0.234$, $p = 0.516$), appreciation ($\rho = 0.224$, $p = 0.534$), and comprehension ($\rho = -0.065$, $p = 0.859$)— showed weak or negligible associations.

Overall, the statistical results provided a complete quantitative account of the experimental findings, encompassing descriptive, inferential, and correlational analyses for both the objective facial-expression data and the subjective questionnaire responses (see [Figure 8.31] for an interactive version of this analysis).



Figure 8.31: Interactive version of the statistical analysis of the data collected through the EEVAN experiment (link).

Finally, this analysis has been integrated with the qualitative feedback provided in the open-ended section of the post-experiment questionnaire. Only three participants —P1, P6, and P10— left written comments, all of which have offered further insight into the subjective experience of interacting with the robot.

P1 emphasized the novelty and emotional resonance of the interaction, noting that «it is interesting to interact with a robot that tries to be emotional». P10 elaborated on a similar theme, highlighting the interplay between music and robotic embodiment as «very interesting». Finally, P6 expressed a desire for a more active interaction with the robot, aimed at understanding «why it makes you listen to certain musical tracks».

Overall, these open comments enrich the quantitative analysis by revealing an underlying curiosity and emotional engagement towards the robot's expressive and musical behavior.

8.3.3.9 Discussion

The results of the exploratory experiment provide a preliminary empirical assessment of the hypothesis that embodied media may elicit higher affective engagement than disembodied ones. The comparison between the embodied condition (C_1) —in which the musical routine was delivered by the humanoid robot— and the non-embodied condition (C_2) —in which the same routine was presented via a computer interface— revealed a statistically significant difference in the participants' expressive behavior, as captured through automatic facial expression recognition.

Quantitatively, the proportion of non-neutral facial expressions was significantly higher in the non-embodied condition than in the embodied

one, with a medium-to-large effect size ($d = -0.729$). This outcome, contrary to the expectation formulated in the alternative hypothesis (H_1), indicates that the humanoid embodiment did not amplify the visible expressivity of participants during the musical interaction. Rather, participants tended to display a greater degree of facial activation when the music was presented by the computer.

However, it must be noted that, in this experimental design, the analysis was restricted to the subset of all non-neutral expressions, considered as a general mark of "emotional activation". Yet neutrality itself can represent a genuine affective state. In this sense, the lower proportion of non-neutral expressions in C_1 may not necessarily imply lesser engagement, but a more regulated or introspective affective stance.

Several factors may account for this inversion of the expected pattern. First, the presence of the humanoid agent might have modulated the participants' emotional display through social or cognitive self-monitoring. In a setting characterized by proximity, anthropomorphic cues, and implied observation, individuals tend to adopt more controlled facial behavior —a phenomenon already documented in many HRI studies (see e.g., [Belkaid et al. 2021; Onnasch et al. 2022; Lombardi et al. 2023]). The robot's embodiment may thus have increased attentional focus and cognitive awareness, while simultaneously inhibiting spontaneous expressivity. Conversely, the computer interface, being disembodied and socially neutral, could have fostered a more relaxed state, allowing participants to manifest emotions more overtly and without evaluative pressure.

Moreover, at least for what concerns its quantitative dimension, consider that this experiment adopted —for practical reasons and design constraints— an implicit PBR-based view of emotions (see [Section 4.3] and [Section 6.1]), identifying it mostly with their measurable physiological-behavioral expressive correlates. In other words, the possibility that emotions may be internally experienced yet remain behaviorally unexpressed has been neglected within the experimental design. It cannot be therefore excluded that, especially under social or cognitive self-regulatory pressure, participants in the embodied condition did experience a comparable or even stronger emotional involvement, but expressed it through subtler or less externally visible cues, such as gaze direction, posture reg-

ulation, or internal attentional focus, which were not captured by the automated facial analysis.

To mitigate this inherent limitation, subjective self-report measures has been namely integrated into the experimental design. The inclusion of post-session questionnaires has been aimed precisely to complement the behavioral data with participants' introspective evaluations, thus offering a more balanced perspective on affective engagement. These subjective responses have revealed a pattern largely consistent with the qualitative expectations of the project: Despite the lower proportion of non-neutral facial expressions in C_1, participants have reported high levels of enjoyment, novelty, and comprehension, together with low boredom ratings in the embodied condition. Descriptively, a majority expressed a preference for the robot-based interaction, suggesting that the embodied experience, although less visibly expressive, was nonetheless perceived as more engaging and meaningful. It remains however unclear wether preference and boredom where directly correlated to the affective dimension of the interaction, and not to other aspects of it.

No significant correlations have conversely emerged between the objective engagement difference (Δ) and subjective ratings, implying that the two sets of measures capture distinct but complementary dimensions of affective response: the former primarily behavioral and outwardly expressive, the latter reflective and interpretative. Despite the higher average boredom in the computer condition, a weak positive trend indicated that participants who were more facially expressive during this setting also reported lower boredom, suggesting that spontaneous expressivity may reflect individual engagement when social inhibition is reduced.

At the same time, the limited interactivity of the EEVAN prototype must be considered as a relevant factor. Although the robot displayed basic expressive behaviors and synchronized playback, the musical routine remained largely unidirectional, offering minimal adaptation to the participant's affective feedback. This structural constraint may have reduced the sense of reciprocal exchange and emotional resonance that a more interactive system could evoke. Such an interpretation, for example, finds qualitative support in participant P6's comment, which explicitly expressed the desire for a more participatory understanding of the

robot's musical selections —a remark that underscores the importance of co-adaptive dynamics in affective musical encounters.

Taken together, these converging observations suggest that, when considered beyond the limitations of facial expressivity and PBR constraints, the results are not entirely at odds with the initial hypothesis (H_1). From a descriptive and phenomenological standpoint, participants appeared to attribute greater novelty, meaning, and enjoyment to the embodied condition, even if these qualities did not translate into overt expressive activation. In this sense, embodiment may have fostered a different quality of engagement —more introspective and relational than expressive— confirming the theoretical stance that affect in human-robot musical interaction cannot be reduced, and consequently studied, to surface-level behavioral manifestations.

Nevertheless, several methodological limitations of the present *exploratory* study must be acknowledged. The experimental sessions were conducted in a working environment, where residual social presence and situational distractions could not be entirely controlled. The study lacked a baseline or silence condition, which would have provided a neutral reference for affective activation. The counterbalancing of conditions was only partial, as the repetition of the same musical piece introduced a second-hearing effect: Participants who first experienced the computer condition might have undergone diminished surprise or emotional arousal upon encountering the same sequence through the robot. Moreover, the musical material itself was not affectively homogeneous throughout the 120-second routine, introducing potential confounds in emotional balance. Also, the six excerpts composing the sequence were not affectively or structurally equivalent, which may have produced additional variability in the temporal distribution of emotional responses independently of the experimental condition. Finally, the small sample size ($N = 10$) and the exploratory nature of the study inherently limit the statistical power and generalizability of the findings.

Future iterations of the experiment should therefore address these methodological constraints by incorporating larger and more diverse samples, affectively balanced (while not identical) musical stimuli, and enhanced interactive capabilities —such as real-time emotional mirroring, adaptive timing, or co-improvisational routines. Such improvements

would allow a more refined evaluation of how embodiment influences affective communication through music and, conversely, how musical interaction can modulate the experiential quality of human–robot engagement. Furthermore, future designs should also explicitly overcome the limitations of the PBR framework by integrating quantitative measures that capture emotion-related dimensions beyond the physiological or behavioral domain. These may include indicators of cognitive appraisal, attentional modulation, or propositional evaluation, thereby enabling a multidimensional assessment of affective experience that accounts for both expressive and reflective components. Finally, future studies should also consider the inherent interindividual variability of emotional response to music (and robots as well) [Section 8.1], which is shaped by personal dispositions, cultural background, and prior listening habits, acknowledging that such variability constitutes a crucial dimension of affective interpretation rather than a source of experimental noise.

In conclusion, while the quantitative data did not confirm a straightforward increase in expressive activation due to embodiment, the integrated analysis of objective, subjective, and qualitative measures reveals a more nuanced reality. The robot's presence reshaped, rather than amplified, the affective experience —transforming it from an externally expressive to a more internally reflective form of engagement. This shift, though subtle, aligns with the broader conceptual foundation of the EEVAN project, reinforcing the view of music as a potentially privileged, yet complex, medium for exploring the interplay between embodiment, emotion, and human-robot relationality.

8.4 Further Developments and Possible Applications

Although the EEVAN project was originally conceived as an exploratory prototype, its realization has already demonstrated the technical and conceptual feasibility of integrating musical and affective processes into human–robot interaction, particularly through the use of a humanoid platform —a shift that remains largely unexplored within the field of mu-

sical robotics, currently focused on non-humanoid models [Section 8.2.2]. Beyond its experimental dimension, the project has revealed the potential of music not merely as an auxiliary element for affective expression, but as an autonomous channel of communication —capable of mediating empathy, regulating interactional dynamics, and enhancing the intelligibility of robotic behavior. In this respect, EEVAN may be regarded as both a technological and epistemological achievement, insofar as it embodies, in operative form, the theoretical premises discussed throughout this dissertation, namely, that affective intelligence in machines cannot be reduced to mere physiological or behavioral imitation, but requires a relational framework that music, with its embodied and intersubjective structure, can effectively instantiate.

From a developmental standpoint, several directions are currently being pursued. The most immediate one concerns the refinement of EEVAN's perceptual and expressive modules, with the aim of improving the synchronization between auditory input, affective recognition, and multimodal feedback. The use of large-scale neural architectures for emotion classification —combined with advanced generative systems for adaptive sound design— could enable the robot to respond to human emotional cues in more nuanced and context-sensitive ways. Parallel to this, the integration of natural language processing frameworks, already under preliminary testing, will allow EEVAN to articulate its responses through dialogic forms of emotional reasoning, expanding its capacity for propositional engagement in the interaction. These developments are expected to support future iterations of the prototype, oriented toward semi-autonomous social performance in educational, therapeutic, and cultural settings.

Indeed, the conceptual and methodological framework developed through EEVAN finds a natural resonance within the broader objectives of *sustainable innovation* promoted by the Italian PNRR —particularly in Mission 4 ("Education and Research") and Mission 6 ("Health") [Section 8.3.1.1]. In the educational sphere, the project outlines potential avenues for interdisciplinary experimentation, in which music, robotics, and digital humanities can be integrated into new didactic formats. As a low-cost and open-source prototype, EEVAN lends itself to adaptation within schools, universities, and public laboratories, where it could serve

as a demonstrative tool for exploring emotional intelligence in machines and for fostering awareness of artificial affectivity among students. Such a framework would be particularly valuable in *inclusive* and *creative* education, where the robot's musical dimension might facilitate engagement, empathy, and curiosity across diverse cognitive and communicative profiles. In this respect, the project anticipates possible implementations that align with the PNRR's strategic vision of combining digital and humanistic competences in the same formative trajectory.

In parallel, EEVAN's modular architecture and open programming environment make it particularly suitable for activities of *coding literacy* within *collaborative learning* environments, encouraging students and teachers to approach robotics and artificial intelligence through direct creative experimentation. Some of these educational modules —focused on the customization of EEVAN's control scripts and interactive behaviors— have already been tested on a small scale within public schools in the Bari area, as part of pilot workshops on digital education and computational thinking, with great public engagement and very positive results.

In healthcare, the same principles suggest promising intersections with current practices of *music-based affective therapy* and *digital rehabilitation*. The project namely highlights how musical mediation through artificial agents could be employed as a complementary modality for emotional stimulation and relational support in care environments. Systems derived from EEVAN might eventually assist therapeutic interventions for neurodivergent or elderly patients, where rhythm, repetition, and affective synchrony may play a crucial role in maintaining cognitive and emotional balance, for example by facilitating relaxation, memory retrieval, or affective regulation. Within this framework, the robot's musical and expressive responsiveness may contribute to the design of non-invasive and patient-centered interfaces, supporting forms of emotional alignment and sensory engagement that are consistent with current trends in personalized and technologically assisted care. By avoiding the reduction of affective behavior to mere PB-based responses, and by incorporating music (and propositionality [Part II]) as a medium of empathetic modulation, such systems could also mitigate some of the classical adverse effects typically exerted by humanoid robots in clinical context

—most notably the *uncanny* [Problem 5]. Preliminary discussions in this direction have already taken place with the pain therapy and music therapy units of the Department of Medicine and Surgery at the University Hospital of Bari, where the potential of affective-musical robotics has been considered for future integration into experimental rehabilitation protocols.

In parallel, the project's reliance on open-source design and 3D-printing technologies naturally connects with ongoing research in low-cost prosthetics, suggesting the possibility of extending the same modular and reproducible logic to the development of personalized assistive devices. Such integration between affective robotics and self-printed prosthetics could enhance not only accessibility and customization in healthcare, but also the emotional and aesthetic dimensions of assistive design, aligning technological functionality with embodied experience.

Beyond education and health, EEVAN also holds relevance in the cultural sector, where public institutions increasingly experiment with forms of interactive and participatory technology. Museums, libraries, and cultural centers (mostly customers and partners of Cube Ltd.) have already shown interest in the system as a vehicle for experiential mediation, using affective robotics to reframe the visitor's encounter with digital heritage. Within these contexts, the robot's musical and expressive capacities function as interpretive interfaces, translating data and narratives into affective performances that encourage empathy, curiosity, and engagement. This convergence between technological embodiment and aesthetic communication directly aligns with the objectives of the PNRR's *Digitalization and Culture* mission, which envisions innovation not only as infrastructural modernization but as the creation of new forms of social and cultural interaction.

In this respect, further collaborations with Cube Ltd. are currently being explored to extend the project toward applications in the field of marketing. The idea is to equip EEVAN with an additional module that transforms affective data into visual representations or sound ("emotion visualization", "emotion sonification and musification"), returned to customers as tangible experiential outputs. Beyond its potential commercial implications, this development would also provide a testing ground for studying how affective and musical rendering can enhance data compre-

hension and engagement in the public discourse, bridging artistic expression, cognitive accessibility, and information design.

Directly related to this, EEVAN's framework could also foster (epistemological reflection on) new methodologies for basic research, particularly within public institutions. In recent years, techniques of *data sonification* and *musification* have been increasingly adopted across the sciences (see e.g., [Diaz-Merced et al. 2011; Newbold et al. 2015; Abbott et al. 2016; Hill et al. 2017; St George et al. 2018; Zanella et al. 2022; Braun et al. 2024; Butler et al. 2025]). Especially where traditional visual methodologies (graphs, diagrams, and imaging systems) prove intractable or inaccessible to visually impaired researchers, auditory representations allow data to be perceived, compared, and interpreted through the temporal and dynamic properties of sound, thereby transforming abstract numerical patterns into perceptually meaningful and experientially navigable structures [Feld 2015; Sawe et al. 2020; Bornmann 2024]. In this context, sonifying or musifying affective data would emerge as an original and fruitful methodology for studying emotions in psychology, neuroscience and medicine. Moreover, given the epistemic and cognitive import of affectivity in understanding and learning processes (see, e.g., [Elgin 2016] and [Problem 4(j)]), EEVAN's sonification and musification module could directly contribute to teaching and educational practices, particularly within the STEAM domain, where the pedagogical potential of music —as an emotional and relational form of communication— has already been acknowledged (see, e.g., [Eramo et al. 2025]).

Further applications also concern musicology, sound studies, acoustics, and electronic music. Within these domains, EEVAN provides a conceptual and experimental framework for investigating how affective computation can model musical meaning and performance dynamics. Its architecture allows for the real-time translation of emotional parameters into sonic structures, offering new tools for the analysis and simulation of expressive behavior in both human and artificial performers. In musicological and sound-studies research, such a system could assist in mapping the relationship between affective intent and acoustic form, while in the field of electronic and experimental music it may serve as a generative instrument, capable of producing compositions driven by emotional logic rather than algorithmic determinism. Through this convergence, EE-

VAN bridges analytical inquiry and creative practice, suggesting new perspectives for understanding and designing the affective dimension of sound.

The future evolution of the EEVAN project will move thus toward the design of open frameworks for affective robotics in the public sector, integrating technical guidelines, ethical protocols, and educational resources. The goal is to establish replicable models of human-robot interaction grounded in transparency, inclusivity, accessibility and sustainability. Within such frameworks, EEVAN's methodological contribution can serve as a reference for interdisciplinary cooperation between universities, research centers, and public administrations, advancing the broader mission of responsible and human-centered innovation.

The project, in general, exemplifies how a philosophical inquiry, when concretely implemented in technology, can generate a virtuous feedback between theoretical reflection, technical application and social transformation —between the history and epistemology of artificial emotions and the public experience of empathy in an increasingly digitized world.

Summary

This chapter has explored the potential of music to enhance machine affectivity and human-robot interaction. After some general philosophical remarks on the relationship between music and emotions and a brief overview of affective musical robotics, it introduced the EEVAN project as an attempt to concretely investigate such potential. EEVAN is a humanoid platform developed upon the open-source, 3D-printable robot *InMoov*. It was built and programmed in collaboration with a private company within the framework of the PNRR (Italy's *National Recovery and Resilience Plan*), which also funded the PhD scholarship. After describing the process of its mechanical and electrical assembly, programming, and software development, the design and results of an experiment were discussed, promisingly pointing to the possibility that music can serve as a privileged channel to sustain affective interaction between humans and robots. Finally, current and future refinements and applications of the project were outlined, with particular attention to those related to education, healthcare, culture, and relevant areas of public administration.

In this sense, EEVAN stands as a concrete realization of the historical and philosophical investigation carried out in this thesis, aimed at understanding the origins of artificial affectivity and the limitations of contemporary implementational practices and epistemological frameworks —limitations for which alternative models have been philosophically proposed toward a more empathetic and inclusive technology, with particular reference to humanoid robotics.

Conclusion

The present research has been conceived and developed within the broader aim of contributing to a comprehensive and philosophically grounded understanding of artificial affectivity. Its main objective has been to outline the historical, conceptual, and methodological conditions under which emotions can be meaningfully represented, modeled, and implemented in computational and robotic systems, while avoiding the reductionist tendencies that, often implicitly and unawarely, constrain this field. By situating the study of artificial emotions at the intersection of history, philosophy, and technological practice, the work has sought to provide the epistemological foundations for designing affective systems —particularly humanoid robots— capable of embodying a more inclusive and genuinely "human" conception of emotional engagement. This *main objective* (MO) [Introduction] —coherently with the framework of the doctoral program [Footnote 1]— has guided the inquiry throughout its different, yet deeply interconnected and interdependent stages, linking theoretical analysis with practical experimentation, while promoting a view of artificial emotions not as simple computational variables, but as complex evaluative processes embedded in social, cultural and normative contexts.

The dissertation has been articulated into three major parts, each addressing a distinct yet complementary level of investigation, through which MO has been progressively developed and substantiated.

[Part I] has offered the (at least to my knowledge) first systematic reconstruction of the historical-epistemological genealogy of artificial affectivity, tracing how emotions became legitimate computational objects within AI and robotics between the 1960s and the 1990s. This part has achieved three main research goals. First, it has identified the conceptual, methodological and cultural transitions that made the idea itself of CME possible; second, through a detailed analysis of early CME and AC models and their gradual implementation in physical robotic systems, it has reconstructed the progressive evolution of the conceptions of emotion underlying CMEs —from cognitive and representational views to subsymbolic, embodied, multimodal, and socially grounded accounts; third, it has identified the origins of some of the most relevant epistemic assumptions and philosophical premises that still inform current research and technical practices in the development of affective technologies.

[Part II] has advanced a critical reassessment of these assumptions and premises by addressing the persistence of a form of *physiological-behavioral reductionism* (PBR) in contemporary affective computing and social robotics, and by proposing a *propositionalist framework* for artificial emotions as a possible path to overcome it. This framework conceptualizes affects, in a non-reductionist fashion, as intentional and evaluative states possessing a cognitive-propositional structure, thereby providing a theoretical basis for designing affective architectures capable of generating, recognizing, and expressing both simple and complex emotions in contextually appropriate and socially intelligible ways.

[Part III] has finally translated and partly operationalized these theoretical, anti-reductionist, insights into an experimental platform —the EEVAN project— which explores the enhancement of robotic affectivity particularly through musical interaction. By integrating auditory and expressive modalities, EEVAN has demonstrated how, especially in humanoid robots, sound and music can function as aesthetic-affective mediators, extending the communicative and interpretive range of robotic emotions beyond merely facial and gestural channels.

Taken together, the three parts establish, with their achievements, a continuous and coherent argumentative trajectory that moves from the historical constitution of artificial emotions to their philosophical redefinition and practical realization within embodied and interactive systems, contributing to the overarching MO of establishing a unified, interdisciplinary framework for understanding and designing affective systems ca-

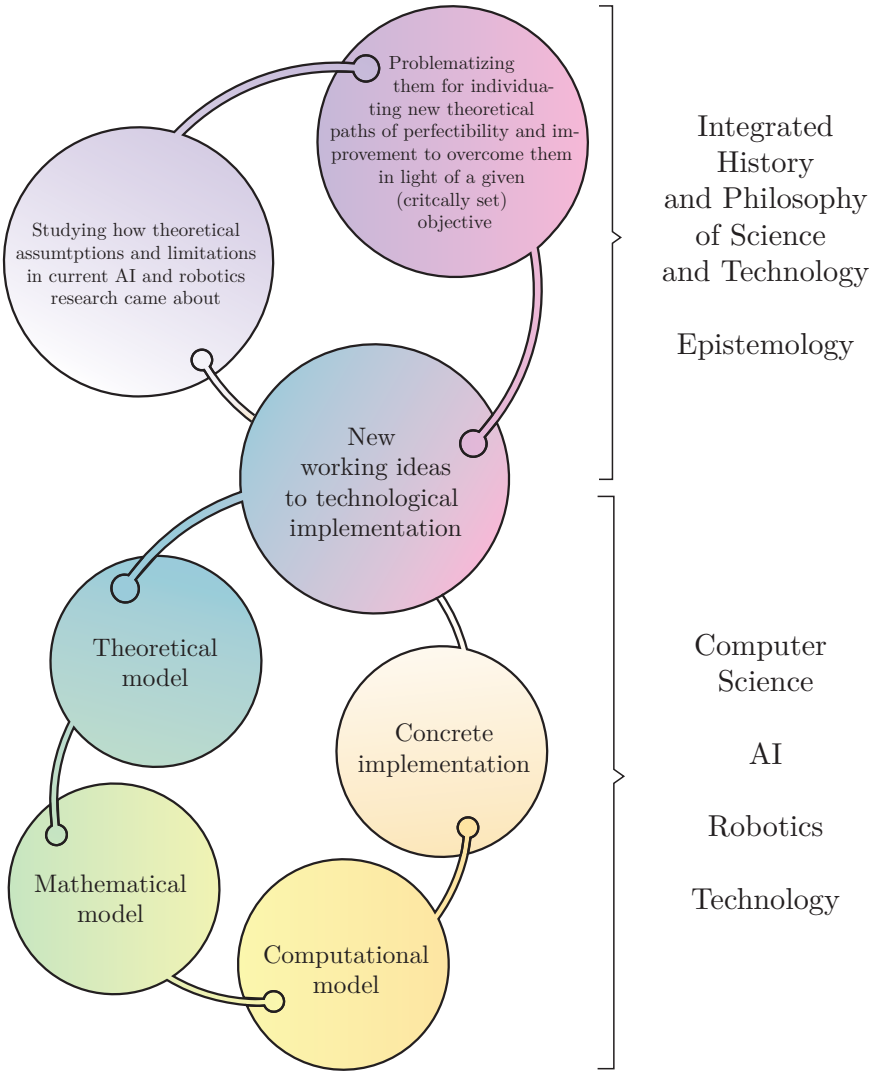


Figure 8.32: The integration of historical, philosophical and epistemological inquiry into the study of science and technology contributes to the identification of the conceptual and methodological tensions embedded in current research practices, which are then critically problematized to expose their limits and potential lines of transformation. From this reflective work, new theoretical insights emerge, which are articulated into frameworks that undergo successive stages of formalization (see e.g., [Durán 2018; Datteri and Schiaffonati 2019]), from the theoretical, then mathematical and finally computational model, to the concrete implementation of a technological device, where abstract hypotheses are materialized and experimentally tested. The outcomes of such realizations, in turn, generate new empirical and conceptual challenges that reinitiate the cycle of reflection and innovation, ensuring a continuous and self-corrective dialogue between knowledge and technology, critique and construction.

pable of more human-like, inclusive, and context-sensitive forms of emotional engagement. Beyond confirming the interdisciplinary and epistemological consistency of the work, these achievement have also demonstrated the methodological relevance of combining humanistic and scientific approaches, revealing how the *integrated historical-philosophical* take on artificial emotions —and artificial intelligence in general— remains today indispensable for understanding the epistemic (and ethical) limitations of current technological approaches, as well as the range of improvement and perfectibility from which new theoretical and practical solutions can emerge [Figure 1].

Building upon the theoretical, historical, and methodological outcomes of the present research, several lines of further investigation can be envisaged to consolidate and extend its results. Firstly, the propositionalist framework outlined in [Part II] is the first theoretical step toward a formalization of the model into a working computational design for affective architectures in robotics. This would entail the development of artificial systems capable of representing and processing emotions as evaluative–intentional states endowed with propositional content. The task requires translating and operationalizing the conceptual structure of propositional emotions into algorithmic models and decision architectures, identifying the minimal computational conditions under which such evaluative representations —*in coordination with* other PB-based or even *musical* modules— may arise and operate coherently within embodied agents.

A second perspective involves establishing, on the basis of the propositionalist framework, new experimental protocols for the systematic investigation of emotional attributions to robots in HRI and robot psychology. These protocols could be used to perform quantitative, reproducible and fine-tuned tests about how human subjects ascribe emotions to robots by directly assessing the role of cognitive —i.e., contextual, social, and propositional— structures in emotional attributions, particularly in the case of *complex affective states*. Establishing such a systematic methodology would allow the validation of the theoretical claims advanced in [Part II] through controlled and replicable experiments.

Third, consider that robots are presently used in science as *epistemic tools* for studying natural phenomena (see e.g., [Datteri and Schiaffonati

2019; Tamborini and Datteri 2022; Datteri and Schiaffonati 2023]), especially in biology, ethology and medicine[127]. These robots serve both a *simulative* and *stimulative* function (see e.g., [Datteri 2025]). In the first case, they reproduce selected features or mechanisms of a target system to generate and test explanatory hypotheses about its functioning; in the second, they operate as controllable sources of stimuli to elicit observable reactions in the empirical target system, allowing the study of behavioral or physiological responses under standardized conditions.

If so, (propositionalist or musical) affective robotic architectures may be used as epistemic models within psychology, neuroscience, and the cognitive sciences for testing hypotheses about human emotions —i.e., their inner causal mechanisms, nature and dynamics— both in a simulative and stimulative sense. Robotic systems may thus simulate specific cognitive-affective architectures or processes, allowing researchers to implement and evaluate competing theoretical models of emotion under controlled computational conditions, thereby assessing their explanatory adequacy and behavioral plausibility. On the other hand, they may function as manipulable, reproducible, standardized and ethically viable sources of affective stimuli, enabling the testing of hypotheses about affective perception, contagion, and empathy with a level of precision and repeatability difficult to achieve in human-human studies. Pursued along these epistemic lines, thus, robotics would not merely imitate emotional processes but actively contribute to their scientific understanding, transforming social robots into genuine epistemic tools for exploring the causal mechanisms, structures, nature and dynamics of emotion itself as a complex —and still controversial— psychophysiological process.

It should be noted, however, that the epistemic status of robots in science is still problematic. The main issues concern the validity and scope of the inferences that can be drawn from robotic models, the conditions under which they can legitimately represent or act as *surrogates* of their natural targets, and the methodological and normative criteria defining

[127]Notable examples are biorobots in evolutionary biology (see e.g., [Datteri 2021; Tamborini 2021; 2024a,b]) —as in the case of Long et al.'s [2006] *Madelaine*, a robotic plesiosaur built to experimentally test functional hypotheses about the evolution of aquatic locomotion in tetrapods; or, in medicine, Fujiso et al.'s [2018] *Swall-E*, a robotic simulation of the human swallowing process enabling non-invasive physiological investigation.

when simulation or stimulation can yield genuine scientific evidence. For these reasons, the use of social robots as epistemic tools in emotion research would also call for further research in the history and philosophy of science and epistemology. For example, it would require the development of a normative framework capable of defining the epistemic conditions under which robotic experiments can provide reliable and non-circular evidence about emotional phenomena —clarifying, in particular, the adequacy of the stimuli delivered, the relevance of the modeled processes to specific theories of emotion, and the criteria for transferring results from robotic to human contexts. Also, it would involve a historical reconstruction of the epistemic use of robots and machines as manipulable models for studying psychophysiological phenomena —which is something that has itself a long history, as seen in [Chapter 5].

Fourth, the historical reconstruction presented in [Part I] could be extended to cover the period from the early 2000s to the present, tracing the evolution of affective computing and robotics in the age of deep learning, large language models, and data-driven affect recognition and synthesis. Although such an extension risks becoming a mere chronicle of today's rapidly evolving technological developments—often too recent to be critically assessed with sufficient historiographical distance— such an extension would offer an even more comprehensive and profound genealogy of artificial affectivity, capturing the epistemic continuities and conceptual shifts linking early CMEs with contemporary practices in affective AI. Moreover, the historical reconstruction of artificial affectivity could be expanded by refining its methodological premises [Introduction] or by introducing new ones. A more culturally oriented or epistemologically oriented history could thus be (re)written, integrating novel historiographical tools and interdisciplinary perspectives suited to analyzing contemporary technoscientific production. Expanding the corpus of sources to include code repositories, design documents, experimental datasets, and corporate research outputs would make it possible to historicize affective AI not only as a scientific endeavor but also as a cultural and, above all, institutional phenomenon, thereby deepening the analysis of how emotions have emerged as computational objects from new and original points of view. In this respect, for instance, a valuable extension of the present research may consist in developing a *global history of*

affective technologies, attentive to the transnational circulation of ideas, models, and practices about artificial emotions and affective robotics across different scientific and cultural contexts. Such an approach would make it possible to trace how diverse conceptions of affect, intelligence, and sociality —emerging especially from Western and Japanese research traditions— have interacted and mutually influenced the design of affective computational systems and robots, revealing both shared epistemic frameworks and culturally specific understandings of emotional embodiment and human-robot communication.

Finally, another important development of the present research may concern the *ethical* dimension of artificial affectivity, to be addressed in continuity with its epistemological premises [Introduction]. Since every model of emotion implies specific conceptions of subjectivity, agency, and relationality, the ethical implications of affective technologies cannot be separated from the theoretical frameworks that sustain them. In particular, the propositionalist account developed in this dissertation — by grounding emotions in evaluative and intentional structures rather than in mere physiological or behavioral correlates— suggests a normative orientation that values interpretative depth, contextual sensitivity, and relational reciprocity over automatic expressivity or affective manipulation. Future research could therefore elaborate an *epistemology-dependent* ethics of artificial affectivity, clarifying how different conceptions of emotional intelligence translate into distinct moral and social consequences for human–machine coexistence. Such a perspective would help articulate criteria for the responsible design of affective systems, ensuring that the pursuit of emotional realism in machines does not entail the instrumentalization of affect, but rather contributes to more transparent, empathetic, and ethically aware forms of technological mediation.

Ultimately, this dissertation has shown that artificial emotions cannot be reduced to a merely computational or engineering problem, but must also be approached as a philosophical and epistemological challenge that calls for a profound reconsideration of what it means to model, interpret, and engage affectively with artificial systems. By tracing their historical formation and conceptual foundations, and by advancing an alternative, non-reductionist framework for their computational and robotic design, the dissertation has presented artificial affectivity as a valuable

instrument for deepening human self-understanding —particularly with regard to how we experience, express, and share emotions within an increasingly technologized world. In this broader sense, the study of artificial affectivity reveals not only a path toward more expressive and emotionally intelligent machines, but also a way of rethinking the place of emotion within technological life itself, fostering forms of *innovation* that are both intellectually and ethically *sustainable*, grounded in empathy, understanding, and the increasing recognition of our shared affective condition across natural and artificial domains.

Bibliography

Abbott, B. P., Abbott, R., Abbott, T. D., et al. (2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". *Physical Review Letters*, (116)6:061102.

Abelson, R. P. (1959). "Modes of Resolution of Belief Dilemmas". *Journal of Conflict Resolution*, (3)4:343–352.

— (1963). "Computer Simulation of "Hot Cognition"". In: *Psychology's New Frontier: Computer Simulation of Personality*. Tomkins, S. S. and Messick, S., eds. John Wiley & Sons, New York, pp. 277–298.

— (1981). "Constraint, Construal, and Cognitive Science". In: *Proceedings of the 3rd Annual Meeting of the Cognitive Science Society, 19-21 August, 1981, Berkeley, California*. Vol. 3.

— (1983). "Whatever Became of Consistency Theory?" *Personality And Social Psychology Bulletin*, (9):37–54.

— and Carroll, J. D. (1965). "Computer Simulation of Individual Belief Systems". *American Behavioral Scientist*, (8)9:24–30.

— and Rosenberg, M. J. (1958). "Symbolic Psycho-Logic: A Model of Attitudinal Cognition". *Behavioral Science*, (3):1–13.

Afzal, S., Ali Khan, H., Jalil Piran, M., et al. (2024). "A Comprehensive Survey on Affective Computing: Challenges, Trends, Applications, and Future Directions". *IEEE Access*, (12):96150–96168.

Ahmed, T. U. and Mishra, D. (2024). "FER-Pep: A Deep Learning Based Facial Emotion Recognition Framework for Humanoid Robot Pepper". In: *AI-HCII 2024: Proceedings of the 5th International Conference on Human-Computer Interaction, Held as a Part of the 26th HCI International Conference, 29-4 July 2024, Washington, DC*. Degen, H. and Ntoa, S., eds. Springer Nature, New York, pp. 199–215.

Ahn, H.-I. (2010). *Modeling and Analysis of Affective Influences on Human Experience, Prediction, Decision Making, and Behavior*. PhD thesis, Massachusetts Institute of Technology, Cambridge (MA).

Alexandre, G. R., Soares, J. M. and Thé, G. A. P. (2020). "Systematic Review of 3D Facial Expression Recognition Methods". *Pattern Recognition*, (100):107–108.

Allport, G. W. (1935). "Attitudes". In: *Handbook of Social Psychology*. Murchison, C., ed. Clark University Press, Worcester (MA), pp. 798–844.

— (1937). *Personality: A Psychological Interpretation*. Henry Holt and Co., New York.

Amari, S. (1967). "A Theory of Adaptive Pattern Classifiers". *IEEE Transactions on Electronic Computers* 3:299–307.

Anderson, J. A. (1977). "Neural Models with Cognitive Implications". *Basic Processes in Reading: Perception and Comprehension*: 27–90.

Anderson, J. R. (1976). *Language, Memory, and Thought*. Erlbaum, Hillsdale (NY).

— (1983). *The Architecture of Cognition*. Harvard University Press, Cambridge (MA).

—, Kline, P. J., and Beasley Jr., C. M. (1979). "A General Learning Theory and Its Application to Schema Abstraction". In: *Psychology of Learning and Motivation*. Bower, G. W., ed. Academic Press, New York, pp. 277–318.

Aoustin, Y., Chevallereau, C. and Laumond, J.-P. (2018). "Historical Perspective of Humanoid Robot Research in Europe". In: *Humanoid Robotics: A Reference*. Goswami, A. and Vadakkepat, P., eds. Springer, New York, pp. 3–8.

APA. (2022). American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing, Washington, DC.

Aquinas, T. (1922). *Summa Theologica*. Eng. tr. by Fathers of the English Dominican Province. Burns Oates and Washbourne, London.

Arabatzis, T. and Schickore, J. (2012). "Introduction: Ways of Integrating History and Philosophy of Science". *Perspectives on Science*, (20)4:395–408.

Arango-Muñoz, S. and Michaelian, K. (2014). "Epistemic Feelings and Epistemic Emotions (Focus Section)". *Philosophical Inquiries*, (2)1:97–122.

Aristotle. (1984a). "De Anima". In: *The Complete Works of Aristotle*. Barnes, J., ed. Eng. tr. by Roberts, W. R. Princeton University Press, Princeton (NJ), pp. 641–692.

— (1984b). "Poetics". In: *The Complete Works of Aristotle*. Barnes, J., ed. Eng. tr. by Roberts, W. R. Princeton University Press, Princeton (NJ), pp. 1448–1469.

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., et al. (1997). "Computational Modeling of Emotion: Explorations through the Anatomy and Physiology of Fear Conditioning". *Trends in Cognitive Sciences*, (1)1:28–34.

Arnold, M. B. (1960). *Emotion and Personality*. Columbia University Press, New York.

Ashok, A., Nguyen, L. and Berns, K. (2025). "A Glimpse into My World: Empathy Towards Emotional Robot Backstories at University". In: *ACM/IEEE 2025: Proceedings of the International Conference on Human-Robot Interaction, 4-6 March, 2025, Melbourne, Australia*. Johal, W. and Lemaignan, S., eds. IEEE, New York, pp. 1227–1231.

Asimov, I. (1950). *I, Robot*. Doubleday, New York.

Averill, J. R. (1990). "Inner Feelings, Works of the Flesh, the Beast Within, Diseases of the Mind, Driving Force, and Putting on a Show: Six Metaphors of Emotion and Their Theoretical Extensions." In: *Metaphors in the History of Psychology*. Leary, D. E., ed. Cambridge University Press, Cambridge, pp. 104–132.

Aviezer, H., Bentin, S., Dudarev, V., et al. (2011). "The Automaticity of Emotional Face-Context Integration". *Emotion*, (11)6:1406–1414.

—, Trope, Y., and Todorov, A. (2012). "Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions". *Science*, (338)6111:1225–1229.

Ax, A. F. (1953). "The Physiological Differentiation Between Fear and Anger in Humans". *Psychosomatic Medicine*, (15)5:433–442.

Axelsson, M., Spitale, M. and Gunes, H. (2024). "Robots as Mental Well-being Coaches: Design and Ethical Recommendations". *Journal of Human-Robot Interaction*, (13)2.

Ayer, A. J. (1936). *Language, Truth and Logic*. Victor Gollancz, London.

Aylett, R. S., Louchart, S., Dias, J., et al. (2005). "FearNot! – An Experiment in Emergent Narrative". In: *IVA 2005: Proceedings of the Fifth International Working Conference on Intelligent Virtual Agents, 12-14 September 2005, Kos, Greece*. Panayiotopoulos, T., Gratch, J., Aylett, R. S., et al., eds. Springer, Berlin, pp. 305–316.

Aymerich-Franch, L. and Ferrer, I. (2023). "Socially Assistive Robots' Deployment in Healthcare Settings: A Global Perspective". *International Journal of Humanoid Robotics*, (20)01:2350002.

Baars, B. J. (1986). *The Cognitive Revolution in Psychology*. The Guilford Press, New York-London.

— (1993). *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge.

Bach, C. P. E. (1753). *Versuch über die wahre Art das Clavier zu spielen*. Volume 1. Christian Friedrich Henning, Berlin.

Baker, D. B., ed. (2012). *The Oxford Handbook of the History of Psychology: Global Perspectives*. Oxford University Press, Oxford.

Balkwill, L.-L. and Thompson, W. F. (1999). "A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues". *Music Perception*, (17)1:43–64.

Ballard, B. S. (2020). "Content and the Fittingness of Emotion". *The Philosophical Quarterly*, (71)4.

Ballard, D. H., Hayhoe, M. M. and Pelz, J. B. (1995). "Memory Representations in Natural Tasks". *Journal of Cognitive Neuroscience*, (7)1:66–80.

Banda, N., Engelbrecht, A. and Robinson, P. (2015). "Feature Reduction for Dimensional Emotion Recognition in Human-Robot Interaction". In: *SSCI 2015: Proceedings of the IEEE Symposium Series on Computational Intelligence, 7-10 December 2015, Cape Town, South Africa*. IEEE, ed. Curran Associates, Inc., Morehouse, pp. 803–810.

Banks, J. (2020). "Theory of Mind in Social Robots: Replication of Five Established Human Tests". *International Journal of Social Robotics*, (12)2:403–414.

Bard, P. (1928). "A Diencephalic Mechanism for the Expression of Rage with Special Reference to the Sympathetic Nervous System". *American Journal of Physiology-Legacy Content*, (84)3:490–515.

— (1934). "Emotion: I. The Neuro-humoral Basis of Emotional Reactions". In: *The Foundations of Experimental Psychology*. Murchison, C., ed. Clark University Press, Worchester (MA), pp. 449–487.

— and Mountcastle, V. B. (1948). "Some Forebrain Mechanisms Involved in Expression of Rage with Special Reference to Suppression of Angry Behavior". *Research Publications-Association for Research in Nervous and Mental Disease*, (27):362–404.

Barlassina, L. and Gordon, R. M. (2017). "Folk Psychology as Mental Simulation". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Summer 2017. Metaphysics Research Lab, Stanford University.

Baron-Cohen, S., Leslie, A. M. and Frith, U. (1985). "Does the Autistic Child Have a "Theory of Mind"?" *Cognition*, (21)1:37–46.

Barrett, L. F. (2006). "Are Emotions Natural Kinds?" *Perspectives on Psychological Science*, (1):28–58.

—, Adolphs, R., Marsella, S., et al. (2019). "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements". *Psychological Science in the Public Interest*, (20)1:1–68.

Barsalou, L. W. (1999). "Perceptual Symbol Systems". *Behavioral and Brain Sciences*, (22)4:577–660.

— (2008). "Grounded Cognition". *Annual Review of Psychology*, (59)1:617–645.

Bartneck, C. (2023). "Godspeed Questionnaire Series: Translations and Usage". In: *International Handbook of Behavioral Health Assessment*. Krägeloh, C. U., Alyami, M. and Medvedev, O. N., eds. Springer, New York, pp. 1–35.

—, Belpaeme, T., Eyssel, F., et al. (2024). *Human-Robot Interaction: An Introduction*. Cambridge University Press, Cambridge (MA).

—, Croft, E., and Kulic, D. (2017). "Measuring the Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety of Robots". In: *Metrics for HRI Workshop, Technical Report*, vol. 471, pp. 37–44.

Bastiaansen, J. A. C. J., Thioux, M. and Keysers, C. (2009). "Evidence for Mirror Systems in Emotions". *Philosophical Transactions of the Royal Society B: Biological Sciences*, (364)1528:2391–2404.

Bates, J. (1992). *Building Emotional Agents*. Technical Report CM UC S-92-143. School of Computer Science, Carnegie Mellon University.

Beaudoin, L. P. (1994). *Goal Processing in Autonomous Agents*. PhD thesis, University of Birmingham, Birminghan (UK).

Bechterew, W. (1885). "Ueber Zwangsbewegungen bei Zerstörung der Hirnrinde". *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, (101)3:473–483.

— (1887). "Die Bedeutung der Sehhügel auf Grund von experimentellen und pathologischen Daten". *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, (110)2:322–365. Springer.

Becker, D. V., Unkelbach, C. and Fiedler, K. (2021). *Inference in Social Cognition*. in *Oxford Research Encyclopedia of Psychology*. https://oxfordre.com/psychology/view/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-834 (Accessed: October 3, 2025).

Becker-Asano, C. (2008). *WASABI: Affect Simulation for Agents with Believable Interactivity*. IOS Press, Amsterdam.

Beer, R. D. (2000). "Dynamical Approaches to Cognitive Science". *Trends in Cognitive Sciences*, (4)3:91–99.

Beilock, S. L. (2008). "Beyond the Playing Field: Sport Psychology Meets Embodied Cognition". *International Review of Sport and Exercise Psychology*, (1)1:19–30.

— and Holt, L. E. (2006). "Expertise and Its Embodiment: Examining the Impact of Sensorimotor Skill Expertise on the Representation of Action-Related Text". *Psychonomic Bulletin & Review*, (13)4:694–701.

Belavkin, R. (2001). "The Role of Emotion in Problem Solving". In: *AISB 2001: Proceedings of the Symposium on Emotion, Cognition and Affective Computing, 10-12 April 2001, York (UK)*. Johnson, C. G., ed. AISB Press, University of York, York (UK), pp. 49–57.

Belkaid, M., Kompatsiari, K., De Tommaso, D., et al. (2021). "Mutual Gaze with a Robot Affects Human Neural Activity and Delays Decision-Making Processes". *Science Robotics*, (6)58:eabc5044.

Bell, G. (2018). "Making Life: A Brief History of Human-Robot Interaction". *Consumption Markets & Culture*, (21)1:22–41.

Benbaji, H. (2013). "How Is Recalcitrant Emotion Possible?" *Australasian Journal of Philosophy*, (91)3:577–599.

Benesh, M. and Weiner, B. (1982). "On Emotion and Motivation: From the Notebooks of Fritz Heider". *American Psychologist*, (37)8:887–895.

Benn, C. (2025). "The Moral Uncanny Valley". *Philosophy & Technology*, (38)2:37.

Bennett, C. C. and Šabanović, S. (2015). "The Effects of Culture and Context on Perceptions of Robotic Facial Expressions". *Interaction Studies*, (16)2:272–302.

Bentley, J. L. (1975). "Multidimensional Binary Search Trees Used for Associative Searching". *Communications of the ACM*, (18)9:509–517.

Bergson, H. (1889). *Essai sur les données immédiates de la conscience*. Félix Alcan, Paris.

Bermúdez, J. and Cahen, A. (2024). "Nonconceptual Mental Content". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Fall 2024. Metaphysics Research Lab, Stanford University.

Bernard, C. (1858). *Leçons sur la physiologie et la pathologie du système nerveux*. J.-B. Baillière et fils, Paris.

Bernays, P. and Schönfinkel, M. (Dec. 1928). "Zum Entscheidungsproblem der mathematischen Logik". *Mathematische Annalen*, (99)1:342–372.

Bernhardt, B. C. and Singer, T. (2012). "The Neural Basis of Empathy". *Annual Review of Neuroscience*, (35):1–23.

Berrios, R. (2019). "What Is Complex/Emotional About Emotional Complexity?" *Frontiers in Psychology*, (10):1606.

Bickle, J. (2009). *The Oxford Handbook of Philosophy and Neuroscience*. Oxford University Press, Oxford.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Blood, A. J. and Zatorre, R. J. (2001). "Intensely Pleasurable Responses to Music Correlate with Activity in Brain Regions Implicated in Reward and Emotion". *Proceedings of the National Academy of Sciences of the United States of America*, (98)20:11818–11823.

Blumberg, B. (1997). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, Massachusetts Institute of Technology, Cambridge (MA).

Boada, J. P., Maestre, B. R. and Genís, C. T. (2021). "The Ethical Issues of Social Assistive Robotics: A Critical Literature Review". *Technology in Society*, (67):101726.

Boden, M. A. (2008). *Mind as Machine: A History of Cognitive Science*. Oxford University Press, Oxford.

Boethius, A. M. S. (1990). *De institutione musica*. Marzi, G. (ed.). Istituto italiano per la storia della musica, Rome.

Bogodistov, Y. and Dost, F. (2017). "Proximity Begins with a Smile, But Which One? Associating Non-duchenne Smiles with Higher Psychological Distance". *Frontiers in Psychology*, (8):1374.

Bohus, K. A., Cesana-Arlotti, N., Martín-Salguero, A., et al. (2023). "The Scope and Role of Deduction in Infant Cognition". *Current Biology*, (33)18:4014–4020.

Bonard, C. (2024). "Emotions Represent Evaluative Properties Unconsciously". *Erkenntnis*.

Bornmann, L. (2024). “The Sound of Science”. *EMBO reports*, (25)9:3743–3747.

Bortolotti, L. (2022). “Delusion”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Summer 2022. Metaphysics Research Lab, Stanford University.

Bower, G. H. and Cohen, P. R. (1982). “Emotional Influences in Memory and Thinking: Data and Theory”. In: *Affect and Cognition: The 17th Annual Carnegie Symposium on Cognition*. Clark, M. S. and Fiske, S. T., eds. Erlbaum, Hillsdale (NY), pp. 291–331.

Bowling, D. L. (2013). “A Vocal Basis for the Affective Character of Musical Mode in Melody”. *Frontiers in Psychology*, (4):464.

Brady, M. S. (2009). “The Irrationality of Recalcitrant Emotions”. *Philosophical studies*, (145)3:413–430.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge (MA).

Braun, R., Tfirn, M. and Ford, R. M. (2024). “Listening to Life: Sonification for Enhancing Discovery in Biological Research”. *Biotechnology and Bioengineering*, (121)10:3009–3019.

Breazeal, C. L. (2000). *Sociable Machines: Expressive Social Exchange between Humans and Robots*. PhD thesis, Massachusetts Institute of Technology, Cambridge (MA).

— (2001). “Emotive Qualities in Robot Speech”. In: *IEEE/RSJ 2001: Proceedings of the IEEE International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium, 29 October - 03 November 2001, Maui, HI, USA*. IEEE, ed. Vol. 3. IEEE, New York, pp. 1388–1394.

— (2002). *Designing Sociable Robots*. MIT Press, Cambridge (MA).

— (2003a). “Emotion and Sociable Humanoid Robots”. *International Journal of Human-Computer Studies*, (59)1-2:119–155.

— (2003b). “Toward Sociable Robots”. *Robotics and Autonomous Systems*, (42)3-4:167–175.

— and Aryananda, L. (2002). “Recognition of Affective Communicative Intent in Robot-Directed Speech”. *Autonomous Robots*, (12)1:83–104.

— and Brooks, R. A. (2005). “Robot Emotion: A Functional Perspective”. In: *Who Needs Emotions? The Brain Meets the Robot*. Fellous, J.-M. and Arbib, M. A., eds. Oxford University Press, Oxford, pp. 271–310.

—, Dautenhahn, K., and Kanda, T. (2016). “Social Robotics”. In: *Springer Handbook of Robotics*. Bruno, S. and Khatib, O., eds. Springer, New York, pp. 1935–1972.

— and Scassellati, B. (1998). “Infant-Like Social Interactions Between a Robot and a Human Caretaker”. *Adaptive Behavior*, (8)1:49–74.

—, Scassellati, B., and Suck Cho, H. (1999). “How to Build Robots that Make Friends and Influence People”. In: *IEEE/RSJ 1999: Proceedings of the International Conference on Intelligent Robots and System, 17-21 October, 1999, Kyongju, Korea*. IEEE, ed. IEEE, New York, pp. 858–863.

— and Velasquez, J. (1998). “Toward Teaching a Robot “Infant” Using Emotive Communication Acts”. In: *Socially Situated Intelligence: Papers from the 1998 Simulated Adaptive Behavior (SAB) Workshop, August, 1998 University of Zurich*. Dautenhahn, K. and Edmonds, B., eds. Technical Report CPM-98-38, Centre for Policy Modelling, Manchester Metropolitan University, pp. 101–112.

Breiman, L., Friedman, J. H., Olshen, R. A., et al. (1984). *Classification and Regression Trees*. Wadsworth, Belmont (CA).

Brentano, F. (1874). *Psychologie vom empirischen Standpunkte*. Duncker & Humblot, Leipzig.

Bretan, M. and Weinberg, G. (2016). “A Survey of Robotic Musicianship”. *Communications of the ACM*, (59)5:100–109.

Bricker, A. M. (2020). “The Neural and Cognitive Mechanisms of Knowledge Attribution: An EEG Study”. *Cognition*, (203):104412.

Broadbent, D. E. (1958). *Perception and Communication*. Pergamon Press, Oxford.

Brooks, R. A. (1986). “A Robust Layered Control System for a Mobile Robot”. *IEEE Journal of Robotics and Automation*, (RA-2):14–23.

— (1989). “A Robot That Walks; Emergent Behaviors from a Carefully Evolved Network”. *Neural Computation*, (1)2:253–262.

— (1990a). “Challenges for Complete Creature Architectures”. In: *From Animals to Animats 1. Proceedings of the First International Conference on Simulation of Adapative Behavior, 24-28 September, 1990, Paris*. Wilson, S. W. and Meyer, J.-A., eds. MIT Press, Cambridge (MA), pp. 434–443.

— (1990b). “Elephants Don’t Play Chess”. *Robotics and Autonomous Systems*, (6)1-2:3–15.

— (1991a). “How to Build Complete Creatures Rather than Isolated Cognitive Simulators”. In: *Architectures for Intelligence. Proceedings of the 22nd Carnegie Mellon Symposium on Cognition*. Van Lehn, K., ed. Psychology Press, New York, pp. 225–240.

— (1991b). “Intelligence Without Reason”. In: *IJCAI 1991: Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, 24-30 August, 1991*. Reiter, R. and Mylopoulos, J., eds. Morgan Kaufmann Publishers, San Francisco, pp. 569–595.

— (1991c). “Intelligence Without Representation”. *Artificial Intelligence*, (47)1-3:139–159.

— (1997). “The Cog Project”. *Journal of the Robotics Society of Japan*, (15)7:968–970.

— (1999). *Cambrian Intelligence: The Early History of the New AI*. MIT Press, Cambridge (MA).

— and Breazeal, C. L. (1998). “Alternative Essences of Intelligence – Lessons from Embodied AI”. In: *AAAI 1998: Proceedings of the Fifteenth National Conference on Artificial Intelligence, 26-30 July, 1998, Madison (WI)*. Rich, C. and Mostow, J., eds. AAAI Press-MIT Press, Cambridge (MA), pp. 961–968.

—, Breazeal, C. L., Marjanović, M., et al. (1998). “The Cog Project: Building a Humanoid Robot”. In: *CMAA 1998: Proceedings of the International Workshop on Computation for Metaphors, Analogy, and Agents, 6-10 April, 1998, Aizu, Japan*. Nehaniv, C. L., ed. Springer, New York, pp. 52–87.

— and Stein, L. A. (1994). “Building Brains for Bodies”. *Autonomous Robots*, (1)1:7–25.

Brown, C. C., Raio, C. M. and Neta, M. (2017). “Cortisol Responses Enhance Negative Valence Perception for Ambiguous Facial Expressions”. *Scientific Reports*, (7):15107.

Brown, S., Martinez, M. J. and Parsons, L. M. (2004). “Passive Music Listening Spontaneously Engages Limbic and Paralimbic Systems”. *Neuroreport*, (15)13:2033–2037.

Bruner, J. (1991). “The Narrative Construction of Reality”. *Critical inquiry*, (18)1:1–21.

Bryson, A. E. and Ho, Y.-C. (1969). *Applied Optimal Control: Optimization, Estimation and Control*. Halsted Press Book, New York.

Buchner, A. (1960). *Mechanical Musical Intruments*. Batchworth Press, London.

Buck, R., Losow, J. I., Murphy, M. M., et al. (1992). “Social Facilitation and Inhibition of Emotional Expression and Communication”. *Journal of Personality and Social Psychology*, (63)6:962–698.

Bui, T. (2002). “Parlee: An Adaptive Plan Based Event Appraisal Model of Emotions”. In: *KI 2002: Proceedings of the 25thAnnual German Conference on Artificial Intelligence, 16–20 September 2002, Aachen, Germany*. Mertsching, B., ed. Springer, Berlin, pp. 129–143.

Burge, T. (2018). “Iconic Representation: Maps, Pictures, and Perception”. In: *The Map and the Territory*. Wuppuluri, S. and Doria, F. A., eds. Springer International Publishing, Cham, pp. 79–100.

Burton, R. (1621). *The Anatomy of Melancholy*. John Lichfield and James Short, Oxford.

Butler, E. A., Egloff, B., Wilhelm, F. H., et al. (2003). “The Social Consequences of Expressive Suppression”. *Emotion*, (3)1:48–67.

Butler, J., Brown, M., Prokocimer, P., et al. (2025). “The Role of Cardiac Acoustic Biomarkers in Monitoring Patients with Heart Failure: A Systematic Literature Review”. *ESC Heart Failure*, (12)2:980–997.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., et al. (2000). “The Psychophysiology of Emotion”. In: *Handbook of Emotions*. Lewis, M. and Haviland, J. M., eds. Guilford Press, New York, pp. 173–191.

Cahn, J. E. (1990). “The Generation of Affect in Synthesized Speech”. *Journal of the American voice I/O society*, (8)1:1–10.

Campeggiani, P. (2023). *Theories of Emotion: Expressing, Feeling, Acting*. Bloomsbury Publishing, London.

Cãnamero, L. and Avila-García, O. (2007). "A Bottom-Up Investigation of Emotional Modulation in Competitive Scenarios". In: *ACII 2007: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction, 12-14 September 2007, Lisbon, Portugal*. Paiva, A. C. R., Prada, R. and Picard, R. W., eds. Springer, Berlin, pp. 398–409.

Cannon, W. B. (1915). *Bodily Changes in Pain, Hunger, Fear, and Rage: An Account of Recent Researches into the Function of Emotional Excitement*. D. Appleton and Company, New York-London.

— (1927). "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory". *The American journal of psychology*, (39)1/4:106–124.

— (1928). "Neural Basis for Emotion Expression". In: *Feeling and Emotions. The Wittenberg Symposium.* Proceedings of the Wittenberg Symposium on Feelings and Emotions. Wittenberg College, Springfield (OH), October 19-23, 1927. Reymert, M. L., ed. Clark University Press, Worchester (MA), pp. 257–269.

— (1929a). *Bodily Changes in Pain, Hunger, Fear, and Rage: An Account of Recent Researches into the Function of Emotional Excitement*. D. Appleton and Company, New York-London.

— (1929b). "Organization for Physiological Homeostasis". *Physiological Reviews*, (9)3:399–431.

— (1929c). "Voodoo Death". *The Medical Review of Reviews*, (3)1:1–10.

— (1932). *The Wisdom of the Body*. W. W. Norton & Company.

— and Britton, S. W. (1925). "Studies on the Conditions of Activity in Endocrine Glands: Xv. Pseudaffective Medulliadrenal Secretion". *American Journal of Physiology - Legacy Content*, (72)2:283–294.

—, Lewis, J. T., and Britton, S. W. (1927). "The Dispensability of the Sympathetic Division of the Autonomic Nervous System". *The Boston Medical and Surgical Journal*, (197)13:514–515.

Čapek, K. (1920). *Rossumovi univerzální roboti (R.U.R.)* Eng. tr. by Selver P. and Playfair N. (1923). *R.U.R. (Rossum's Universal Robots): A Fantastic Melodrama*. Doubleday Page & Co., Garden City, (NY). Aventinum, Prague.

Carnap, R. (1928). *Der Logische Aufbau der Welt*. Weltkreis-Verlag, Berlin-Schlachtensee.

— (1932). "Überwindung der Metaphysik durch logische Analyse der Sprache". *Erkenntnis*, (2)1:219–241.

Cassirer, E. (1923). *Philosophie der symbolischen Formen*. Bruno Cassirer Verlag, Berlin.

Cavallo, F., Semeraro, F., Fiorini, L., et al. (2018). "Emotion Modelling for Social Robotics Applications: A Review". *Journal of Bionic Engineering*, (15):185–203.

Celeghin, A., Gelder, B. de and Tamietto, M. (2015). "From Affective Blindsight to Emotional Consciousness". *Consciousness and Cognition*, (36):414–425.

Cesana-Arlotti, N., Martin, A., Teglas, E., et al. (2018). "Precursors of Logical Reasoning in Preverbal Human Infants". *Science*, (359)6381:1263–1266.

Challacombe, B. J., Khan, M. S., Murphy, D., et al. (2006). "The History of Robotics in Urology". *World Journal of Urology*, (24):120–127.

Chang, H. (2007). "Scientific Progress: Beyond Foundationalism and Coherentism". *Royal Institute of Philosophy Supplements*, (61):1–20.

Chapuis, A. and Droz, E. (1949). *Les automates: figures artificielles d'hommes et d'animaux: histoire et technique*. Eng. tr. by Chapuis, A. and Droz, E. (1958). *Automata: A Historical and Technological Study*. Editions du Griffon, Neuchâtel. Editions du griffon, Neuchatel.

Charcot, J.-M. (1872). *Leçons sur les maladies du système nerveux faites à la Salpêtrière*. Adrien Delahaye, Paris.

Charpentier, M.-A. (1971). *Règles de composition*. University of Iowa Press, Iowa City.

Charton, É. d. M. (1882). "Vaucanson". *Le Magasin pittoresque*, (50):121–122.

Cheetham, M., Suter, P. and Jäncke, L. (2011). "The Human Likeness Dimension of the "Uncanny Valley Hypothesis": Behavioral and Functional MRI Findings". *Frontiers in Human Neuroscience*, (5).

Chen, Y.-H., Ceccarelli, M. and Yan, H.-S. (2018). "A Historical Study and Mechanical Classification of Ancient Music-Playing Automata". *Mechanism and Machine Theory*, (121):273–285.

Chimisso, C. (2024). "Historical Epistemology: A Broader and More Complex View". In: *The Oxford Handbook of Modern French Philosophy*. Sinclair, M. and Whistler, D., eds. Oxford University Press, pp. 201–218.

Church, A. (1936). "An Unsolvable Problem of Elementary Number Theory". *American Journal of ;athematics*, (58)2:345–363.

Churchland, P., Ramachandran, V. and Sejnowski, T. (1994). "A Critique of Pure Vision". In: *Large-Scale Neuronal Theories of the Brain*. Koch, C. and Davis, J., eds. MIT Press, Cambridge (MA), pp. 23–60.

Churchland, P. M. (1981). "Eliminative Materialism and the Propositional Attitudes". *The Journal of Philosophy*, (78)2:67–90.

Clark, A. (1998). *Being there: putting brain, body, and world together again*. Press, Cambridge (MA).

— and Chalmers, D. (1998). "The Extended Mind". *Analysis*, (58)1:7–19.

Clark, M. S. (1982). "A Role for Arousal in the Link Between Feeling States, Judgments, and Behavior". In: *Affect and Cognition: The 17th Annual Carnegie Symposium On Cognition*. Clark, M. S. and Fiske, S. T., eds. Erlbaum, Hillsdale (NY), pp. 263–289.

Clippinger, J. H. J. (1977). *Meaning and Discourse: A Computer Model of Psychoanalytic Speech and Cognition*. Johns Hopkins University Press, Baltimore.

Cochran, R., Lee, F. and Chown, E. (2006). "Modeling Emotion: Arousal's Impact on Memory". In: *CogSci 2006: Proceedings of the 28thAnnual Conference of the Cognitive Science Society, 26–29 July 2006, Vancouver, Canada*. Sun, R. and Miyake, N., eds. CSS Press, Toronto, pp. 1133–1138.

Cohen, D. J. and Volkmar, F. R., eds. (1997). *Handbook of Autism and Pervasive Developmental Disorders*. John Wiley & Sons.

Cohen, H. F. (1984). *Quantifying Music: The Science of Music at the First Stage of Scientific Revolution, 1580-1650*. Reidel, Dordrecht.

Cohen, M. and Massaro, D. (1990). "Synthesis of Visible Speech". *Behaviour Research Methods, Instruments and Computers*, (22)2:260–263.

Colby, K. M. (1973). "Simulations of Belief Systems". In: *Computer Models of Thought and Language*. Schank, R. C. and Colby, K. M., eds. WH Freeman & Co., New York, pp. 251–286.

— (1975). *Artificial Paranoia*. Pergamon Press, New York.

— (1981). "Modeling a Paranoid Mind". *Behavioral and Brain Sciences*, (4)4:515–534.

— and Gilbert, J. P. (1964). "Programming a Computer Model of Neurosis". *Journal of Mathematical Psychology*, (1):405–417.

—, Weber, S., and Hilf, F. D. (1971). "Artificial Paranoia". *Artificial intelligence*, (2)1:1–25.

Colombetti, G. (2014). *The Feeling Body: Affective Science Meets the Enactive Mind*. MIT Press, Cambridge (MA).

— (2017). "The Embodied and Situated Nature of Moods". *Philosophia*, (45):1437–1451.

Comte, A. (1830). *Cours de philosophie positive*. Bureau du Cours de Philosophie Positive, Paris.

Congdon, C., Huber, M., Kortenkamp, D., et al. (1993). "Carmel Versus Flakey: A Comparison of Two Winners". *AI Magazine*, (14)1:49–49.

Conti, F. (2002). "Claude Bernard's *Des Fonctions du Cerveau*: An *ante litteram* Manifesto of the Neurosciences?" *Nature Reviews Neuroscience*, (3)12:979–985.

Cook, N. and Dibben, N. (2013). "Emotion in Culture and History. Perspectives from Musicology". In: *The cultural Study of Music: A Critical Introduction*. Middleton, R. and Finnegan, R., eds. Routledge, London, pp. 45–72.

Cordeschi, R. (2002). *The Discovery of the Artificial Behavior, Mind and Machines Before and Beyond Cybernetics*. Springer, Berlin.

Corti, M. and Cerrone, N. (2020). *Deep Convolutional Networks for Real-Time Analysis of Human Emotions to Extend Robotic Empathy*. Master Thesis. Politecnico di Milano, Milano. https://www.politesi.polimi.it/handle/10589/154244.

Cosentino, S. and Takanishi, A. (2021). "Human–Robot Musical Interaction". In: *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*. Miranda, E. R., ed. Springer International Publishing, Cham, pp. 799–822.

Coupland, D. (1995). *Microserfs*. Regan Books, New York.

Cover, T. and Hart, P. (1967). "Nearest Neighbor Pattern Classification". *IEEE Transactions on Information Theory*, (13)1:21–27.

Cowan, N. (2001). "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity". *Behavioral and Brain Sciences*, (24)1:87–114.

Coyne, A. K., Murtagh, A. and McGinn, C. (2020). "Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction". In: *HRI 2020: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, 23-26 March, 2020, Cambridge (UK)*. Young, J. O. and Belpaeme, T., eds. ACM/IEEE, New York, pp. 491–498.

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press, Cambridge.

— (1947). "Theory of the Human Operator in Control Systems 1: I. The Operator as an Engineering System". *British Journal of Psychology.General Section*, (38)2:56–61.

— (1948). "Theory of the Human Operator in Control Systems. II. Man as an Element in a Control System". *British Journal of Psychology*, (38)3:142.

— (1966). *The Nature of Psychology: A Selection of Papers, Essays and Other Writings*. Cambridge University Press, Cambridge.

Crane, T. (2009). "Intentionalism". In: *The Oxford Handbook of Philosophy of Mind*. Beckermann, A., McLaughlin, B. P. and Walter, S., eds. Oxford University Press, pp. 474–493.

— and Farkas, K. (2022). "The Limits of the Doxastic". In: *Oxford Studies in Philosophy of Mind*. Kriegel, U., ed. Vol. 2. Oxford University Press, Oxford, pp. 36–57.

Crowell, C. R., Deska, J. C., Villano, M., et al. (2019). "Anthropomorphism of Robots: Study of Appearance and Agency". *JMIR human factors*, (6)2:e12629.

Crowther-Heyck, H. (1999). "George A. Miller, Language, and the Computer Metaphor of Mind". *History of Psychology*, (21):37–64.

Cucciniello, I., Sangiovanni, S., Maggi, G., et al. (2023). "Mind Perception in HRI: Exploring Users' Attribution of Mental and Emotional States to Robots with Different Behavioural Styles". *International Journal of Social Robotics*, (15)5:867–877.

d'Allemagne, H. R. (1902). *Histoire des jouets*. Librairie Hachette, Paris.

D'Arms, J. and Jacobson, D. (2000). "The Moralistic Fallacy: On the 'Appropriateness' of Emotions". *Philosophical and Phenomenological Research*, (61)1:65–90.

— and Jacobson, D. (2003). "VIII. The Significance of Recalcitrant Emotion (or, Anti-Quasi–judgmentalism)". *Royal Institute of Philosophy Supplements*, (52):127–145.

Damasio, A. R. (1994). *Descartes' Error; Emotion, Reason, and the Human Brain*. Avon Books, New York.

— (1999). *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, San Diego.

— (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Harcourt Brace, San Diego.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray, London.

Daston, L. and Galison, P. (2007). *Objectivity*. Zone Books, New York.

Datteri, E. (2008). "Intelligenza artificiale: elaborazione simbolica e coordinazione senso-motoria". In: *La mente. Tradizioni filosofiche, prospettive scientifiche, paradigmi contemporanei*. Gensini, S. and Rainone, A., eds. Carocci, Roma.

— (2021). "Interactive Biorobotics". *Synthese*, (198)8:7577–7595.

— (2025). "Models for Surrogative Stimulation". *Erkenntnis*.

— and Schiaffonati, V. (2019). "Robotic Simulations, Simulations of Robots". *Minds and Machines*, (29)1:109–125.

— and Schiaffonati, V. (2023). "Computer Simulations and Surrogative Reasoning for the Design of New Robots". *Synthese*, (202)1:5.

Dautenhahn, K. (1998). "The Art of Designing Socially Intelligent Agents: Science, Fiction, and The Human in The Loop". *Applied Artificial Intelligence*, (12)7-8:573–617.

— (2007). "Socially Intelligent Robots: Dimensions of Human-Robot Interaction". *Philosophical Transactions of the Royal Society B: Biological Sciences*, (362)1480:679–704.

Davidson, R. J. (1998). "Affective Style and Affective Disorders: Perspectives from Affective Neuroscience". *Cognition & Emotion*, (12)3:307–330.

— and Sutton, S. K. (Apr. 1995). "Affective Neuroscience: The Emergence of a Discipline". *Current Opinion in Neurobiology*, (5)2:217–224.

Davies, S. (1994). *Musical Meaning and Expression*. Cornell University Press, Ithaca (NY).

— (2003). *Themes in the Philosophy of Music*. Oxford University Press, Oxford.

De, S., Bostan, I. and Sastry, N. (2025). "Making Social Platforms Accessible: Emotion-Aware Speech Generation with Integrated Text Analysis". In: *Social Networks Analysis and Mining*. Aiello, L. M., Chakraborty, T. and Gaito, S., eds. Springer Nature Switzerland, Cham, pp. 101–116.

De Marchi, S., Gena, C. and Lieto, A. (2024). "You Got the Feeling: Attributing Affective States to Dialogical Social Robots". In: *AI-HCI 2024: Proceedings of the 5th International Conference on Artificial Intelligence in Human-Computer Interaction, held as Part of the 26th HCI International Conference, HCII 2024, 29 June - 4 July, 2024, Washington, DC*. Degen, H. and Ntoa, S., eds. Springer Nature Switzerland, Cham, pp. 216–230.

De Somer, E. and Heymans, J. F. (1912). "Méthode pour conserver à l'état de survie la tête isolée des animaux mammifères". *Journal de Physiologie et de Pathologie Générale*, (14)6:1138–1142.

Dear, P. (1995). "Cultural History of Science: An Overview with Reflections". *Science, Technology, & Human Values*, (20)2:150–170.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)*, (39)1:1–22.

Dennett, D. C. (1971). "Intentional Systems". *Journal of Philosophy*, (68)4:87–106.

Dennett, D. C., ed. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Bradford Books, Cambridge (MA).

Deonna, J. A. and Teroni, F. (2008). *Qu'est-ce qu'une émotion?* Vrin, Paris.

— and Teroni, F. (2015). "Emotions as Attitudes". *Dialectica*, (69)3:293–311.

— and Teroni, F. (2024). "Emotions and Their Correctness Conditions: A Defense of Attitudinalism". *Erkenntnis*, (89)1:45–64.

Descartes, R. (1649). *Traité des passions de l'âme*. Henry Le Gras, Paris.

— (1650). *Compendium musicae*. Gisbert van Zyll & Theodor van Ackersdijk, Trajectum ad Rhenum.

Deweese-Boyd, I. (2023). "Self-Deception". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Fall 2023. Metaphysics Research Lab, Stanford University.

Dewey, J. (1934). *Art as Experience*. Minton, Balch & Co., New York.

Di Natale, A. F., Simonetti, M. E., La Rocca, S., et al. (2023). "Uncanny Valley Effect: A Qualitative Synthesis of Empirical Research to Assess the Suitability of Using Virtual Faces in Psychological Research". *Computers in Human Behavior Reports*, (10):100288.

Diamond, A. (1990). "Developmental Time Course in Human Infants and Infant Monkeys, and the Neural Bases of Inhibitory Control in Reaching". In: *The Development and Neural Bases of Higher Cognitive Functions*. Vol. 608. New York Academy of Sciences, pp. 637–676.

Diaz-Merced, W. L., Candey, R. M., Brickhouse, N., et al. (2011). "Sonification of Astronomical Data". *Proceedings of the International Astronomical Union*, (7)S285:133–136.

Dibattista, L. (2015). "La caratterizzazione fisiologica del contagio emotivo nel pensiero di Georges Dumas". In: *Emozioni infettive. Saggi storico-filosofici sul contagio emozionale*. Dibattista, L., ed. Mimesis Edizioni, Milano, pp. 51–71.

— (2021). "Introduzione a Dumas". In: *L'espressione delle emozioni*. In Dumas, G. (2021). L'espressione delle emozioni. It. tr. by Di Vittorio, P., Dibattista, L., editor. ETS, Bologna, pp. 5–25.

Dick, P. K. (1968). *Do Androids Dream of Electric Sheep?* Doubleday, New York (NY).

Diel, A. and Lewis, M. (2022). "The Deviation-from-Familiarity Effect: Expertise Increases Uncanniness of Deviating Exemplars". *PLOS ONE*, (17)9:e0273861. Goldwater, M. B., ed. (Visited on 08/19/2025).

— and Lewis, M. (2024). "Rethinking the Uncanny Valley as a Moderated Linear Function: Perceptual Specialization Increases the Uncanniness of Facial Distortions". *Computers in Human Behavior*, (157):108254.

—, Weigelt, S., and Macdorman, K. F. (2021). "A Meta-Analysis of the Uncanny Valley's Independent and Dependent Variables". *Journal of Human-Robot Interaction*, (11)1.

Dixon, T. (2023). *The History of Emotions: A Very Short Introduction*. Oxford University Press, Oxford.

Dokic, J. and Lemaire, S. (2015). "Are Emotions Evaluative Modes?" *Dialectica*, (69)3:271–292.

Doorn, E. A. van, Kleef, G. A. van and Pligt, J. van der. (2015). "Deriving Meaning from Others' Emotions: Attribution, Appraisal, and the Use of Emotions as Social Information". *Frontiers in Psychology*, (6).

Doran, J. E. (1968). "Experiments with Pleasure-Seeking Automaton". In: *Machine Intelligence 3*. Michie, D., ed. Edinburgh University Press, Edinburgh, pp. 195–216.

Döring, S. A. (2007). "Seeing What to Do: Affective Perception and Rational Motivation". *Dialectica*, (61)3:363–394.

Doyon, A. and Liaigre, L. (1966). *Jacques Vaucanson, mécanicien de génie*. Presses Universitaires de France, Paris.

Dreyfus, H. L. (1967). "Why Computers Must Have Bodies in Order to Be Intelligent". *The Review of Metaphysics*: 13–32.

— (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, Cambridge (MA).

Duchenne, G.-B.-A. (1862). *Mécanisme de la physionomie humaine, ou Analyse électro-physiologique de l'expression des passions*. Jules Renouard, Paris.

Duffy, B. R., Rooney, C., O'Hare, G. M., et al. (1999). "What is a Social Robot?" In: *AICS 1999: Proceedings of the 10th Irish Conference on Artificial Intelligence & Cognitive Science, 1-3 September, 1999 University College Cork*. Bridge, B., Byrne, R., O'Sullivan, B., et al., eds. Springer, Berlin, pp. 1–7.

Dumas, G. (1895). *Les états intellectuels dans la mélancolie*. Alcan, Paris.

— (1900). *La tristesse et la joie*. Alcan, Paris.

— (1906). *Le sourire et l'expression des émotions*. Alcan, Paris.

— (1922). "L'expression des émotions". *Revue Philosophique de la France et de l'Étranger*, (93):32–72.

— (1923). *Traité de psychologie. 2 Voll.* Alcan, Paris.

— (1926a). "Introduction à l'étude de l'expression des émotions". *Revue Philosophique de la France et de l'Étranger*, (101):223–259.

— (1926b). "Les méthodes dans l'étude de l'expression des émotions". *Revue Philosophique de la France et de l'Étranger*, (102):107–150.

— (1932). "La mimique des aveugles". *Bulletin de l'Académie de Médecine*, (107):607–610.

— (1948). *La vie affective. Physiologie, psychologie, socialisation*. Presses Universitaires de France, Paris.

— (1949). *Nouvelle traité de psychologie. 7 Voll.* Presses Universitaires de France, Paris.

— and Malloizel, L. (1910). "De l'expression polyglandulaire desémotions (Communication à la Société de Psychologie, 3 décembre 1909)". *Journal de Psychologie Normale et Pathologique*, (7):62–66.

Dunlap, K., ed. (1922). *The Emotions*. Williams & Wilkins Company, Baltimore.

Dupuy, J.-P. (2009). *On the Origins of Cognitive Science: The Mechanization of the Mind*. MIT Press, Cambridge (MA).

Durán, J. M. (2018). *Computer Simulations in Science and Engineering: Concepts - Practices - Perspectives*. Springer International Publishing, Cham.

Dyer, M. G. (1982). *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. MIT Press, Cambridge (MA).

— (1983). "The Role of Affect in Narratives". *Cognitive Science*, (7)3:211–242.

Eimler, S. C., Krämer, N. C. and Pütten, A. M. von der. (2011). "Empirical Results on Determinants of Acceptance and Emotion Attribution in Confrontation with a Robot Rabbit". *Applied Artificial Intelligence*, (25)6:503–529.

Ekman, P. (1980). "Biological and Cultural Contributions to Body and Facial Movement". In: *Anthropology of the Body*. Blacking, J., ed. Academic Press, London, pp. 34–84.

— (1985). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W. W. Norton & Company, New York.

— (1992). "An Argument for Basic Emotions". *Cognition & Emotion*, (6)3-4:169–200.
— (1994). "All Emotions Are Basic". In: *The Nature of Emotion*. Ekman, P. and Davidson, R., eds. Oxford University Press, Oxford, pp. 15–19.
— (1999). "Basic Emotions". In: *Handbook of Cognition and Emotion*. Dalgleish, T. and Power, M., eds. John Wiley & Sons, Sussex, pp. 45–60.
— (2003a). *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books/Henry Holt and Co., New York.
— (2003b). *Emotions Revealed: Understanding Faces and Feelings*. Weidenfeld & Nicolson, London.
— and Cordaro, D. (2011). "What Is Meant by Calling Emotions Basic". *Emotion Review*, (3):364–370.
— and Friesen, W. V. (1971). "Constants Across Cultures in the Face and Emotion". *Journal of Personality and Social Psychology*, (17)2:124.
— and Friesen, W. V. (1975). *Unmasking the Face*. Prentice Hall, Englewood Cliffs.
— and Friesen, W. V. (1978). *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.
—, Friesen, W. V., Ellsworth, P., et al. (1972). *Emotion in the Human Face. Guidelines for Research and an Integration of Findings*. Pergamon Press, Oxford.
—, Sorenson, E. R., and Friesen, W. V. (1969). "Pan-Cultural Elements in Facial Displays of Emotion". *Science*, (164)3875:86–88.
Elgin, C. Z. (2016). "Emotion and Understanding". In: *Epistemology and Emotions*. Doguoglu, U. and Brun, G., eds. Routledge, New York, pp. 33–49.
— (2017). *True Enough*. MIT Press, Cambridge (MA).
Elias, N. (1976). *Über den Prozeß der Zivilisation: Soziogenetische und psychogenetische Untersuchungen*. Suhrkamp, Frankfurt am Main.
Elliott, C. (1992). *The Affective Reasoner: A Process Model of Emotions in a Multiagent System*. PhD thesis, Northwestern University, Evanston (IL).
Embgen, S., Luber, M., Becker-Asano, C., et al. (2012). "Robot-Specific Social Cues in Emotional Body Language." In: *IEEE RO-MAN 2012: Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication, 9-13 September, 2012, Paris, France*. IEEE, ed. IEEE, pp. 1019–1025.
Ephron, N., director (1998). *You've Got Mail [Moving Picture]*. Warner Brothers. Lauren Shuler Donner Productions, Beverly Hills.
Eramo, G., Pastore, S., De Tullio, M., et al. (2025). "The Sound of Science: A Sonification Learning Experience in an Italian Secondary School". *Frontiers in Education*, (9).
Essa, I. A. and Pentland, A. P. (1997). "Coding, Analysis, Interpretation, and Recognition of Facial Expressions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (19)7:757–763.
European Union. (2021). "Regulation (EU) 2021/241 of the European Parliament and of the Council of 12 February 2021 establishing the Recovery and Resilience Facility". Official Journal of the European Union, L 57, 18 February 2021, pp. 17–75. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R0241.
Faraj, Z., Selamet, M., Morales, C., et al. (2021). "Facially Expressive Humanoid Robotic Face". *HardwareX*, (9):e00117.
Faught, W. S. (1978). *Motivation and Intentionality in a Computer Simulation Model of Paranoia*. Springer, Basel.
Febvre, L. (1941). "La sensibilité et l'histoire: Comment reconstituer la vie affective d'autrefois?" *Annales d'histoire sociale (1939-1941)*, (3)2:5–20.
Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig.
Feest, U. and Sturm, T. (2011). "What (Good) is Historical Epistemology?" *Erkenntnis*, (75)3:285–302.
Feld, S. (2015). "Acoustemology". In: *Keywords in Sound*. David, N. and Sakakeeny, M., eds. Duke University Press, Durham, pp. 12–21.
Fenigstein, A. (2001). "Paranoia". In: *The Disorders. Specialty Articles from the Encyclopedia of Mental Health*. Friedman, H. S., ed. Academic Press, Cambridge (MA), pp. 301–310.

Fernald, A., Taeschner, T., Dunn, J., et al. (1989). “A Cross-Language Study of Prosodic Modifications in Mothers’ and Fathers’ Speech to Preverbal Infants”. *Journal of Child Language*, (16)3:477–501.

Ferrell, C. (1996). “Orientation Behavior Using Registered Topographic Maps”. In: *From Animals to Animats 4. Proceedings of the Fourth International Conference on Simulation of Adpative Behavior, 9-13 September, 1996, Cape Cod (MA)*. MIT Press, Cambridge (MA), pp. 94–103.

— and Kemp, C. (1996). “An Ontogenetic Perspective to Scaling Sensorimotor Intelligence”. In: *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium, 9-11 November, 1996, Massachusetts Institute of Technology, Cambridge (MA)*. AAAI Press, Stanford, pp. 102–108.

Ferrigno, S., Huang, Y. and Cantlon, J. F. (2021). “Reasoning Through the Disjunctive Syllogism in Monkeys”. *Psychological Science*, (32)2:292–300.

Festinger, L. (1957). *Theory of Cognitive Dissonance*. Row Peterson.

Fiacconi, C. M., Owen, A. M. and Kohler, S. (2014). “Nature And Extent of Person Recognition Impairments Associated with Capgras Syndrome in Lewy Body Dementia”. *Frontiers in Human Neuroscience*, (8):726.

Filippini, C. and Merla, A. (2023). “Systematic Review of Affective Computing Techniques for Infant Robot Interaction”. *International Journal of Social Robotics*, (15)3:393–409.

—, Perpetuini, D., Cardone, D., et al. (2021). “Improving Human-Robot Interaction by Enhancing NAO Robot Awareness of Human Facial Expression”. *Sensors*, (21)19:6438.

Fink, J. (2012). “Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction”. In: *ICSR 2012: Proceedings of the 4th International Conference on Social Robotics, 29-31 October, 2012, China*. Sam Ge, S., Khatib, O., Cabibihan, J.-J., et al., eds. Springer, Berlin-Heidelberg, pp. 199–208.

Finnegan, R. (2003). “Music, Experience and the Anthropology of Emotion”. In: *The Cultural Study of Music: A Critical Introduction*. Clayton, M., Herbert, T. and Middleton, R., eds. Routledge, London-New York, pp. 353–363.

Fiorini, L., D’Onofrio, G., Sorrentino, A., et al. (2024). “The Role of Coherent Robot Behavior and Embodiment in Emotion Perception and Recognition During Human-Robot Interaction: Experimental Study”. *JMIR Human Factors*, (11)1:e45494.

Fisch, M. (2025). *Refexive Emotions: Shame, Humor, Humility*. Springer, New York.

Fischer, K., Jung, M., Jensen, L. C., et al. (2019). “Emotion Expression in HRI – When and Why”. In: *HRI 2019: Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction, 11-14 March, 2019, Daegu Republic of Korea*. Kim, J. and Tapus, A., eds. ACM/IEEE, New York, pp. 29–38.

Fiske, S. T. and Taylor, S. E. (1991). *Social Cognition*. Mcgraw-Hill Book Company Inc., Columbus (OH).

Fitzpatrick, S., Byrne, A. K., Headley, A., et al. (2025). “Neural Correlates of the Uncanny Valley Effect for Robots and Hyper-Realistic Masks”. *PLOS ONE*, (20)3:e0311714.

Fix, E. and Hodges, J. L. (1951). “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (Report)”. *USAF School of Aviation Medicine, Randolph Field, Texas*.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press, Cambridge (MA).

— (1980). “Methodological Solipsism as a Research Strategy in Cognitive Psychology”. *Behavioral and Brain Sciences*, (3):63–73.

— and Pylyshyn, Z. W. (1981). “How Direct Is Visual Perception?: Some Reflections on Gibson’s ”Ecological Approach”". *Cognition*, (9):139–196.

Fonagy, P. (1991). “Thinking About Thinking: Some Clinical and Theoretical Considerations in the Treatment of a Borderline Patient”. *International Journal of Psychoanalysis*, (72)4:639–656.

Forrester, J. W. (1968). *Principles of Systems*. Wright-Allen, Cambridge (MA).

Fowler, C. B. (1967). “The Museum of Music: A History of Mechanical Instruments”. *Music Educators Journal*, (54)2:45–49.

François-Franck, C. A. (1905). *Cours du Collège de France de 1880 à 1904 Et Travaux du Laboratoire de 1875 à 1904*. O. Doin, Paris.

Frank, M. G. and Stennett, J. (2001). "The Forced-Choice Paradigm and the Perception of Facial Expressions of Emotion". *Journal of Personality and Social Psychology*, (80)1:75–85.

Frederiksen, M. R., Fischer, K. and Matarić, M. (2022). "Robot Vulnerability and the Elicitation of User Empathy". In: *RO-MAN 2022: Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication, 29 August - 02 September, 2022, Naples (IT)*. IEEE, ed. IEEE, pp. 52–58.

Freedheim, D. K. and Weiner, I. B., eds. (2003). *Handbook of Psychology: Volume 1, History of Psychology*. John Wiley & Sons, New York.

Frege, G. (1884). *Die Grundlagen der Arithmetik: Eine Untersuchung über den Begriff der Zahl*. Koebner, Breslau.

Freud, S. (1900). *Die Traumdeutung*. Franz Deuticke, Leipzig.

— (1914). *Zur Einführung des Narzißmus*. Internationaler Psychoanalytischer Verlag, Leipzig-Wien-Zürich.

Freud, S. S. (1919). "Das Unheimliche". *Imago*, (5)6:297–394.

— (1925). "On Narcissism: An Introduction". In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Strachey, J., ed. Hogarth Press, London, pp. 67–102.

— (2003). *The Uncanny*. Eng. tr. by McLintok, D. Penguin Books, London.

Frevert, U. (2011). *Emotions in History – Lost and Found*. Central European University Press, Budapest.

— (2024). *Writing the History of Emotions: Concepts and Practices, Economies and Politics*. Bloomsbury Publishing, London.

Frigg, R. and Nguyen, J. (2017). "Scientific Representation Is Representation-As". In: *Philosophy of Science in Practice*. Chao, H.-K. and Reiss, J., eds. Springer International Publishing, Cham, pp. 149–179.

— and Nguyen, J. (2021). "Scientific Representation". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Winter 2021. Metaphysics Research Lab, Stanford University.

Frijda, N. H. (1986). *The Emotions*. Cambridge University Press, Cambridge.

— (1988). "The Laws of Emotion". *American Psychologist*, (43)5:349–358.

— and Mesquita, B. (1993). "The Social Roles and Functions of Emotions". In: *Emotion and Culture: Empirical Studies of Mutual Influence*. Kitayama, S. and Markus, H. R., eds. American Psychological Association, Washington, DC, pp. 51–87.

— and Swagerman, J. (1987). "Can Computers Feel? Theory and Design of an Emotional System". *Cognition and Emotion*, (1)3:235–257.

Fronzuto, G. (2008). *Organi di Roma. Gli organi delle quattro basiliche maggiori*. Leo S. Olschki, Firenze.

Fujiso, Y., Perrin, N., Van Der Giessen, J., et al. (2018). "Swall-E: a robotic in-vitro simulation of human swallowing". *PLoS One*, (13)12:e0208193.

Fukushima, K. (1969). "Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements". *IEEE Transactions on Systems Science and Cybernetics*, (5)4:322–333.

— (1975). "Cognitron: A Self-Organizing Multilayered Neural Network". *Biological cybernetics*, (20)3:121–136.

— (1980). "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". *Biological Cybernetics*, (36)4:193–202.

Fum, D. and Stocco, A. (2004). "Memory, Emotion, and Rationality: An ACT-R Interpretation for Gambling Task Results". In: *ICCM 2004: Proceedings of the Sixth International Conference on Cognitive Modeling, 30 July-1 August 2004, Pittsburgh (PA), USA*. Lovett, M. C., Schunn, C. D., Lebiere, C., et al., eds. Psychology Press, London, pp. 106–111.

Fux, J. J. (1725). *Gradus ad parnassum sive manuductio ad compositionem musicae regularem*. Johann Peter van Ghelen, Wien.

Gabor, D. (1956). "Models in Cybernetics". In: *I Modelli nella Tecnica. Atti del convegno di Venezia, 1-4 ottobre 1955*. Vol. 2. Accademia Nazionale dei Lincei, Roma, pp. 236–274.

Gadanho, S. and Custódio, L. (2003). "Asynchronous Learning by Emotions and Cognition". In: *SAB 2002: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior, 12-16 July 2002, Cambridge (MA), USA*. Hallam, B., Floreano, D., Hallam, J., et al., eds. MIT Press, Cambridge (MA), pp. 224–225.

Galen. (1963). *On the Passions and Errors of the Soul*. Eng. tr. by Harkins, Paul W. Ohio State University Press, Columbus.

Gallese, V. (2005). "Embodied Simulation: From Neurons to Phenomenal Experience". *Phenomenology and the Cognitive Sciences*, (4)1:23–48.

— (2007). "Embodied Simulation: From Mirror Neuron Systems to Interpersonal Relations". *Novartis Foundation Symposium*, (278):3–12.

— and Goldman, A. (1998). "Mirror Neurons and the Simulation Theory of Mind-Reading". *Trends in Cognitive Sciences*, (2)12:493–501.

Garcia, S., Gomez-Donoso, F. and Cazorla, M. (2024). "Enhancing Human–Robot Interaction: Development of Multimodal Robotic Assistant for User Emotion Recognition". *Applied Sciences*, (14)24.

Gasparetto, A. and Scalera, L. (2019). "A Brief History of Industrial Robotics in the 20th Century". *Advances in Historical Studies*, (8):24–35.

Gasteiger, N., Lim, J., Hellou, M., et al. (2024). "A Scoping Review of the Literature On Prosodic Elements Related to Emotional Speech in Human-Robot Interaction". *International Journal of Social Robotics*, (16)4:659–670.

Gay, P. (1984–1998). *The Bourgeois Experience: Victoria to Freud*. 5 Vols. Oxford University Press, New York.

Gazzaniga, M. S. and LeDoux, J. E. (1978). *The Integrated Mind*. Plenium Press, New York-London.

Gebhard, P. (2005). "ALMA: A Layered Model of Affect". In: *AAMAS 2005: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, 25-29 July 2005, Utrecht, Netherlands*. Wooldridge, M., McBurney, P. and Parsons, S., eds. ACM Press, New York, pp. 29–36.

Gelder, B. de, Vroomen, J., Pourtois, G., et al. (1999). "Non-Conscious Recognition of Affect in the Absence of Striate Cortex." *Neuroreport*, (10)18.

Gendler, T. S. (2008). "Alief and Belief". *The Journal of Philosophy*, (105)10:634–663.

Geng, H., Xu, P., Aleman, A., et al. (2024). "Dynamic Organization of Large-Scale Functional Brain Networks Supports Interactions Between Emotion and Executive Control". *Neuroscience Bulletin*, (40)7:981–991.

George, M. S., Ketter, T. A., Gill, D. S., et al. (1993). "Brain Regions Involved in Recognizing Facial Emotion or Identity: An Oxygen-15 Pet Study." *The Journal of Neuropsychiatry and Clinical Neurosciences*, (5)4:384–394.

—, Ketter, T. A., and Post, R. M. (1994). "Prefrontal Cortex Dysfunction in Clinical Depression". *Depression*, (2)2:59–72.

Gervasi, R., Barravecchia, F., Mastrogiacomo, L., et al. (2023). "Applications of Affective Computing in Human-Robot Interaction: State-of-Art and Challenges for Manufacturing". *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, (237)6-7:815–832.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Waveland Press, Inc., Prospect Heights.

— (1979). *The Ecological Approach to Visual Perception*. Houghton-Mifflin, Boston.

Giere, R. N. (1973). "History and Philosophy of Science: Intimate Relationship or Marriage of Convenience?" *The British Journal for the Philosophy of Science*, (24)3:282–297.

Gigerenzer, G. and Goldstein, D. G. (1996). "Mind as Computer: Birth of a Metaphor". *Creativity Research Journal*, (9)2-3:131–144.

Giralt, G., Sobek, R. and Chatila, R. (1979). "A Multi-Level Planning and Navigation System for a Mobile Robot: A First Approach to Hilare". In: *IJCAI-79: Proceedings of the 6th International Joint Conference on Artificial Intelligence, August 20-23, 1979, Tokio*. Buchanan, B. G., ed. Morgan Kaufmann Publishers, San Francisco, pp. 335–337.

Glenberg, A. M. (1997). "What Memory is for". *Behavioral and brain sciences*, (20)1:1–19.

— and Kaschak, M. P. (2002). "Grounding Language in Action". *Psychonomic Bulletin & Review*, (9):558–565.

Glickstein, M. (2017). *Neuroscience: A Historical Introduction*. MIT Press.

Glüxam, D. (2020). *"Aus der Seele muß man spielen...": Über die Affekttheorie in der Musik des 17. und 18. Jahrhunderts und ihre Auswirkung auf die Interpretation*. Hollitzer Wissenschaftsverlag, Wien.

Gmytrasiewicz, P. J. and Lisetti, C. L. (2000). "Using Decision Theory to Formalize Emotions in Multi-Agent Systems". In: *ICMAS 2000: Proceedings of the Fourth International Conference on MultiAgent Systems, 10-12 July 2000, Boston (MA), USA*. Durfee, E. H., ed. IEEE Computer Society, Los Alamitos (CA), pp. 31–38.

Gödel, K. (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". *Monatshefte für Mathematik und Physik*, (38):173–198.

— (1995). "Some Basic Theorems on the Foundations of Mathematics and Their implications". *Collected Works, Volume III: Unpublished Essays and Lectures*: 304–323. Feferman, S., Dawson Jr, J. W., Goldfarb, W., et al., eds. Originally presented as the Josiah Willard Gibbs Lecture, December 26, 1951.

Goel, S., Jara-Ettinger, J., Ong, D. C., et al. (2024). "Face and Context Integration in Emotion Inference Is Limited and Variable Across Categories and Individuals". *Nature Communications*, (15)1:2443.

Goldie, P. (2000). *The Emotions: A Philosophical Exploration*. Oxford University Press, Oxford.

Goldman, A. I. (2012). "Theory of Mind". In: *The Oxford Handbook of Philosophy of Cognitive Science*. Margolis, E., Samuels, R. and Stich, S. P., eds. Oxford University Press, Oxford, pp. 402–424.

Goldman, E. J., Baumann, A. E., Pare, L., et al. (2025). "Children's Attribution of Mental States to Humans and Social Robots Assessed with the Theory of Mind Scale". *Scientific Reports*, (15)1.

Goldstein, I. (2002). "Are Emotions Feelings?: A Further Look at Hedonic Theories of Emotions". *Consciousness & Emotion*, (3)1:21–33.

Gonzalez-Aguirre, J. A., Osorio-Oliveros, R., Rodríguez-Hernández, K. L., et al. (2021). "Service Robots: Trends and Technology". *Applied Sciences*, (11)22.

Goodman, N. (1968). *Languages of Art: An Approach to a Theory of Symbols*. Bobbs-Merrill, Indianapolis-New York.

Gopnik, A. and Wellman, H. M. (1992). "Why the Child's Theory of Mind Really Is a Theory". *Mind & Language*, (5):145–171.

Gordon, R. M. (1973). "Judgmental Emotions". *Analysis*, (34)2:40–48.

— (1986). "Folk Psychology as Simulation". *Mind & Language*, (1)2:158–171. (Visited on 08/01/2025).

Goswami, A. and Vadakkepat, P., eds. (2018). *Humanoid Robotics: A Reference. Part I: History of Humanoid Robots*, pages 2-52. Springer, New York.

Gottman, J. M., Katz, L. F. and Hooven, C. (2013). *Meta-Emotion: How Families Communicate Emotionally*. Routledge, New York.

Grágeda, N., Busso, C., Alvarado, E., et al. (2025). "Speech Emotion Recognition in Real Static and Dynamic Human-Robot Interaction Scenarios". *Computer Speech & Language*, (89):101666.

Gratch, J. (2000). "Emile: Marshalling Passions in Training and Education". In: *AGENTS 2000: Proceedings of the Fourth International Conference on Autonomous Agents, 3–7 June 2000, Barcelona, Spain*. Sierra, C., Gini, M. L. and Rosenschein, J. S., eds. ACM Press, New York, pp. 325–332.

— and Marsella, S. (2004). "A Domain-Independent Framework for Modeling Emotion". *Cognitive Systems Research*, (5)4:269–306.

Gray, H. M., Gray, K. and Wegner, D. M. (2007). "Dimensions of Mind Perception". *Science*, (315)5812:619.

Greco, M. and Stenner, P. (2008). *Emotions: A Social Science Reader*. Routledge, London.

Green, B. F. (1954). "Attitude Measurement". In: *Handbook of Social Psychology*. Lindzey, G., ed. Addison-Wesley, Cambridge, pp. 335–369.

Greene, P. H. (1982). "Why Is It Easy to Control Your Arms?" *Journal of Motor Behavior*, (14)4:260–286.

Gregory, R. L. (1972). "Seeing as Thinking: An Active Theory of Perception". *Times Literary Supplement*, (23):707–708.

Griffiths, P. E. (2003). "III. Basic Emotions, Complex Emotions, Machiavellian Emotions". *Royal Institute of Philosophy Supplements*, (52):39–67.

Gross, C. G. (1998). "Claude Bernard and the Constancy of the Internal Environment". *The Neuroscientist*, (4)5:380–385.

Grossberg, S. (1976). "Adaptive Pattern Classification and Universal Recoding: I. Parallel Development and Coding of Neural Feature Detectors". *Biological cybernetics*, (23)3:121–134.

Grossmann, I., Huynh, A. C. and Ellsworth, P. C. (2016). "Emotional Complexity: Clarifying Definitions and Cultural Correlates." *Journal of Personality and Social Psychology*, (111)6:895–916.

Grout, D. J. and Palisca, C. V. (2001). *A History of Western Music*. W. W. Norton & Company, New York.

Grundke, A., Stein, J.-P. and Appel, M. (2022). "Mind-Reading Machines: Distinct User Responses to Thought-Detecting and Emotion-Detecting Robots". *Technology, Mind, and Behavior*, (3)1.

Grzankowski, A. (2013). "Non-Propositional Attitudes". *Philosophy Compass*, (8)12:1123–1137.

— (2015). "Not all Attitudes Are Propositional". *European Journal of Philosophy*, (23)3:374–391.

— (2025). "Do Emotions Represent Values and How Can We Tell?" *Mind & Language*.

— and Montague, M. (2018). *Non-Propositional Intentionality*. Oxford University Press, Oxford.

Guizzo, E. (2020). "What Is a Robot?" *IEEE Robotics Guide*. IEEE, New York. Available at: https://robots.ieee.org/learn/what-is-a-robot/ (Accessed: July 15, 2025).

Gullahorn, J. T. and Gullahorn, J. E. (1964). "Computer Simulation of Human Interaction in Small Groups". In: *AFIPS 1964: Proceedings of the Spring Joint Computer Conference, 21-23 April, 1964, Washington, DC, USA*. Roller, H. R., ed. ACM Press, New York, pp. 103–113.

Gullahorn, J. T. and Gullahorn, J. E. (1963). "A Computer Model of Elementary Social Behavior". In: *Computers and Thought*. Feldman, J. and Feigenbaum, E. A., eds. Mcgraw-Hill Book Company Inc., New York, pp. 375–386.

Gunderson, K. (1963). "Interview with a Robot". *Analysis*, (23)6:136–142.

— (1968). "Robots, Consciousness, and Programmed Behaviour". *The British Journal for the Philosophy of Science*, (19)2:109–122.

— (1971). *Mentality and Machines*. Doubleday, New York.

Gunther, Y. H., ed. (2003). *Essays on Nonconceptual Content*. MIT Press, Cambridge (MA).

— (2004). "The Phenomenology and Intentionality of Emotion". *Philosophical Studies*, (117)1-2:43–55.

Hadj-Bouziane, F., Bell, A. H., Knusten, T. A., et al. (2013). "Amygdala Lesions Disrupt Modulation of Functional Responses in Visual Cortex by Emotional Valence". *Proceedings of the National Academy of Sciences*, (109)52:E2865–E2874.

Hadwin, J. and Perner, J. (1991). "Pleased and Surprised: Children's Cognitive Theory of Emotion". *British Journal of Developmental Psychology*, (9)2:215–234.

Halberstadt, J., Winkielman, P., Niedenthal, P. M., et al. (2009). "Emotional Conception: How Embodied Emotion Concepts Guide Perception and Facial Action". *Psychological Science*, (20)10:1254–1261.

Hanslick, E. (1854). *Vom musikalisch Schönen*. Eng. tr. by Cohen, G. (1891), *The Beautiful in Music*. Novello and Company, London. Weigel, Leipzig.

Harbou, T. von. (1925). *Metropolis*. Eng. tr. by McMillan, D. D. R. (2023). Metropolis. Wildside Press, Cabin John. August Scherl, Berlin.

Hareli, S. and Parkinson, B. (2008). "What's social about social emotions?" *Journal for the Theory of Social Behaviour*, (38)2:131–156.

Harris, P. L., Johnson, C. N., Hutton, D., et al. (1989). "Young Children's Theory of Mind and Emotion". *Cognition and Emotion*, (3)4:379–400.

Hashimoto, K. and Takanishi, A. (2018). "WABIAN and Other Waseda Robots". In: *Humanoid Robotics: A Reference*. Goswami, A. and Vadakkepat, P., eds. Springer Netherlands, Dordrecht, pp. 265–289.

Hashimoto, S., Narita, S., Kasahara, H., et al. (1997). “Humanoid Robot-Development of an Information Assistant Robot Hadaly”. In: *RO-MAN 1997: Proceedings 6th IEEE International Workshop on Robot and Human Communication, 19 September - 1 October, 1997, Tohoku University, Sendai, Japan*. IEEE, ed. IEEE, New York, pp. 106–111.

—, Narita, S., Kasahara, H., et al. (2002). “Humanoid Robots in Waseda University—Hadaly-2 and WABIAN”. *Autonomous Robots*, (12):25–38.

Haslam, N. (2006). “Dehumanization: An Integrative Review”. *Personality and Social Psychology Review*, (10)3:252–264.

al–Ḥassan, A. Y. (1986). *Islamic Technology. An Illustrated History*. Cambridge University Press and Unesco, Paris.

Hatfield, G. (1988). “Representation and Content in Some (Actual) Theories of Perception”. *Studies in History and Philosophy of Science Part A*, (19)2:175–214.

Haugeland, J. (1978). “The Nature and Plausibility of Cognitivism”. *Behavioral and Brain Sciences*, (1)2:215–226.

— (1985). *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge (MA).

Hay, E. L. and Diehl, M. (2011). “Emotion Complexity and Emotion Regulation Across Adulthood”. *European Journal of Ageing*, (8)3:157–168.

Healey, J. and Picard, R. (1998). “Digital Processing of Affective Signals”. In: *ICASSP 1998: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 6. IEEE, pp. 3749–3752.

Hebb, D. O. (1949). *The Organization of Behavior. A Neuropsychological Theory*. John Wiley & Sons, New York.

Hegel, F., Muhl, C., Wrede, B., et al. (2009). “Understanding Social Robots”. In: *ACHI 2009: Proceedings of the 2nd International Conferences on Advances in Computer-Human Interactions, 1-7 February, 2009, Cancun, Mexico*. IEEE, ed. IEEE, New York, pp. 169–174.

Heidegger, M. (1927). *Sein und Zeit*. Max Niemayer Verlag, Tübingen.

— (1955). “Zur Seinsfrage”. In: *Gesamtausgabe, Band IX (1976)*. Vittorio Klostermann, Frankfurt am Main, pp. 385–426.

Heider, F. (1946). “Attitudes and Cognitive Organization”. *The Journal of Psychology*, (21)1:107–112.

— and Simmel, M. (1944). “An Experimental Study of Apparent Behavior”. *The American Journal of Psychology*, (57)2:243–259.

Heinisch, J. S., Kirchhoff, J., Busch, P., et al. (2024). “Physiological Data for Affective Computing in HRI with Anthropomorphic Service Robots: The AFFECT-HRI Data Set”. *Scientific Data*, (11)1:333.

Heise, D. R. (1977). “Social Action as the Control of Affect”. *Behavioral Science*, (22)3:163–177.

— (1979). *Understanding Events. Affect and the Construction of Social Action*. Cambridge University Press, Cambridge.

— (1987). “Affect Control Theory: Concepts and Model”. *The Journal of Mathematical Sociology*, (13)1-2:1–33.

Helmholtz, H. (1862). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. F. Vieweg & So., Braunschweig.

— (1867). *Handbuch der physiologischen Optik*. Leopold Voss, Leipzig.

Herbart, J. F. (1825). *Psychologie als Wissenschaft: Neu gegründet auf Erfahrung, Metaphysik und Mathematik*. August Wilhelm Unzer, Königsberg.

Herring, E., Matthew Jones, K., Kiprijanov, K. S., et al., eds. (2019). *The Past, Present, and Future of Integrated History and Philosophy of Science*. 1st ed. Routledge.

Herzog, O., Forchhammer, N., Kong, P., et al. (2022). “The Influence of Robot Designs on Human Compliance and Emotion: A Virtual Reality Study in the Context of Future Public Transport”. *ACM Transactions on Human-Robot Interaction*, (11)2:1–17.

Hieida, C. and Nagai, T. (2022). “Survey and Perspective on Social Emotions in Robotics”. *Advanced Robotics*, (36)1-2:17–32.

Hilbert, D. (1900). “Mathematische Probleme”. *Nachrichten von der Koniglichen Gesellschaft der Wissenschaften zu Gottingen*.

— (Dec. 1930). “Probleme der Grundlegung der Mathematik”. *Mathematische Annalen*, (102)1:1–9.

— and Ackermann, W. (1928). *Grundzüge der theoretischen Logik*. Springer, Berlin.

Hill, D. R., ed. (1998). *Studies in Medieval Islamic Technology: From Philo to Al-Jazarī, from Alexandria to Diyār Bakr*. Ashgate Publishing, Burlington.

Hill, E., Cherston, J., Goldfarb, S., et al. (2017). “ATLAS Data Sonification: A New Interface for Musical Expression”. In: *ICHEP 2016: Proceedings of 38th International Conference on High Energy Physics (PoS), 3-10 August 2016, Chicago*.

Hirai, K. (1998). “Current and Future Perspective of Honda Humanoid Robot”. In: *Robotics Research: Proceedings of the Eighth International Symposium, 4-7, October, 1997, Shonan, Japan*. Shirai, Y. and Hirose, S., eds. Springer, Berlin, pp. 446–450.

—, Hirose, M., Haikawa, Y., et al. (1998). “The Development of Honda Humanoid Robot”. In: *ICRA 1998: Proceedings of the 1998 IEEE International Conference on Robotics and Automation, 20-20 May, 1998, Katholieke Universiteit Leuven, Leuven, Belgium*. IEEE, ed. Vol. 2. IEEE, New York, pp. 1321–1326.

Hirose, M. and Ogawa, K. (2007). “Honda Humanoid Robots Development”. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, (365)1850:11–19.

—, Takenaka, T., Gomi, H., et al. (1997). “Humanoid Robot”. *Journal of the robotics society of Japan*, (15)7:983–985.

Hobbes, T. (1651). *Leviathan or The Matter, Forme and Power of a Common Wealth Ecclesiastical and Civil*. Andrew Crooke, London.

Hobson, R. P. (1993). *Autism and the Development of Mind*. Erlbaum, Mahwah (NJ).

Hochschild, A. R. (1979). “Emotion Work, Feeling Rules, and Social Structure”. *American Journal of Sociology*, (85)3:551–575.

Hoffman, G. and Weinberg, G. (2010). “Shimon: An Interactive Improvisational Robotic Marimba Player”. In: *CHI EA 2010: Extended Abstracts of the Conference on Human Factors in Computing Systems, 10-15 April, 2010, Atlanta*. Mynatt, E. and Schoner, D., eds. Association for Computing Machinery, New York, pp. 3097–3102.

— and Weinberg, G. (2011). “Interactive Improvisation with a Robotic Marimba Player”. *Autonomous Robots*, (31)2-3:133–153.

Hoffmann, E. T. (1817). “Der Sandmann”. In: *Nachtstücke*. Callots Manier Verlag, Berlin.

Holland, O. (2003). “The First Biologically Inspired Robots”. *Robotica*, (21)4:351–363.

Homans, G. C. (1958). “Social Behavior as Exchange”. *American Journal of Sociology*, (63)6:597–606.

— (1961). *Social Behavior: Its Elementary Forms*. Harcourt, Brace and World, New York:

Hovy, E. H. (1986). “Putting Affect into Text”. In: *Proceedings of the 8th Annual Conference of the Cognitive Science Society, 15-17 August, 1986, Amhurst*. Erlbaum, Englewood Cliffs (NJ), pp. 669–675.

Hudlicka, E. (2005). “A Computational Model of Emotion and Personality: Applications to Psychotherapy Research and Practice”. In: *CyberTherapy 2005: Proceedings of the Tenth Annual CyberTherapy Conference, 6-10 June 2005, Basel, Switzerland*. Wiederhold, B. K. and Bullinger, A. H., eds. Special Issue of CyberPsychology & Behavior, 8(4), pp. 1–7.

— and Fellous, J. M. (1996). *Review of Computational Models of Emotion*. Technical Report 9612. Arlington (MA).

Hufendiek, R. (2015). *Embodied Emotions: A Naturalist Approach to a Normative Phenomenon*. Routledge, New York.

— (2018). “Explaining Embodied Emotions – with and Without Representations”. *Philosophical Explorations*, (21)2:319–331.

Hume, D. (1739). *A Treatise of Human Nature*. John Noon, London.

Hung, L., Zhao, Y., Alfares, H., et al. (2025). “Ethical Considerations in the Use of Social Robots for Supporting Mental Health and Wellbeing in Older Adults in Long-Term Care”. *Frontiers in Robotics and AI*, (12):1560214.

Husserl, E. (1922). *Ideen zu einer reinen Phänomenologie und Phänomenologischen Philosophie*. Max Niemeyer, Halle.

— (1929). “Cartesianische Meditationen: Eine Einleitung in die Phänomenologie”. In: *Husserliana: Edmund Husserl Gesammelte Werke, Band I (1950)*. Strasser, S., ed. Martinus Nijhoff, Den Haag.

— (1936). “Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie”. In: *Husserliana: Edmund Husserl Gesammelte Werke, Band VI (1954)*. Biemel, W., ed. Martinus Nijhoff, Den Haag.

Hutto, D. D. (2012). “Truly Enactive Emotion”. *Emotion Review*, (4)2:176–181.

— and Ravenscroft, I. (2021). “Folk Psychology as a Theory”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Fall 2021. Metaphysics Research Lab, Stanford University.

Huxley, A. (1932). *Brave New World*. Harper & Row, New York and Evanston.

International Federation of Robotics. (2024). *World Robotics Report 2024*. https://ifr.org/img/worldrobotics/Press_Conference_2024.pdf. International Federation of Robotics, Frankfurt.

Irie, R. E. (1997). “Multimodal Sensory Integration for Localization in a Humanoid Robot”. In: *IJCAI 1997: Proceedings of Second IJCAI Workshop on Computational Auditory Scene Analysis (CASA 1997), 23-29 August, Nagoya*. Pollack, M. E., ed. Morgan Kaufmann Publishers, San Francisco, pp. 55–59.

Ivakhnenko, A. G. (1971). “Polynomial Theory of Complex Systems”. *IEEE Transactions on Systems, Man, and Cybernetics*, (SMC-1)4:364–378.

— and Lapa, V. G. (1967). *Cybernetics and Forecasting Techniques*. American Elsevier Publishing Company, New York.

Izard, C. E. (1968). “Cross-Cultural Research Findings on Development in Recognition of Facial Behavior”. *Proceedings of the 76th Annual Convention of the American Psychological Association, San Francisco*, (3)727.

— (1969). “The Emotions and Emotion Constructs in Personality and Culture Research”. In: *Handbook of Modern Personality Theory*. Cattell, R. B. and Dreger, R. M., eds. John Wiley & Sons, New York, pp. 496–510.

— (1991). *The Psychology of Emotions*. Springer Science & Business Media, New York.

— (1993). “Four Systems for Emotion Activation: Cognitive and Noncognitive Processes”. *Psychological Review*, (100)1:68.

Jack, R. E., Garrod, O. G. B., Yu, H., et al. (2012). “Facial Expressions of Emotion Are Not Culturally Universal”. *Proceedings of the National Academy of Sciences*, (109)19:7241–7244.

Jackson, M., ed. (2011). *The Oxford Handbook of the History of Medicine*. Oxford University Press, Oxford.

Jacob, P. (2023). “Intentionality”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Spring 2023. Metaphysics Research Lab, Stanford University.

Jakob, C. (1906). “Localización del alma y de la inteligencia”. *El Libro*, (1):151.

James, W. (1884). “What is an Emotion?” *Mind*, (9)34:188–202.

— (1890). *The Principles of Psychology*. Henry Holt and Co., New York.

— (1894). “The Physical Basis of Emotion”. *Psychological Review*, (101)2:205–210.

— and Lange, C. G. (1922). *The Emotions*. Williams & Wilkins Company, Baltimore.

Janet, P. M. F. (1901). “Une Extatique. Conférence faite à l’Institut Psychologique International le 25 mai 1901”. *Bulletin de l’Institut Psychologique International*, (1):209–240.

Jastrzab, L. E., Chaudhury, B., Ashley, S. A., et al. (2024). “Beyond Human-Likeness: Socialness Is More Influential When Attributing Mental States to Robots”. *iScience*, (27)6:110070.

al–Jazarī. (1989). الهندسية الحيل معرفة في كتاب [*Kitab fi Ma’rifat al-Hiyal al-Handasiyya*]. Eng. tr. by Hill D.R. *The Book Knowledge of Ingenious Mechanical Devices*. Elite Publishers, Islamabad.

Jefferson, G. (1949). “The Mind of Mechanical Man”. *British Medical Journal*, (1)4616:1105.

Jentsch, E. (1906). “Zur Psychologie des Unheimlichen”. *Psychiatrisch-neurologische Wochenschrift*, (22)23:195–205.

Jentsch, E. A. (1997). “On the Psychology of the Uncanny”. Trans. by R. Sellars. *Angelaki: Journal of the Theoretical Humanities*, (2)1:7–16. Eng. tr. by Sellars, Roy.

Jeon, H. and Lee, S.-H. (2018). “From Neurons to Social Beings: Short Review of the Mirror Neuron System Research and Its Socio-Psychological and Psychiatric Implications”. *Clinical Psychopharmacology and Neuroscience*, (16)1:18–31.

Jing, H., Lun, X., Dan, L., et al. (2015). "Cognitive Emotion Model for Eldercare Robot in Smart Home". *China Communications*, (12)4:32–41.

Johnson, M. H. (1987). *The Body in the Mind: The Bodily Basis of Imagination, Reason and Meaning, Chicago: Chic*. University of Chicago Press, Chicago.

— (1993). "Constraints on Cortical Plasticity". In: *Brain Development and Cognition: A Reader*. Johnson, M. H., ed. Blackwell, Oxford, pp. 703–721.

Johnson-Laird, P. N. and Oatley, K. (1989). "The Language of Emotions: An Analysis of a Semantic Field". *Cognition and Emotion*, (3)2:81–123.

— and Oatley, K. (1998). "Basic Emotions, Rationality, and Folk Theory". In: *Consciousness and Emotion in Cognitive Science*. Clark, A. and Toribio, J., eds. Routledge, pp. 289–311.

Johnston, M. (2001). "The Authority of Affect". *Philosophy and Phenomenological Research*, (63)1:181.

Jung, M. F. (2017). "Affective Grounding in Human-Robot Interaction". In: *HRI 2017: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, 6-9 March, 2017, Vienna*. Mutlu, B. and Tscheligi, M., eds. ACM/IEEE, New York, pp. 263–273.

Juslin, P. N., Sloboda, J., Juslin, P. N., et al., eds. (2011). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, Oxford, New York.

Kabacińska, K., Prescott, T. J. and Robillard, J. M. (2021). "Socially Assistive Robots as Mental Health Interventions for Children: A Scoping Review". *International Journal of Social Robotics*, (13)5:919–935.

Kalan, S., Chauhan, S., Coelho, R. F., et al. (2010). "History of Robotic Surgery". *Journal of Robotic Surgery*, (4):141–147.

Kamide, H., Kawabe, K., Shigemi, S., et al. (2014). "Relationship Between Familiarity and Humanness of Robots – Quantification of Psychological Impressions Toward Humanoid Robots". *Advanced Robotics*, (28)12:821–832.

Kant, I. (1788). *Kritik der praktischen Vernunft*. Johann Friedrich Hartknoch, Riga.

— (1797). *Die Metaphysik der Sitten*. Friedrich Nicolovius, Königsberg.

— (1798). *Anthropologie in pragmatischer Hinsicht*. Friedrich Nicolovius, Königsberg.

Kato, I. (1974). "Information-Power Machine with Senses and Limbs (WABOT-1)". In: *CISM-IFToMM 1974: Proceedings of the First CISM-IFToMM Symposium on Theory and Practice of Robots and Manipulators, 5-8 September, 1973, Udine, Italy*. Morecki, A., Bianchi, G. and K dzior, K., eds. Vol. 1. Springer, Berlin, pp. 11–24.

—, Ohteru, S., Shirai, K., et al. (June 1987). "The Robot Musician "Wabot-2" (Waseda Robot-2)". *Robotics*, (3)2:143–155.

Kätsyri, J., De Gelder, B. and Takala, T. (2019). "Virtual Faces Evoke Only a Weak Uncanny Valley Effect: An Empirical Investigation with Controlled Virtual Face Images". *Perception*, (48)10:968–991.

—, Förger, K., Mäkäräinen, M., et al. (2015). "A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley of Eeriness". *Frontiers in Psychology*, (6).

Katz, J. E., Rice, R. E. and Aspden, P. (2001). "The Internet, 1995-2000: Access, civic involvement, and social interaction". *American behavioral scientist*, (45)3:405–419.

Kauppinen, A. (2022). "Moral Sentimentalism". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Spring 2022. Metaphysics Research Lab, Stanford University.

— (2024). "How Emotions Grasp Value". *Philosophical Issues*, (34)1:213–233.

Kelso, J. A. S. (1999). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press, Cambridge (MA).

Kemper, T. D. (1978). *A Social Interactional Theory of Emotions*. John Wiley & Sons-Interscience Publication, New York.

Kenny, A. (1963). *Action, Emotion and Will*. Routledge & Kegan Paul, New York.

Kerner, G. (1982). "Emotions Are Judgments of Value". *Topoi*, (1):52–56.

Kety, S. S. (1970). "The Biogenic Amines in the Central Nervous System: Their Possible Roles in Arousal, Emotion and Learning". In: *The Neurosciences: Second Study Program*. Schmitt, F. O., ed. The Rockefeller University Press, New York, pp. 324–336.

— (1972). “Brain Catecholamines, Affective States and Memory”. In: *The Chemistry of Mood, Motivation, and Memory. The Proceedings of an Interdisciplinary Conference on the Chemistry of Mood, Motivation, and Memory held at the University of California, San Francisco, in October 1971*. Schmitt, F. O., ed. Plenium Press, New York-London, pp. 65–80.

Kiilavuori, H., Peltola, M. J., Sariola, V., et al. (2022). “Being Watched by a Humanoid Robot and a Human: Effects on Affect-Related Psychophysiological Responses”. *Biological Psychology*, (175):108451.

King, P. (2009). “Emotions in Medieval Thought”. In: *The Oxford Handbook of Philosophy of Emotion*. Goldie, P., ed. Oxford University Press, Oxford, pp. 167–188.

— (2012). “Emotions”. In: *The Oxford Handbook of Aquinas*. Davies, B., ed. Oxford University Press, Oxford, pp. 209–226.

Kirby, R., Forlizzi, J. and Simmons, R. (2010). “Affective Social Robots”. *Robotics and Autonomous Systems*, (58)3:322–332. (Visited on 08/11/2025).

Kircher, A. (1650). *Musurgia universalis sive ars magna consoni et dissoni*. Francesco Corbelletti, Rome.

Kivy, P. (1980). *The Corded Shell. Reflections on Musical Expression*. Princeton University Press, Princeton (NJ).

— (1989). *Sound Sentiment: An Essay on the Musical Emotions, Including the Complete Text of the Corded Shell*. Temple University Press, Philadelphia.

— (2002). *Introduction to a Philosophy of Music*. Clarendon Press.

— (2007). *Music, Language, and Cognition: And Other Essays in the Aesthetics of Music*. Clarendon Press.

Klausen, T. A., Farhadi, U., Vlachos, E., et al. (2022). “Signalling Emotions with a Breathing Soft Robot”. In: *RoboSoft 2022: Proceedings of the 5th IEEE International Conference on Soft Robotics (RoboSoft), 4-8 April, 2022, Edinburgh*. IEEE, ed. IEEE, New York, pp. 194–200.

Klein, J., Moon, Y. and Picard, R. W. (2002). “This Computer Responds to User Frustration: Theory, Design, and Results”. *Interacting with Computers*, (14)2:119–140.

Klein, S. B. and Nichols, S. (2012). “Memory and the Sense of Personal Identity”. *Mind*, (121)483:677–702.

Kline, R. R. (2015). *The Cybernetics Moment: Or Why We Call Our Age the Information Age*. Johns Hopkins University Press, Baltimore.

Klüber, K., Schwaiger, K. and Onnasch, L. (2025). “Affect-Enhancing Speech Characteristics for Robotic Communication”. *International Journal of Social Robotics*, (17)2:315–333.

Klüver, H. and Bucy, P. C. (1939). “Preliminary Analysis of Functions of the Temporal Lobes in Monkeys”. *Archives of Neurology & Psychiatry*, (42)6:979–1000.

Knudsen, E. I. and Knudsen, P. F. (1985). “Vision Guides the Adjustment of Auditory Localization in Young Barn Owls”. *Science*, (230):545–548.

Koelsch, S. (2005). “Neural Substrates of Processing Syntax and Semantics in Music”. *Current Opinion in Neurobiology*, (15)2:207–212.

—, Fritz, T., Cramon, D. Y., et al. (2006). “Investigating Emotion with Music: An fMRI Study”. *Human Brain Mapping*, (27)3:239–250.

Koetsier, T. (2001). “On The Prehistory of Programmable Machines: Musical Automata, Looms, Calculators”. *Mechanism and Machine Theory*, (36)5:589–603.

Koffka, K. (1936). *Principles of Gestalt Psychology*. Lund Humphries, London.

Köhler, W. (1930). *Gestalt Psychology*. Bell, London.

Konijn, E. A. and Hoorn, J. F. (2020). “Differential Facial Articulacy in Robots and Humans Elicit Different Levels of Responsiveness, Empathy, and Projected Feelings”. *Robotics*, (9)4:92.

Kowalczuk, Z. and Czubenko, M. (2016). “Computational Approaches to Modeling Artificial Emotion – An Overview of the Proposed Solutions”. *Frontiers in Robotics and AI*, (3):21.

Kravljaca, N. and Stanley, J. (2024). “Smile, You’ll Feel Better: Testing the Facial Feedback Hypothesis in Young and Older Adults”. *Innovation in Aging*, (8)Suppl 1:608.

Kreibig, S. D. (2010). “Autonomic Nervous System Activity in Emotion: A Review”. *Biological Psychology*, (84)3:394–421.

Kremer, W. F. (1947). "Autonomic and Somatic Reactions Induced by Stimulation of the Cingular Gyrus in Dogs". *Journal of Neurophysiology*, (10)5:371–379.

Kripke, S. A. (2011). *Philosophical Troubles: Collected Papers*. Vol. 1. Oxford University Press, Oxford.

Krumhansl, C. L. (1997). "An Exploratory Study of Musical Emotions and Psychophysiology". *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Experimentale*, (51)4:336–353.

Krumhuber, E. G., Hyniewska, S. and Orlowska, A. (2023). "Contextual Effects on Smile Perception and Recognition Memory". *Current Psychology*, (42)8:6077–6085.

—, Likowski, K., and Weyers, P. (2014). "Facial Mimicry of Spontaneous and Deliberate Duchenne and Non-Duchenne Smiles". *Journal of Nonverbal Behavior*, (38).

Krzyzaniak, M. (2012). "Prehistory of Musical Robots". *Journal of Human-Robot Interaction*, (1)1:78–95.

Kuhn, T. S. (1962). *The Structure of Scientific Revolution*. The University of Chicago Press, Chicago.

Kühne, R. and Peter, J. (2023). "Anthropomorphism in Human-Robot Interactions: A Multidimensional Conceptualization". *Communication Theory*, (33)1:42–52.

Kusuda, Y. (2008). "Toyota's Violin-Playing Robot". *Industrial Robot. International Journal of Robotics Research and Application*, (35)6:504–506.

Laird, A. (2009). "The Von Neumann Architecture". *Computer Science*, (319):360–8771.

Lakatos, G., Gácsi, M., Konok, V., et al. (2014). "Emotion Attribution to a Non-Humanoid Robot in Different Social Situations". *PloS One*, (9)12:e114207.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago.

Lamb, R. (1987). "Objectless Emotions". *Philosophy and Phenomenological Research*, (48)1:107.

Lane, T. (2018). "A Short History of Robotic Surgery". *The Annals of The Royal College of Surgeons of England*, (100)6:5–7.

Lange, C. G. (1885). *Om sindsbevaegelser; et psyko-fysiologisk studie*. Ger. tr. by Kurella, H. (1887) *Über Gemütsbewegungen: Eine psycho-physiologische Studie*. Theodor Thomas Verlag, Leipzig. Fr. tr. (from Ger.) by Dumas, G. (1895). *Les émotions: étude psychophysiologique*. Alcan, Paris. Jacob Lunds Forlag, Kjobenhavn.

Langer, S. K. (1941). *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art*. Harvard University Press, Cambridge (MA).

— (1979). *Feeling and Form: A Theory of Art*. Routledge and Kegan, New York.

Langevin, G. (2012). *InMoov. Open-Source 3D Printed Life-Size Robot*. https://inmoov.fr/.

Langley, J. N. (1889). "On the Physiology of the Salivary Secretion". *Journal of Physiology*, (10):245–310.

Larsen, J. T. and McGraw, A. P. (2014). "The Case for Mixed Emotions". *Social and Personality Psychology Compass*, (8)6:263–274.

Lazarus, R. S. (1974). "Psychological Stress and Coping in Adaptation and Illness". *The International Journal of Psychiatry in Medicine*, (5)4:321–333.

— (1982). "Thoughts on the Relations Between Emotion and Cognition". *American Psychologist*, (37)9:1019.

— (1984). "On the Primacy of Cognition". *American Psychologist*, (39):124–129.

— (1991). *Emotion and adaptation*. Oxford University Press, Oxford.

LeCun, Y. (1985). "A Learning Scheme for Asymmetric Threshold Networks". *Cognitiva 1985: Proceedings, 4-7 June, 1985, Paris*, (85)537:599–604.

Lederberg, J. (1964). *DENDRAL-64. A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs. Part I - Notational Algorithm for Tree Structures*. Tech. rep. NASA Interim Report CR.57029.

Lederberg, J. (1965). *DENDRAL-64. A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs. Part II -Topology of Cyclic Graphs*. Tech. rep. NASA Interim Report CR.68898.

Lederberg, J. (1969). *DENDRAL-64. A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs. Part III - Complete Chemical Graphs; Embedding Rings in Trees*. Tech. rep. NASA Interim Report CR.123176.

LeDoux, J. E. (1984). "Cognition and Emotion". In: *Handbook of Cognitive Neuroscience*. Gazzaniga, M. S., ed. Springer, Boston, pp. 357–368.

— (1986). "Neurobiology of Emotion". In: *Mind and Brain*. LeDoux, J. E. and Hirst, W., eds. Cambridge University Press, New York.

— (1998). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster, Inc., New York.

— (2014). "Coming to Terms with Fear". *Proceedings of the National Academy of Sciences of the United States of America*, (111)8:2871–2878.

— and Hirst, W., eds. (1986). *Mind and Brain*. Cambridge University Press, New York.

Lemmings, D. and Brooks, A. (2014). "The Emotional Turn in the Humanities and Social Sciences". In: *Emotions and Social Change. Historical and Sociological Perspectives*. Lemmings, D. and Brooks, A., eds. Routledge, New York, pp. 3–18.

Leventhal, H. (1980). "Toward a Comprehensive Theory of Emotion". In: *Advances in Experimental Social Psychology*. Vol. 13. Elsevier, pp. 139–207.

Levinson, J., ed. (2011). *Music, Art, and Metaphysics: Essays in Philosophical Aesthetics*. Oxford University Press, Oxford.

— (2018). *Music in the Moment*. Cornell University Press, Ithaca (NY).

Lewin, K. Z. (1936). *A Dynamic Theory of Personality: Selected Papers*. Eng. tr. by Adam, Donald D. and Zenzer, Karl F. Mcgraw-Hill Book Company Inc., New York.

— (1947). "Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change". *Human Relations*, (1)1:5–41.

Lewis, D. (1983). "Extrinsic Properties". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, (44)2:197–200.

Lewis, M. and Cãnamero, L. (2013). "Are Discrete Emotions Useful in Human-Robot Interaction? Feedback from Motion Capture Analysis". In: *ACII 2013: Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction, 2-5 September, 2013, Geneva, Switzerland*. Pun, T. and Nijholt, A., eds. IEEE, New York, pp. 97–102.

Lim, H.-o. and Takanishi, A. (2007). "Biped Walking Robots Created at Waseda University". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, (365)1850:49–64.

Lin, J., Spraragen, M. and Zyda, M. (2012). "Computational Models of Emotion and Cognition". *Advances in Cognitive Systems*, (2)1:59–76.

Lindquist, K. A., Barrett, L. F., Bliss-Moreau, E., et al. (2006). "Language and the Perception of Emotion." *Emotion*, (6)1:125–138.

—, Wager, T. D., Kober, H., et al. (2012). "The Brain Basis of Emotion: A Meta-Analytic Review". *Behavioral and Brain Sciences*, (35)3:121–143.

Linnainmaa, S. (1970). "Alogritmin kumulatiivinen pyöristysvirhe yksittäisten pyöristysvirheiden taylor-kehitelmänä (The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors)". PhD thesis. Helsinki: University of Helsinki.

Liu, X., Dong, J. and Jeon, M. (2023). "Robots' "Woohoo" and "Argh" Can Enhance Users' Emotional and Social Perceptions: An Exploratory Study on Non-lexical Vocalizations and Nonlinguistic Sounds". *ACM Transactions on Human-Robot Interaction*, (12)4:1–20.

Lloyd, S. (1982). "Least Squares Quantization in PCM". *IEEE Transactions on Information Theory*, (28)2:129–137.

Locke, J. (1689). *An Essay Concerning Human Understanding*. Thomas Basset, London.

Loehlin, J. C. (1962). "The Personality of Aldous". *Discovery*, (23):23–26.

— (1963). "A Computer Program That Simulates Personality". In: *Computer Simulation of Personality*. Tomkins, S. S. and Messick, S. S., eds. John Wiley & Sons, New York, pp. 189–211.

— (1965). ""Interpersonal" Experiments with a Computer Model of Personality." *Journal of Personality and Social Psychology*, (2)4:580–584.

— (1968). *Computer Models of Personality*. Random House, New York.

Logins, A. (2024). "Evidence and Emotions: The Problem of Propositional Justification". *American Philosophical Quarterly*, (61)2:99–108.

Lombardi, G., Sciutti, A., Rea, F., et al. (2024). "Humanoid Facial Expressions as a Tool to Study Human Behaviour". *Scientific Reports*, (14)1:133.

Lombardi, M., Roselli, C., Kompatsiari, K., et al. (2023). "The Impact of Facial Expression and Communicative Gaze of a Humanoid Robot on Individual Sense of Agency". *Scientific Reports*, (13)1:10113.

Long, J. H., Schumacher, J., Livingston, N., et al. (2006). "Four Flippers or Two? Tetrapodal Swimming with an Aquatic Robot". *Bioinspiration & Biomimetics*, (1)1:20.

Long, M. (2022). "Cultural History and Modern Science". *The Historical Journal*, (65)3:856–868. (Visited on 01/13/2026).

Lorenz, K. (1981). *The Foundations of Ethology*. Springer, Berlin.

Lovelace, A. K. (1843). "Sketch of the Analytical Engine Invented by Charles Babbage with Notes by the Translator". *Scientific Memoirs*, (3)29:666–731.

Lutz, C. A. (1988). *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago Press, Chicago.

Ma, H. and Yarosh, S. (2023). "A Review of Affective Computing Research Based on Function-Component-Representation Framework". *IEEE Transactions on Affective Computing*, (14)2:1655–1674.

MacDorman, K. F. (2005a). "Androids as an Experimental Apparatus: Why is There an Uncanny Valley and Can We Exploit It". *Android Science*, (3):106–118.

— (2005b). "Mortality Salience and the Uncanny Valley". In: *IEEE-RAS 2005: Proceedings of the 5th International Conference on Humanoid Robots, 5-7 December, 2005, Tsukuba, Japan*. IEEE, ed. IEEE, New York, pp. 399–405.

— (2006). "Subjective Ratings of Robot Video Clips for Human Likeness, Familiarity, and Eeriness: An Exploration of the Uncanny Valley". In: *ICCS-CogSci 2006: Proceedings of the Long Symposium Toward Social Mechanisms of Android Science, 26 July, 2006. Vancouver, Canada*. Cowley, S., Kerstin, D., Hiroshi, I., et al., eds., pp. 26–29.

— and Chattopadhyay, D. (2016). "Reducing Consistency in Human Realism Increases the Uncanny Valley Effect; Increasing Category Uncertainty Does Not". *Cognition*, (146):190–205.

— and Entezari, S. O. (2015). "Individual Differences Predict Sensitivity to the Uncanny Valley". *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, (16)2:141–172.

— and Ishiguro, H. (2006). "The Uncanny Advantage of Using Androids in Cognitive and Social Science Research". *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, (7)3:297–337.

— and Kageki, N. (2012). "The Uncanny Valley [From the Field]". *IEEE Robotics & Automation Magazine*, (19)2:98–100. Eng. tr. of Mori, M. (1970). 不気味の谷 [*Bukimi No Tani Gensho*]. *Energy*, 7(4):33–35.

MacLean, P. D. (1949). "Psychosomatic Disease and the" Visceral Brain": Recent Developments Bearing on the Papez Theory of Emotion". *Biopsychosocial Science and Medicine*, (11)6:338–353.

— (1952). "Some Psychiatric Implications of Physiological Studies on Frontotemporal Portion of Limbic System (visceral Brain)". *Electroencephalography and Clinical Neurophysiology*, (4)4:407–418.

— (1955). "The Limbic System (visceral Brain) and Emotional Behavior". *AMA Archives of Neurology & Psychiatry*, (73)2:130–134.

Maes, P. (1991). "Learning Behavior Networks from Experience". In: *ECAL 1991: Proceedings of the 1st European Concerence on Artificial Life (Toward a Practice of Autonomous Systems), 11-13 December, 1991, Paris*. Varela, F. J. and Bourgine, P., eds. Bradford Books, Bradford (PA), pp. 48–57.

Mahon, J. E. (2016). "The Definition of Lying and Deception". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Metaphysics Research Lab, Stanford University.

Maier, M., Leonhardt, A., Blume, F., et al. (2025). "Neural Dynamics of Mental State Attribution to Social Robot Faces". *Social Cognitive and Affective Neuroscience*, (20)1.

Mäkäräinen, M., Kätsyri, J. and Takala, T. (2014). "Exaggerating Facial Expressions: A Way to Intensify Emotion or a Way to the Uncanny Valley?" *Cognitive Computation*, (6)4:708–721.

Malfaz, M. and Salichs, M. A. (2006). "Using Emotions for Behavior-Selection Learning". In: *ECAI 2006: Proceedings of the 17thEuropean Conference on Artificial Intelligence, 29 August–1 September 2006, Riva del Garda, Italy*. Brewka, G., Coradeschi, S., Perini, A., et al., eds. IOS Press, Amsterdam, pp. 697–698.

Malle, B. F. (2007). "Attributions as Behaviour Explanations: Towards a New Theory". In: *Current Themes and Perspectives in Social Psychology*. Chadee, D. and Hunter, J., eds. SOCS, The University of the West Indies, Augustine Trinidad, pp. 3–26.

— (2011). "Time to Give Up the Dogmas of Attribution". In: *Advances in Experimental Social Psychology*. Zanna, M. P., ed. Vol. 44. Elsevier, pp. 297–352.

— (2022). "Attribution Theories: How People Make Sense of Behavior". In: *Theories in Social Psychology*. Chadee, D., ed. John Wiley & Sons.

Mandler, G. (1975). *Mind and Emotion*. John Wiley & Sons, New York.

Manzi, F., Massaro, D., Di Lernia, D., et al. (2021). "Robots Are Not All the Same: Young Adults' Expectations, Attitudes, and Mental Attribution to Two Humanoid Social Robots". *Cyberpsychology Behavior and Social Networking*, (24)5:307–314.

Margócsy, D. (2015). "Androids in the Enlightenment: Mechanics, Artisans, and Cultures of the Self". *Annals of Science*, (72)3.

Marinier, R. P., Laird, J. E. and Lewis, R. L. (2009). "A Computational Unification of Cognitive Behavior and Emotion". *Cognitive Systems Research*, (10)1:48–69.

Marinier, R. P. and Laird, J. E. (2008). "Emotion-Driven Reinforcement Learning". In: *CogSci 2008: Proceedings of the 30thAnnual Meeting of the Cognitive Science Society, 23-26 July 2008, Washington, DC, USA*. Love, B. C., McRae, K. and Sloutsky, V. M., eds. Vol. 30. Curran Associates, Inc., Red Hook (NY), pp. 115–120.

Marjanović, M. J., Scassellati, B. and Williamson, M. M. (1996). "Self-Taught Visually-Guided Pointing for a Humanoid Robot". In: *From Animals to Animats 4. Proceedings of the Fourth International Conference on Simulation of Adpative Behavior, 9-13 September, 1996, Cape Cod (MA)*. Cliff, D., Husbands, P., Meyer, J.-A., et al., eds. MIT Press, Cambridge (MA), pp. 35–44.

Marmpena, M., Garcia, F. and Lim, A. (2020). "Generating Robotic Emotional Body Language of Targeted Valence and Arousal with Conditional Variational Autoencoders". In: *HRI 2020: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, 23-26 March, 2020, Cambridge (UK)*. Young, J. O. and Belpaeme, T., eds. ACM/IEEE, New York, pp. 357–359.

—, Lim, A., and Dahl, T. S. (2018). "How Does the Robot Feel? Perception of Valence and Arousal in Emotional Body Language". *Paladyn, Journal of Behavioral Robotics*, (9)1:168–182.

Marr, D. (1976). "Early Processing of Visual Information". *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, (275)942:483–519.

— (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco.

Marsella, S. C. (2003). "Interactive Pedagogical Drama: Carmen's Bright IDEAS Assessed". *Intelligent Virtual Agents, Springer*: 183–194.

—, Gratch, J., and Petta, P. (2010). "Computational Models of Emotion". In: *A Blueprint for Affective Computing - A Sourcebook and Manual*. Scherer, K. R., Banziger, T. and Roesch, E., eds. Oxford University Press, Oxford, pp. 21–46.

Marshall, D. and Weatherson, B. (2023). "Intrinsic vs. Extrinsic Properties". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Fall 2023. Metaphysics Research Lab, Stanford University.

Mataric, M. J. (2007). *The Robotics Primer*. MIT Press, Cambridge (MA).

Mattheson, J. (1739). *Der vollkommene Capellmeister*. Christian Herold, Hamburg.

Maxwell, J. C. (1867). "On Governors". *Proceedings of the Royal Society of London*, (16):270–283.

McCarthy, J. (1986). "Applications of Circumscription to Formalizing Common-Sense Knowledge". *Artificial intelligence*, (28)1:89–116.

— and Hayes, P. (1969). "Some Philosophical Problems From the Standpoint of Artificial Intelligence". In: *Machine Intelligence 4*. Meltzer, B. and Michie, D., eds. Edinburgh University Press, pp. 463–502.

McCulloch, W. S. and Pitts, W. (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". *The Bulletin of Mathematical Biophysics*, (5):115–133.

McDermott, D. (1987). "We've Been Framed: Or, Why AI Is Innocent of the Frame Problem". In: *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Pylyshyn, Z. W., ed. Ablex Publishing, New York, pp. 113–122.

McDougall, W. (1921). *An Introduction to Social Psychology*. John W. Luce & Co., Boston.

McGrath, M. and Frank, D. (2024). "Propositions". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Fall 2024. Metaphysics Research Lab, Stanford University.

McIntosh, D. N. (1996). "Facial Feedback Hypotheses: Evidence, Implications, and Directions". *Motivation and Emotion*, (20):121–147.

Mei, S., Marsella, S. C. and Pynadath, D. V. (2006). "Thespian: Modeling Socially Normative Behavior in a Decision-Theoretic Framework". In: *IVA 2006: Proceedings of the Sixth International Conference on Intelligent Virtual Agents, 21-23 August, 2006, Marina Del Rey (CA)*. Gratch, J., Young, M., Aylett, R., et al., eds. Springer, Berlin, pp. 369–382.

Meinertsen, B. R. (2024). "Relational Properties: Definition, Reduction, and States of Affairs". *Ratio*, (37)2-3:178–190. (Visited on 09/08/2025).

Menabrea, L. F. (1842). *Notions sur la machine analytique de M. Charles Babbage*. Bibliothèque Universelle de Genève, Genève.

Menary, R., ed. (2010). *The Extended Mind*. MIT Press, Cambridge (MA).

Mendez, M. F. and Shapira, J. S. (2009). "Altered Emotional Morality in Frontotemporal Dementia". *Cognitive Neuropsychiatry*, (14)3:165–179.

Mendonça, D. (2013). "Emotions about Emotions". *Emotion Review*.

Menninghaus, W., Wagner, V., Wassiliwizky, E., et al. (2019). "What Are Aesthetic Emotions?" *Psychological review*, (126)2:171.

Merleau-Ponty, M. (1942). *La Structure du comportement*. Presses Universitaires de France, Paris.

— (1945). *Phénoménologie de la perception*. Gallimard, Paris.

— (1964). *L'Œil et l'Esprit*. Gallimard, Paris.

Merriam, A. P. (1964). *The Anthropology of Music*. Northwestern University Press, Evanston, Illinois.

Meyer, J.-J. (2006). "Reasoning about Emotional Agents". *International Journal of Intelligent Systems*, (21):601–619.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. University of Chicago Press, Chicago.

— (1957). "Meaning in Music and Information Theory". *The Journal of Aesthetics and Art Criticism*, (15)4:412–424.

— (1967). *Music, the Arts, and Ideas: Patterns and Predictions in Twentieth-Century Culture*. University of Chicago Press, Chicago.

Michaelian, K. and Sutton, J. (2017). "Memory". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Summer 2017. Metaphysics Research Lab, Stanford University.

Mignard, M. (1909). *La Joie Passive: Étude de Psychologie Pathologique*. Alcan, Paris.

Miller, G. A. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological review*, (63)2:81.

— (1986). "Interview with George A. Miller". In: In Baars, Bernard J. (1986). *The Cognitive Revolution in Psychology*. The Guilford Press, New York-London, pages 200–223.

— (2003). "The Cognitive Revolution: A Historical Perspective". *Trends in Cognitive Sciences*, (7)3:141–144.

—, Galanter, E., and Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Holt, Rinehart and Winston, Inc., New York.

Minsky, M. L. (1952). "A Neural-Analogue Calculator Based Upon a Probability Model of Reinforcement". *Harvard University Psychological Laboratories*. Cambridge (MA).

— (1985). *The Society of Mind*. Simon & Schuster, Inc., New York.

— (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon and Schuster, New York.

— and Papert, S. (1969). *An Introduction to Computational Geometry*. MIT Press, Cambridge (MA).

Mishra, C., Skantze, G., Hagoort, P., et al. (2024). "Perception of Emotions in Human and Robot Faces: Is the Eye Region Enough?" In: *ICSR 2024: Proceedings of the 16th International Conference on Social Robotics and AI, 23-26 October, 2024, Odense, Denmark*. Palinko, O., Bodenhagen, L., Cabibihan, J.-J., et al., eds. Springer, New York, pp. 290–303.

Mitchell, J. (2019). "Emotional Experience and Propositional Content". *Dialectica*, (73)4:535–561.

— (2020). "On the Non-Conceptual Content of Affective-Evaluative Experience". *Synthese*, (197)7:3087–3111.

Moffat, D. and Frijda, N. H. (1994). "Where There's a Will There's an Agent". In: *IJCAI-ATAL 1995: Proceedings of the International Workshop on Agent Theories, Architectures, and Languages, 19-20 August, 1999, Montreal, Canada*. Wooldridge, M. and Jennings, N. R., eds. Springer, Berlin, pp. 245–260.

Moors, A. (2022). *Demystifying Emotions: A Typology of Theories in Psychology and Philosophy*. Cambridge University Press.

Moravec, H. (1976). "Cart Project Progress Report". Notes NASA contract NASW 2916.

Moravec, H. P. (1982). "The CMU Rover". In: *AAAI-82: Second National Conference on Artificial Intelligence, August 18-20, 1982, Carnegie Mellon University and University of Pittsburgh, Pittsburgh*. Morgan Kaufmann Publishers, Los Altos (CA), pp. 377–380.

— (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge (MA).

— (1999). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press, Oxford.

Mori, M. (1970). "不気味の谷 [*Bukimi No Tani Gensho*]". *Energy*, (7)4:33–35. Eng. tr. by MacDorman, K. F. and Minato, T. (2012). *The Uncanny Valley [From the Field]*. IEEE Robotics & Automation Magazine, (19)2:98–100.

Morita, T., Iwata, H. and Sugano, S. (1999). "Development of Human Symbiotic Robot: WENDY". In: *ICRA 1999: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, 10-15 May, 1999, Detroit, MI, USA*. IEEE, ed. Vol. 4. IEEE, New York, pp. 3183–3188.

—, Shibuya, K., and Sugano, S. (1998). "Design and Control of Mobile Manipulation System for Human Symbiotic Humanoid: Hadaly-2". In: *ICRA 1998: Proceedings of the 1998 IEEE International Conference on Robotics and Automation, 20-20 May 1998, Katholieke Universiteit Leuven, Leuven, Belgium (Cat. No.98CH36146)*. IEEE, ed. Vol. 2. IEEE, New York, pp. 1315–1320.

Morton, A. (2009). "Epistemic Emotions". In: *The Oxford Handbook of Philosophy of Emotion*. Goldie, P., ed. Oxford University Press, Oxford, pp. 385–400.

Moser, U., Pfeifer, R., Schneider, W., et al. (1982). "Experiences with Computer Simulation of Dream Processes". In: *ESRS 1982: Proceedings of the 6th European Congress on Sleep Research, 23-26 March, 1982, Zürich*. Koella, W. P., ed. Karger, Basel, pp. 30–44.

—, Zeppelin, I. von, and Schneider, W. (1970). "Computer Simulation of a Model of Neurotic Defense Processes". *Behavioral Science*, (15):194–202.

Mosso, A. (1884). *La paura*. Treves, Milano.

Mueller, E. T. and Dyer, M. G. (1985). "Daydreaming in Humans and Computers". In: *IJCAI 1985: Proceedings of the 9th International Joint Conference on Artificial Intelligence, 18-23 August, 1985, Los Angeles, CA*, pp. 278–280.

Müller, J. M. (2017). "How (Not) to Think of Emotions as Evaluative Attitudes". *Dialectica*, (71)2:281–308.

— (2020). "Response-Dependent Normative Properties and the Epistemic Account of Emotion". *The Journal of Value Inquiry*, (54)3:355–364.

Murillo-Lara, A. and Muñoz-Serna, C. A. (2019). "An Approach to the Non-Conceptual Content of Emotions". *Revista de Filosofia Aurora*, (31)54.

Murphy, F. C., Nimmo-Smith, I. and Lawrence, A. D. (2003). "Functional Neuroanatomy of Emotion: A Meta-Analysis". *Cognitive, Affective, and Behavioral Neuroscience*, (3):207–233.

Mūsā, i. S. B. (1979). الهندسية الحيل معرفة في كتاب [*Kitab fi Ma'rifat al-Hiyal al-Handasiyya*]. Eng. tr. by Hill, D.R. (1979). The Book Knowledge of Ingenious Mechanical Devices. D. Reidel Publishing Company, Dordrecht, Boston and London.

Musson, A. E. and Robinson, E. (1989). *Science and Technology in the Industrial Revolution*. Taylor & Francis, New York.

Nääs, H., Thellman, S. and Ziemke, T. (2025). "Human Preferences for Cognitive and Emotional Capabilities in Robots Across Different Application Domains". *Frontiers in Robotics and AI*, (12).

Nagel, T. (1974). "What Is It Like to Be a Bat?" *The Philosophical Review*, (83)4:435–450.

Nahlowsky, J. W. (1884). *Das Gefühlsleben in Seinen Wesentlichsten Erscheinungen*. Verlag von Veit & Comp., Leipzig.

Nanay, B. (2020). "Vicarious Representation: A New Theory of Social Cognition". *Cognition*, (205):104451.

El-Nasr, M. and Seif, T. R. (2000). "FLAME-Fuzzy Logic Adaptive Model of Emotions". *Autonomous Agents and Multi-Agent Systems*, (3):219–257.

Neches, R. (1981). "Hpm: A Computational Formalism for Heuristic Procedure Modification". In: *IJCAI 1981: Proceedings of the 7th International Joint Conference on Artificial Intelligence, 24-28 August, 1981, University of British Columbia*. Vol. 1. Morgan Kaufmann Publishers, San Francisco, pp. 283–288.

Negroponte, N. (1995). *Being Digital*. Alfred A. Knopf, Inc., New York.

Neisser, U. G. (1963). "The Imitation of Man by Machine: the View That Machines Will Think as Man Does Reveals Misunderstanding of the Nature of Human Thought". *Science*, (139)3551:193–197.

— (1967). *Cognitive Psychology*. Meredith Publishing Company, New York.

Nettl, B. (1983). *The Study of Ethnomusicology*. University of Illinois Press, Chicago.

Newbold, J. W., Hunt, A. and Brereton, J. (2015). "Chemical Spectral Analysis Through Sonification". In: Institute of Electronic Music and Acoustics (IEM).

Newell, A., Shaw, J. C. and Simon, H. A. (1959). "Report on a General Problem Solving Program". In: *IFIP 1959: Proceedings of the 1st International Conference on Information Processing, UNESCO, 15-20 June, 1959, Paris*. Vol. 256, p. 64.

— and Simon, H. A. (1956). "The Logic Theory Machine-A Complex Information Processing System". *IRE Trans. Inf. Theory*, (2):61–79.

— and Simon, H. A. (1961). "Computer Simulation of Human Thinking: A Theory of Problem Solving Expressed as a Computer Program Permits Simulation of Thinking Processes." *Science*, (134)3495:2011–2017.

Niedenthal, P. M., Winkielman, P., Mondillon, L., et al. (2009). "Embodiment of Emotion Concepts". *Journal of Personality and Social Psychology*, (96)6:1120–1136.

Nietzsche, F. (1889). *Götzen-Dämmerung oder Wie man mit dem Hammer philosophirt*. Verlag C. G. Naumann, Leipzig.

Nilsson, N. J., ed. (1984). *Shakey the Robot*, Technical Note 323, Artificial Intelligence Center Computer Science and Technology Division, SRI International, University of Stanford, Menlo Park (CA).

— (2009). *The Quest for Artificial Intelligence*. Cambridge University Press, Cambridge.

Nussbaum, M. C. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press, Cambridge.

Oatley, K. and Johnson-Laird, P. N. (1987). "Towards a Cognitive Theory of Emotions". *Cognition and Emotion*, (1):29–50.

Öhman, A., Carlsson, K., Lundqvist, D., et al. (2007). "On the Unconscious Subcortical Origin of Human Fear". *Physiology & Behavior*, (92)1:180–185.

Ojha, S., Vitale, J. and Williams, M.-A. (2021). "Computational Emotion Models: A Thematic Review". *International Journal of Social Robotics*, (13)6:1253–1279.

Okon-Singer, H., Hendler, T., Pessoa, L., et al. (2015). "The Neurobiology of Emotion-Cognition Interactions". *Frontiers in Human Neuroscience*, (9):58.

Ong, D. C., Zaki, J. and Goodman, N. D. (2019). "Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap". *Topics in Cognitive Science*, (11)2:338–357.

Onnasch, L., Kostadinova, E. and Schweidler, P. (2022). "Humans Can't Resist Robot Eyes – Reflexive Cueing With Pseudo-Social Stimuli". *Frontiers in Robotics and AI*, (9).

Ortony, A., Clore, G. L. and Collins, A. (1987). "The Cognitive Structure of Emotions". *Cognition and Emotion*, (1):5–37.

—, Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.

Palomäki, J., Kunnari, A., Drosinou, M., et al. (2018). "Evaluating The Replicability Of The Uncanny Valley Effect". *Heliyon*, (4)11.

Panksepp, J. (1982). "Toward a General Psychobiological Theory of Emotions". *Behavioral and Brain Sciences*, (5)3:407–422.

— (1988). "Brain Emotional Circuits and Psychopathologies". In: *Emotions and Psychopathology*. Clynes, M. and Panksepp, J., eds. Springer, Boston, pp. 37–76.

— (1989a). "The Neurobiology of Emotions: Of Animal Brains and Human Feelings". In: *Handbook of Social Psychophysiology*. Wagner, H. and Manstead, A., eds. John Wiley & Sons, Oxford, pp. 5–26.

— (1989b). "The Psychobiology of Emotions: The Animal Side of Human Feelings". *Experimental Brain Research Series* 18:31–55.

— (1995). "The Emotional Sources of "Chills" Induced by Music". *Music Perception*, (13)2:171–207.

— (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, Oxford.

Papadopoulos, E. (2007). "Heron of Alexandria (c. 10-85 AD)". In: *Distinguished Figures in Mechanism and Machine Science: Their Contributions and Legacies, Part 1*. Ceccarelli, M., ed. Springer Netherlands, Dordrecht, pp. 217–245.

Papez, J. W. (1937). "A Proposed Mechanism of Emotion". *Archives of Neurology & Psychiatry*, (38)4:725–743.

Pardo, J. V., Pardo, P. J. and Raichle, M. E. (1993). "Neural Correlates of Self-Induced Dysphoria". *The American Journal of Psychiatry*, (150)5:713–719.

Park, G., Vasey, M. W., Kim, G., et al. (2016). "Trait Anxiety Is Associated with Negative Interpretations When Resolving Valence Ambiguity of Surprised Faces". *Frontiers in Psychology*, (7):1164.

Parkinson, B. (1996). "Emotions Are Social". *British Journal of Psychology*, (87)4:663–683.

—, Fischer, A., and Manstead, A. S. R. (2005). *Emotion in Social Relations: Cultural, Group, and Interpersonal Processes*. Psychology Press, New York.

Pavlov, I. P. (1927). *Лекции о работе больших полушарий головного мозга* [*Lekcii o rabote bol'shih polusharij golovnogo mozga*]. State Publishing House, Moscow. Eng. tr. by Anrep, G. V. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, London.

— and Schumows-Simanowsky, E. O. (1895). "Die Innervation der Magendrüsen Beim Hunde". *Archiv für Physiologie*: 59–78.

Paz, E. B., Ceccarelli, M., Otero, J. E., et al. (2010). *A Brief Illustrated History of Machines And Mechanisms*. Springer Science & Business Media, Dordrecht.

Peak, H. (1955). "Attitude and Motivation". In: *Nebraska Symposium on Motivation*. Jones, M., ed. University of Nebraska Press, Lincoln, pp. 149–189.

Pei, G., Li, H., Lu, Y., et al. (2024). "Affective Computing: Recent Advances, Challenges, and Future Trends". *Intelligent Computing*, (3):0076.

Peña, D. and Tanaka, F. (2020). "Human Perception of Social Robot's Emotional States via Facial and Thermal Expressions". *ACM Transactions on Human-Robot Interaction*, (9)4:1–19.

Peretz, I. (1990). "Processing of Local and Global Musical Information by Unilateral Brain-Damaged Patients". *Brain: A Journal of Neurology*, (113)4:1185–1205.

Perez-Gaspar, L.-A., Caballero-Morales, S.-O. and Trujillo-Romero, F. (2016). "Multimodal Emotion Recognition with Evolutionary Computation for Human-Robot Interaction". *Expert Systems Applications*, (66)C:42–61.

Perregaux, C. (1906). *Les Jaquet-Droz et leurs automate*. Wolfrath & Sperlé, Neuchâtel.

— and Perrot, F. L. (1916). *Les Jaquet-Droz et Leschot*. Attinger, Neuchâtel.

Peskin, J. and Scassellati, B. (1997). "Image Stabilization Through Vestibular and Retinal Feedback". In: *Research Abstracts*. Brooks, R. A., ed. MIT Artificial Intelligence Laboratory, Cambridge (MA).

Pessoa, L. (2017). "A Network Model of the Emotional Brain". *Trends in Cognitive Sciences*, (21)5:357–371.

Petersen, K., Solis, J. and Takanishi, A. (2010). "Musical-Based Interaction System for the Waseda Flutist Robot". *Autonomous Robots*, (28)4:471–488.

Pfeifer, R. (1982). *Cognition and Emotion: An Information Processing Approach*. Tech. rep. CIP Working Paper No. 436. Department of Psychology, Carnegie-Mellon University.

Pfeifer, R. (1988). "Artificial Intelligence Models of Emotion". In: *Cognitive Perspectives on Emotion and Motivation*. Hamilton, V., Bower, G. H. and Frijda, N. H., eds. Vol. 44. NATO ASI Series. Springer, Dordrecht, pp. 287–320.

— (1994). "The "Fungus Eater Approach" to Emotion: A View from Artificial Intelligence". *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, (1)2:42–57.

— and Nicholas, D. W. (1985). "Toward Computational Models of Emotion". In: *Selected and Updated Papers from the Proceedings of the 1982 European Conference on Progress in Artificial Intelligence*. Steels, L. and Campbell, J. A., eds. John Wiley & Sons, New York, pp. 184–192.

Phan, K. L., Wager, T. D., Taylor, S. F., et al. (2002). "Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI". *NeuroImage*, (16):331–348.

Picard, R. W. (1995). "Affective Computing - MIT Media Laboratory Perceptual Computing Section Technical Report No. 321". *Cambridge, MA*, (2139):92.

— (1997). *Affective Computing*. MIT Press, Cambridge (MA).

—, Vyzas, E., and Healey, J. (2001). "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (23)10:1175–1191.

Piéron, H. (1907). "La théorie des émotions et les données actuelles de la physiologie". *Journal de psychologie normale et pathologique*, (4):439–451.

Pitcher, G. (1965). "Emotions". *Mind*, (74)295:326–346.

Pitt, D. (2022). "Mental Representation". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Fall 2022. Metaphysics Research Lab, Stanford University.

Plamper, J. (2010). "The History of Emotions: An Interview with William Reddy, Barbara Rosenwein, and Peter Stearns". *History and Theory*, (49)2:237–265.

Planck, M. (1948). *Wissenschaftliche Selbstbiographie: Mit einem Bildnis und der von Max von Laue gehaltenen Traueransprache*. Barth, Leipzig.

Plato. (1997a). *Complete Works*. Edited by Cooper, John M. Hackett Publishing Company, Indianapolis.

— (1997b). "Phaedrus". In: *Complete Works*. Cooper, J. M., ed. Eng. tr. by Nehamas, A. and Woodruff, P. Hackett Publishing Company, Indianapolis, pp. 506–556.

— (1997c). "Republic". In: *Complete Works*. Cooper, J. M., ed. Eng. tr. by Grube G. M. and Reeve C. D. Hackett Publishing Company, Indianapolis, pp. 971–1223.

Plutchik, R. (1962). *The Emotions: Facts, Theories and a New Model*. Crown Publishing Group/ Random House, New York.

— (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York.

— (1994). *The Psychology and Biology of Emotions*. Harper Collins, New York.

Postman, N. (1992). *Technopoly. The Surrender of Tulture to Technology*. Knopf, New York.

Powers, W. T. (1973). *Behavior: The Control of Perception*. Aldine, Chicago.

Prati, E., Borsci, S., Peruzzini, M., et al. (2022). "A Systematic Literature Review of User Experience Evaluation Scales for Human-Robot Collaboration". In: *Advances in Transdisciplinary Engineering*. Moser, B. R., Koomsap, P. and Stjepandić, J., eds., pp. 13–23.

Preece, J., Rogers, Y., Sharp, H., et al. (1994). *Human-Computer Interaction*. Addison-Wesley Longman Ltd., London.

Premack, D. and Woodruff, G. (1978). "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences*, (1)4:515–526.

Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. Philosophy of mind series. Oxford University Press, Oxford.

Proffitt, D. R. (2006). "Embodied Perception and the Economy of Action". *Perspectives on Psychological Science*, (1)2:110–122.

Putnam, H. (1960). "Minds and Machines". In: *Dimensions of Mind*. Hook, S., ed. New York University Press, New York, pp. 148–178.

Pylyshyn, Z. W. and LePore, E. (1999). *What is Cognitive Science?* Blackwell, Hoboken.

Quantz, J. J. (1752). *Versuch einer Anweisung die Flöte traversiere zu spielen*. Eng. tr. by Reilly, E. R. (1966). *On Playing the Flute*. Faber and Faber, London. Johann Friedrich Voß Verlag, Berlin.

Quesque, F. and Apperly, I. e. (2024). "Defining Key Concepts for Mental State Attribution". *Communications Psychology*, (2)1:29.

Quilty-Dunn, J. and Green, E. J. (2023). "Perceptual Attribution and Perceptual Reference". *Philosophy and Phenomenological Research*, (106)2:273–298.

Quinlan, J. R. (1986). "Induction of Decision Trees". *Machine learning*, (1):81–106.

— (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco.

Quoidbach, J., Gruber, J., Mikolajczak, M., et al. (2014). "Emodiversity and the Emotional Ecosystem". *Journal of Experimental Psychology: General*, (143)6:2057.

Rank, S. (2009). *Behaviour Coordination for Models of Affective Behaviour*. PhD thesis, Technische Universität Wien, Wien.

Ratcliffe, M. (2019). "Emotional Intentionality". *Royal Institute of Philosophy Supplement*, (85):251–269.

Ravasio, M. (2025). "History of Western Philosophy of Music: Antiquity to 1800". In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Metaphysics Research Lab, Stanford University.

Rawal, N. and Stock-Homburg, R. M. (2022). "Facial Emotion Expressions in Human-Robot Interaction: A Survey". *International Journal of Social Robotics*, (14)7:1583–1604.

Rayati, K., Feizi, A., Beigy, A., et al. (2023). "Real-Time Imitation of Human Head Motions, Blinks and Emotions by Nao Robot: A Closed-Loop Approach". In: *ICRoM 2023: Proceedings of the 11th RSI International Conference on Robotics and Mechatronics, 19-21 December, 2023, Tehran, Iran.* IEEE, ed. IEEE, New York, pp. 794–800.

Read, R. and Belpaeme, T. (2014). "Situational Context Directs How People Affectively Interpret Robotic Non-Linguistic Utterances". In: *HRI 2014: Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction, 3-6 March, Bielefeld*. IEEE, ed. ACM/IEEE, New York, pp. 41–48.

Reddy, W. M. (1997). "Against Constructionism: The Historical Ethnography of Emotions". *Current Anthropology*, (38)3:327–351.

— (2001). *The Navigation of Feeling: A Framework for the History of Emotions*. Cambridge University Press, Cambridge.

Reeves, B. and Nass, C. I. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Center for the Study of Language and Information*. Cambridge University Press, Cambridge.

Reichardt, J. (1978). *Robots: Fact, Fiction, and Prediction*. Penguin Books, New York.

Reilly, S. W. (2006). *Modeling What Happens Between Emotional Antecedents and Emotional Consequents*. Charles River Analytics, Cambridge (MA).

Reiter, R. (2001). *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, Cambridge (MA).

Rensink, R. A., O'regan, K. J. and Clark, J. J. (1997). "To See or Not to See: The Need for Attention to Perceive Changes in Scenes". *Psychological Science*, (8)5:368–373.

Repubblica Italiana. (2021). "Decreto-Legge 6 maggio 2021, n. 59: Misure urgenti relative al Fondo complementare al Piano nazionale di ripresa e resilienza e altre misure urgenti per gli investimenti". Gazzetta Ufficiale della Repubblica Italiana, Serie Generale n. 108, 7 maggio 2021. https://www.gazzettaufficiale.it/eli/id/2021/05/07/21G00070/SG.

— (2022). "Decreto Ministeriale 9 aprile 2022, n. 351: Riparto di 2.500 borse di dottorato di durata triennale per la frequenza di percorsi di dottorato in programmi specificamente dedicati e declinati". Ministero dell'Università e della Ricerca. https://www.mur.gov.it/it/atti-e-normativa/decreto-ministeriale-n351-del-09-04-2022.

Reuten, A., Dam, M. van and Naber, M. (2018). "Pupillary Responses to Robotic and Human Emotions: The Uncanny Valley and Media Equation Confirmed". *Frontiers in Psychology*, (9).

Rheinberger, H.-J. (2010). *On Historicizing Epistemology: An Essay*. Stanford University Press, Stanford (CA).

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. Perseus Books, New York.

Ribot, T. (1896). *La psychologie des sentiments*. Félix Alcan, Paris.

Richardson, D. C. and Dale, R. (2005). "Looking to Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension". *Cognitive Science*, (29)6:1045–1060.

Riskin, J. (2021). *The Restless Clock. A History of the Centuries Long Argument over What Makes Living Things Tick*. University of Chicago Press, Chicago and London.

Roberts, R. (2001). *Emotions: An Essay in Aid of Moral Psychology*. Cambridge University Press, Cambridge.

Robinson, A. M. (2018). "Let's Talk About Stress: History of Stress Research". *Review of General Psycholoy*, (22)3:334–342.

Robinson, J., ed. (1997). *Music and Meaning*. Cornell University Press, Ithaca (NY).

— (2005). *Deeper Than Reason: Emotion and Its Role in Literature, Music, and Art*. Clarendon Press-Oxford University Press, Oxford-New York.

Robinson, N. L., Connolly, J., Suddery, G., et al. (2021). "A Humanoid Social Robot to Provide Personalized Feedback for Health Promotion in Diet, Physical Activity, Alcohol and Cigarette Use: A Health Clinic Trial". In: *RO-MAN 2021: Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication, 8 August-12 August, 2021, Vancouver, Canada*. IEEE, ed. IEEE, New York, pp. 720–726.

Rock, I. (1975). *An Introduction to Perception*. Macamillan Publishing Co., Inc.

— (1983). *The Logic of Perception*. Bradford Books, Cambridge.

Roeser, S. and Todd, C. (2014). *Emotion and value*. Oxford University Press.

Rojas-Quintero, J. A. and Rodríguez-Liñán, M. C. (2021). "A Literature Review of Sensor Heads for Humanoid Robots". *Robotics and Autonomous Systems*, (143):103834.

Rolls, E. T. (1990). "A Theory of Emotion, and Its Application to Understanding the Neural Basis of Emotion". *Cognition & Emotion*, (4)3:161–190.

Romashov, V. (June 2018). *How Response Format Artificially Increases Emotion Recognition Rates*. Masters Thesis, Kozminski University, Warsaw.

Rosaldo, M. Z. (1980). *Knowledge and Passion*. Cambridge University Press, Cambridge.

Roselli, C., Lapomarda, L. and Datteri, E. (2025). "How Culture Modulates Anthropomorphism in Human-Robot Interaction: A Review". *Acta Psychologica*, (255):104871.

Roseman, I. J. (1996). "Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory". *Cognition & Emotion*, (10)3:241–278. (Visited on 09/08/2025).

—, Spindel, M. S., and Jose, P. E. (1990). "Appraisals of Emotion-Eliciting Events: Testing a Theory of Discrete Emotions." *Journal of Personality and Social Psychology*, (59)5:899.

Rosenberg, M. J. (1956). "Cognitive Structure and Attitudinal Affect". *The Journal of Abnormal and Social Psychology*, (53)3:367.

— and Abelson, R. P. (1960). "An Analysis of Cognitive Balancing". In: *Attitude Organization and Change: An Analysis of Consistency Among Attitude Components*. Rosenberg, M. J., Hovland, C. I., McGuire, W. J., et al., eds. Yale University Press, New Haven, pp. 112–163.

Rosenblatt, F. (1958). “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. *Psychological review*, (65)6:386.

— (1962). *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. partan Books, Washington, DC.

Rosenthal-von der Pütten, A. M., Arndt, J., Pieczykolan, A., et al. (2025). “Within, Between, Forced Choice, or Likert Scale? How Methodological Decisions Influence Recognition Rates in HRI Recognition Studies”. *International Journal of Social Robotics*, (17)4:693–706.

— and Krämer, N. C. (2014). “How Design Characteristics of Robots Determine Evaluation and Uncanny Valley Related Responses”. *Computers in Human Behavior*, (36):422–439.

—, Schulte, F. P., Eimler, S. C., et al. (2013). “Neural Correlates of Empathy Towards Robots”. In: *HRI 2013: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, 3-6 March, 2013, Tokio*. Kuzuoka, H. and Evers, V., eds. ACM/IEEE, New York, pp. 215–216.

—, Schulte, F. P., Eimler, S. C., et al. (2014). “Investigations on Empathy Towards Humans and Robots Using fMRI”. *Computers in Human Behavior*, (33):201–212.

Rosenwein, B. H. (1998). *Anger’s Past: The Social Uses of an Emotion in the Middle Ages*. Cornell University Press, Ithaca (NY).

— (2002). “Worrying About Emotions in History”. *The American Historical Review*, (107)3:821–845.

Rosheim, M. E. (2006). *Leonardo’s Lost Robots*. Springer, Berlin.

Rossi, M. and Tappolet, C. (2019). “What Kind of Evaluative States Are Emotions? The Attitudinal Theory Vs. the Perceptual Theory of Emotions”. *Canadian Journal of Philosophy*, (49)4:544–563.

Rouse, J. (1993). “What Are Cultural Studies of Scientific Knowledge?” *Configurations*, (1)1:1–22.

Roy, D. and Pentland, A. (1996). “Automatic Spoken Affect Analysis and Classification”. In: *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, 14-16 October, 1996, Killington, Vermont*. IEEE, ed. IEEE, New York, pp. 363–367.

Ruffman, T. and Keenan, T. R. (1996). “The Belief-Based Emotion of Suprise: The Case for a Lag in Understanding Relative to False Belief.” *Developmental Psychology*, (32)1:40.

Rumelhart, D. E. (1984). “The Emergence of Cognitive Phenomena from Sub-Symbolic Processes”. In: *Proceedings of the 6th Annual Meeting of the Cognitive Science Society, 28-30 June, 1984, Boulder, Colorado*. Vol. 6.

—, Hinton, G. E., and Williams, R. J. (1986a). “Learning Representations by Back-Propagating Errors”. *Nature*, (323)6088:533–536.

— McClelland, J. L., and Group, P. R., eds. (1986b). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition (2 Volumes)*. MIT Press, Cambridge (MA).

Russell, B. and Whitehead, A. N. (1913). *Principia Mathematica*. Cambridge University Press, Cambridge.

Russell, J. A. (1980). “A Circumplex Model of Affect”. *Journal of Personality and Social Psychology*, (39)6:1161–1178.

— (1991). “Culture and the Categorization of Emotions”. *Psychological Bulletin*, (110)3:426.

— (1994). “Is There Universal Recognition of Emotion from Facial Expression? A Review of the Cross-Cultural Studies”. *Psychological bulletin*, (115)1:102.

— (2003). “Core Affect and the Psychological Construction of Emotion”. *Psychological Review*, (110):145–172.

— and Barrett, L. F. (1999a). “Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant”. *Journal of Personality and Social Psychology*, (76)5.

— and Barrett, L. F. (1999b). “The Structure of Current Affect: Controversies and Emerging Consensus”. *Current Directions in Psychological Science*, (8):10–14.

— and Mehrabian, A. (1974a). *An Approach to Environmental Psychology*. MIT Press, Cambridge.

— and Mehrabian, A. (1974b). “Distinguishing Anger and Anxiety in Terms of Emotional Response Factors”. *Journal of Consulting and Clinical Psychology*, (42)1:79–83.

— and Mehrabian, A. (1977). “Evidence for a Three-Factor Theory of Emotions”. *Journal of Research in Personality*, (11)3:273–294.

Russell, S. and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson, London.

Sacco, G. (2024). “Recalcitrant Emotions: The Problems of Perceptual Theories”. *Ratio*, (37)2-3:156–167.

Sachs, E. and Brendler, S. J. (1948). “Some Effects of Stimulation of the Orbital Surface of the Frontal Lobe in the Dog and Monkey”. In: *Federation Proceedings*. Vol. 7, p. 107.

Sacks, O. (2008). *Musicophilia. Tales of Music and the Brain*. Knopf Doubleday Publishing Group.

Saffiotti, A., Ruspini, E. and Konolige, K. (1993). *A Fuzzy Controller for Flakey, An Autonomous Mobile Robot*. Technical Report 529, SRI Artificial Intelligence Center, Menlo Park (CA).

Samuel, A. L. (1959). “Some Studies in Machine Learning Using the Game of Checkers”. *IBM Journal of Research and Development*, (3)3:210–229.

Savery, R., ed. (2023). *Sound and Robotics: Speech, Non-Verbal Audio and Robotic Musicianship*. Chapman and Hall/CRC, New York.

—, Rogel, A., and Weinberg, G. (2021a). “Emotion Musical Prosody for Robotic Groups and Entitativity”. In: *RO-MAN 2021: Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication*. IEEE, ed. IEEE, New York, pp. 440–446.

—, Rose, R., and Weinberg, G. (2019a). “Establishing Human-Robot Trust Through Music-Driven Robotic Emotion Prosody and Gesture”. In: *RO-MAN 2019: Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication, 14-18 October, 2019, New Delhi*. IEEE, ed. IEEE, New York, pp. 1–7.

—, Rose, R., and Weinberg, G. (2019b). “Finding Shimi’s Voice: Fostering Human-Robot Communication with Music and a NVIDIA Jetson TX2”. In: *LAC 2019: Proceedings of the 17th Linux Audio Conference, 23-26 March, 2019, Stanford University*. Michon, R., ed. LAC, Stanford, pp. 101–105.

—, Zahray, L., and Weinberg, G. (2021b). “Emotional Musical Prosody for the Enhancement of Trust: Audio Design for Robotic Arm Communication”. *Paladyn, Journal of Behavioral Robotics*, (12)1:454–467.

Sawabe, T., Honda, S., Sato, W., et al. (2022). “Robot Touch with Speech Boosts Positive Emotions”. *Scientific Reports*, (12)1:6884.

Sawe, N., Chafe, C. and Treviño, J. (2020). “Using Data Sonification to Overcome Science Literacy, Numeracy, and Visualization Barriers in Science Communication”. *Frontiers in Communication*, (5).

Saygin, A. P., Chaminade, T., Ishiguro, H., et al. (2012). “The Thing That Should Not Be: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions”. *Social Cognitive and Affective Neuroscience*, (7)4:413–422.

Scarantino, A. and Sousa, R. de. (2021). “Emotion”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Summer 2021. Metaphysics Research Lab, Stanford University.

Scassellati, B. (1996). “Mechanisms of Shared Attention for a Humanoid Robot”. In: *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium, 9-11 November, 1996, Massachusetts Institute of Technology, Cambridge (MA)*. AAAI Press, Stanford, pp. 102–106.

— (1998). “Building Behaviors Developmentally: A New Formalism”. In: *Proceedings of the Integrating Robotics Research: Papers from the 1998 AAAI Spring Symposium, 23-25 March 1998, Palo Alto (CA)*. Yanco, H. A., ed. AAAI Press, Stanford, pp. 136–141.

Schaal, S. (2018). “Historical Perspective of Humanoid Robot Research in the Americas”. In: *Humanoid Robotics: A Reference*. Goswami, A. and Vadakkepat, P., eds. Springer, New York, pp. 9–17.

Schachter, S. and Singer, J. (1962). “Cognitive, Social, and Physiological Determinants of Emotional State.” *Psychological Review*, (69)5:379–399.

Schalkoff, R. (1992). *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, New York.

Schechtman, M. (2011). “Memory and Identity”. *Philosophical Studies*, (153)1:65–79.

Scheirer, J., Fernandez, R., Klein, J., et al. (2002). “Frustrating the User on Purpose: A Step Toward Building an Affective Computer”. *Interacting with Computers*, (14)2:93–118.

Scherer, K. R. (1984). “On the Nature and Function of Emotion: A Component Process Approach”. In: *Approaches to Emotion*. Ekman, P. and Scherer, K. R., eds. Lawrence Erlbaum, New York, pp. 293–317.

— (1999). “Appraisal Theory”. In: *Handbook of Cognition and Emotion*. Dalgleish, T. and Power, M., eds. John Wiley & Sons, Sussex.

—, Dieckmann, A., Unfried, M., et al. (2021). “Investigating Appraisal-Driven Facial Expression and Inference in Emotion Communication.” *Emotion*, (21)1:73.

Schiano, D. J., Ehrlich, S. M., Rahardja, K., et al. (2000). “Face to Interface: Facial Affect in (hu)man and Machine”. In: *CHI 2000: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Turner, T. and Szwillus, G., eds. ACM, New York, pp. 193–200.

Schildkraut, J. J. (1974). “Biogenic Amines and Affective Disorders”. *Annual Review of Medicine*, (25)1:333–348.

— and Kety, S. S. (1967). “Biogenic Amines and Emotion: Pharmacological Studies Suggest a Relationship Between Brain Biogenic Amines and Affective State.” *Science*, (156)3771:21–30.

Schindler, I., Hosoya, G., Menninghaus, W., et al. (2017a). “Measuring Aesthetic Emotions: A Review of the Literature and a New Assessment Tool”. *PloS one*, (12)6:e0178899.

Schindler, S., Zell, E., Botsch, M., et al. (2017b). “Differential Effects of Face-Realism and Emotion on Event-Related Brain Potentials and Their Implications for the Uncanny Valley Theory”. *Scientific Reports*, (7):45003.

Schlick, M. (1918). *Allgemeine Erkenntnislehre*. Julius Springer, Berlin.

Scholl, R. and Räz, T. (2016). “Towards a Methodology for Integrated History and Philosophy of Science”. In: *The Philosophy of Historical Case Studies*. Sauer, T. and Scholl, R., eds. Springer International Publishing, Cham, pp. 69–91.

Schopenhauer, A. (1819). *Die Welt als Wille und Vorstellung*. Johann Wilhelm Zinke, Leipzig.

Schultz, D. and Schultz, S. E. (2013). *A History of Modern Psychology*. Cengage Learning, Boston.

Schwind, V. (2015). “Historical, Cultural, and Aesthetic Aspects of the Uncanny Valley”. In: *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*. Misselhorn, C., ed. Springer International Publishing, Cham, pp. 81–107.

Schwitzgebel, E. (2024). “Belief”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. and Nodelman, U., eds. Spring 2024. Metaphysics Research Lab, Stanford University.

Scotus, J. D. (1968). *Opera Omnia*. Georg Olms, Hildesheim.

Searle, J. (1983). *Intentionality. An Essay in The Philosophy of Mind*. Cambridge University Press, Cambridge (MA).

Seashore, C. E. (1938). *Psychology of Music*. Mcgraw-Hill Book Company Inc., New York and London.

Selye, H. (1956). *The Stress of Life*. Mcgraw-Hill Book Company Inc., New York.

Seneca, L. A. (2015). *Letters on Ethics: To Lucilius*. Eng. tr. by Graver, M. and Long, A. A. University of Chicago Press, Chicago.

Sergi, G. (1894). *Dolore e piacere. Storia naturale dei sentimenti*. Fratelli Dumolard Editori, Milano.

Sertoli, E. (1883). “Contribution à la physiologie générale des muscles lisses”. *Archives italiennes de biologie*, (3):80–100.

Setiawan, S. A., Yamaguchi, J., Hyon, S. H., et al. (1999). “Physical Interaction between Human and a Bipedal Humanoid Robot-realization of Human-Follow Walking”. In: *ICRA 1999: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, 10-15 May, 1999, Detroit, MI, USA*. IEEE, ed. Vol. 1. IEEE, New York, pp. 361–367.

Shah, J., Vyas, A. and Vyas, D. (2014). “The History of Robotics in Surgical Specialties”. *American Journal of Robotic Surgery*, (1)1:12–20.

Shah, N. and Velleman, J. D. (2005). “Doxastic Deliberation”. *The Philosophical Review*, (114)4:497–534.

Shanahan, M. (2016). “The Frame Problem”. In: *The Stanford Encyclopedia of Philosophy*. Zalta, E. N., ed. Spring 2016. Metaphysics Research Lab, Stanford University.

— and Baars, B. (2005). “Applying Global Workspace Theory to the Frame Problem”. *Cognition*, (98)2:157–176.

Shannon, C. E. (1937). “A Symbolic Analysis of Relay and Switching Circuits”. *Electrical Engineering*, (57)12:713–723.

— (1948). “A Mathematical Theory of Communication”. *Bell System Technical Journal*, (27)3:379–423.

— (1951). “Presentation of a Maze-Solving Machine”. In: *Transaction of the Eighth Cybernetics Conference, 15-16 March 1951, New York*. Von Foerster, H., ed. Josiah Macy, Jr. Foundation, New York, pp. 173–180.

— (1952a). “Featured: Better Mouse: A Robot Rodent Masters Mazes”. *Life Magazine,* July 28, 1952, pp. 45–46.

— (1952b). “Featured: This Mouse is Smarter than You Are”. *Popular Science Monthly,* 1952, March. 160(3):99–101.

— (1953a). In: *The Potentialities of Computers*. Typescript, April 3, 1953, Bell Laboratories. In Sloane, N. J. A. and Wyner, Aaron D., editors (1993). *Claude Elwood Shannon. Collected Papers*. John Wiley & Sons, New York, pages 691-694.

— (1953b). “Computers and Automata”. *Proceedings of the IRE*, (41)10:1234–1241.

— and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana and Chicago.

Shao, M., Snyder, M., Nejat, G., et al. (2020). “User Affect Elicitation with a Socially Emotional Robot”. *Robotics*, (9)2:44.

Shapiro, L. (2019). *Embodied Cognition*. Routledge, London-New York.

— and Spaulding, S., eds. (2024). *The Routledge Handbook of Embodied Cognition*. Routledge, London-New York.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press, Oxford.

Sheridan, T. B. (2016). “Human-Robot Interaction: Status and Challenges”. *Human Factors*, (58)4:525–532.

— (2020). “A Review of Recent Research in Social Robotics”. *Current Opinion in psychology*, (36):7–12.

Sherrington, C. S. (1900). “Experiments on the Value of Vascular and Visceral Factors for the Genesis of Emotion”. *Proceedings of the Royal Society of London*, (66)424-433:390–403.

Shkurti, F. (2023). “The History and Future of Human-Robot Communication”. In: *The SAGE Handbook of Human-Machine Communication*. Guzman, A. L., Jones, S. and McEwen, R., eds. SAGE Publications, New York, pp. 54–62.

Shortliffe, E. H., Davis, R., Axline, S. G., et al. (1975). “Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System”. *Computers and Biomedical Research*, (8)4:303–320.

Shu, L., Xie, J., Yang, M., et al. (2018). “A Review of Emotion Recognition Using Physiological Signals”. *Sensors (Basel)*, (18)7:2074.

Siciliano, B. and Khatib, O., eds. (2016). *Springer Handbook of Robotics*. Springer, New York.

Simon, H. A. (1967). “Motivational and Emotional Controls of Cognition”. *Psychological Review*, (74)1:29.

Singer, T., Seymour, B., O’Doherty, J., et al. (2004). “Empathy for Pain Involves the Affective but Not Sensory Components of Pain”. *Science*, (303)5661:1157–1162.

Singh, P. (2005). *EM-ONE: An Architecture for Reflective Commonsense Thinking*. PhD thesis, Massachusetts Institute of Technology, Cambridge (MA).

Skinner, B. F. (1953). *Science and Human Behavior*. 92904. Macmillan and Co., New York.

Sloboda, J. A. (1991). “Music Structure and Emotional Response: Some Empirical Findings”. *Psychology of Music*, (19)2:110–120.

Sloman, A. (2001). “Varieties of Affect and the CogAff Architecture Schema”. In: *AISB 2001: Proceedings of the Symposium on Emotion, Cognition and Affective Computing, 10-12 April 2001, York (UK)*. Johnson, C. G., ed. AISB Press, University of York, York (UK), pp. 39–48.

— and Croucher, M. (1981). “Why Robots Will Have Emotions”. In: *IJCAI 1981: Proceedings of the 7th International Conference on AI, 24-28 August, 1981, University of British Columbia*. Hayes, P. J., ed. Vol. 176. IJCAI, Vancouver, pp. 197–202.

Smidt, K. E. and Suvak, M. K. (2015). “A Brief, but Nuanced, Review of Emotional Granularity and Emotion Differentiation Research”. *Current Opinion in Psychology*, (3):48–51.

Smith, L. B. and Samuelson, L. K. (2003). “Different Is Good: Connectionism and Dynamic Systems Theory Are Complementary Emergentist Approaches to Development”. *Developmental Science*, (6)4:434–439.

— and Thelen, E. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Press, Cambridge (MA).

— and Thelen, E. (2003). "Development as a Dynamic System". *Trends in Cognitive Sciences*, (7)8:343–348.

Smith, M. B. (1956). *Opinions and Personality*. John Wiley & Sons, New York.

Smith, W. K. (1944). "The Results of Stimulation of the Uncus and Adjacent Portions of the Hippocampal Gyrus". In: *Federation Proceedings*. Vol. 3, p. 43.

— (1945). "The Functional Significance of the Rostral Cingular Cortex as Revealed by Its Responses to Electrical Excitation". *Journal of Neurophysiology*, (8)4:241–255.

Smolensky, P. (1988). "On the Proper Treatment of Connectionism". *Behavioral and Brain Sciences*, (11)1:1–23.

Söderlund, M. (2022). "Service Robots with (Perceived) Theory of Mind: An Examination of Humans' Reactions". *Journal of Retailing and Consumer Services*, (67).

Solis, J., Chida, K., Suefuji, K., et al. (2006). "The Development of the Anthropomorphic Flutist Robot at Waseda University". *International Journal of Humanoid Robotics*, (03)02:127–151.

—, Takanishi, A., and Hashimoto, K. (2010). "Development of an Anthropomorphic Saxophone-Playing Robot". In: *Brain, Body and Machine*. Angeles, J., Boulet, B., Clark, J. J., et al., eds. Springer, Berlin, Heidelberg, pp. 175–186.

—, Taniguchi, K., Ninomiya, T., et al. (2009). "Refining the Flute Sound Production of the Waseda Flutist Robot the Mechanical Design of the Artificial Organs Involved During the Flute Playing". *Mechanism and Machine Theory*, (44)3:527–540.

Solomon, R. C. (1976). *The Passions: Emotions and the Meaning of Life*. Hackett Publishing, Indianapolis.

— (2003). *What is an Emotion?: Classic and Contemporary Readings*. Oxford University Press, Oxford.

Somenzi, V. and Cordeschi, R., eds. (1994). *La filosofia degli automi. Origini dell'intelligenza artificiale*. Bollati Boringhieri, Torino.

Spiegel, E. A., Miller, H. R. and Oppenheimer, M. J. (1940). "Forebrain and Rage Reactions". *Journal of Neurophysiology*, (3)6:538–548.

Spinoza, B. (1677). *Ethica more geometrico demonstrata*. Jan Rieuwertsz, Amsterdam.

Spitale, M., Axelsson, M., Jeong, S., et al. (2025). "Past, Present, and Future: A Survey of The Evolution of Affective Robotics For Well-being". *IEEE Transactions on Affective Computing*: 1–17.

— and Gunes, H. (2022). "Affective Robotics For Wellbeing: A Scoping Review". In: *ACIIW 2022: Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos 17-21 October 2022, Nara Japan*. Baird, A., Tzirakis, P., Brooks, J. L., et al., eds. IEEE, New York, pp. 1–8.

Spitzer, M. (2020). *A History Of Emotion In Western Music: A Thousand Years From Chant To Pop*. Oxford University Press, Oxford.

St George, J., Kim, S. Y. and Bischof, H.-P. (2018). "Sonification of Simulated Black Hole Merger Data". In: *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*. The Steering Committee of The World Congress in Computer Science, pp. 3–9.

Staffa, M., D'Errico, L., Sansalone, S., et al. (2023). "Classifying Human Emotions in HRI: Applying Global Optimization Model to EEG Brain Signals". *Frontiers in Neurorobotics*, (17):1191127.

Staller, A. and Petta, P. (2001). "Introducing Emotions into the Computational Study of Social Norms: A First Evaluation". *Journal of Artificial Societies and Social Simulation*, (4)1:U27–U60.

Stearns, P. N. and Matt, S. J. M. (2014). *Doing Emotions History*. University of Illinois Press, Champaign.

— and Stearns, C. Z. (1985). "Emotionology: Clarifying the History of Emotions and Emotional Standards". *The American Historical Review*, (90)4:813–836.

Steels, L. and Brooks, R. A., eds. (1995). *The Artificial Life Route to Artificial Intelligence Building Embodied, Situated Agents*. Routledge, New York.

Stewart, C. C. (1900). “Mammalian Smooth Muscle–The Cat’s Bladder”. *American Journal of Physiology*, (4):179–200.

Stich, S. P. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. MIT Press, Cambridge (MA).

Stock-Homburg, R. (2022). “Survey of Emotions in Human-Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research”. *International Journal of Social Robotics*, (14)2:389–411.

Stone, W. L. (2005). “The History of Robotics”. In: *Robotics and Automation Handbook*. Kurfess, T. R., ed. Vol. 414. CRC Press, Boca Raton, FL.

Stroop, J. R. (1935). “Studies of Interference in Serial Verbal Reactions”. *Journal of Experimental Psychology*, (18)6:643.

Stumpf, C. (1883). *Tonpsychologie*. Hirzel, Leipzig.

Sturm, T. (2011). “Historical Epistemology or History of Epistemology? The Case of the Relation Between Perception and Judgment”. *Erkenntnis*, (75)3:303–324.

Sugano, S. and Kato, I. (1987). “Wabot-2: Autonomous Robot with Dexterous Finger-Arm–Finger-Arm Coordination Control and Keyboard Performance”. In: *ICRA 1987: Proceedings of the 1987 IEEE International Conference on Robotics and Automation, 31 March-3 April, 1987, Raleigh NC (USA)*. IEEE, ed. Vol. 4. IEEE, New York, pp. 90–97.

Susino, M. and Schubert, E. (2019). “Cultural Stereotyping of Emotional Responses to Music Genre”. *Psychology of Music*, (47)3:342–357.

Sutherland, A. (1898). *The Origin and Growth of the Moral Instinct*. Longmans, Green and Co., New York-Bombay.

Suzuki, Y., Galli, L., Ikeda, A., et al. (2015). “Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography”. *Scientific Reports*, (5)1:15924.

Swagerman, J. (1987). “The Artificial Concern Realization System (ACRES): A Computer Model of Emotion”. PhD thesis. University of Amsterdam, Psychology Department.

Takahashi, H., Matsuura, M., Koeda, M., et al. (2008). “Brain Activations During Judgments of Positive Self-Conscious Emotion and Positive Basic Emotion: Pride and Joy”. *Cerebral Cortex*, (18)4:898–903.

Takanishi, A. (1991). “Development of Control Method for Biped Walking under Unknown External Force Acting in Lateral Plane”. In: *RSJ 1991: Proceedings of the Annual Conference of the Robotics Society of Japan, Tsukuba, Japan, 1991*, pp. 321–324.

Takanishi, A., Ishida, M., Yamazaki, Y., et al. (1985). “The Realization of Dynamic Walking by the Biped Walking Robot WL-10 RD”. *Journal of the Robotics Society of Japan*, (3)4:325–336.

Tamborini, M. (2021). “The Material Turn in the Study of Form: From Bio-Inspired Robots to Robotics-Inspired Morphology”. *Perspectives on Science*, (29)5:643–665.

— (2024a). *Biorobotik zur Einführung*. Junius Verlag, Hamburg.

—, ed. (2024b). *Die Philosophie der Bio-Robotik*. Meiner, Leipzig, pp. 1–10.

— and Datteri, E. (2022). “Is Biorobotics Science? Some Theoretical Reflections”. *Bioinspiration & Biomimetics*, (18)015005.

Tangney, J. P., Stuewig, J. and Mashek, D. J. (2007). “Moral Emotions and Moral Behavior”. *Annual review of psychology*, (58):345–372.

Tappolet, C. (2000). *Émotions et valeurs*. Presses Universitaires de France, Paris.

— (2016). *Emotions, Value, and Agency*. Oxford University Press, Oxford.

— (2020). “Emotions Inside Out: The Nonconceptual Content of Emotions”. In: *Concepts in Thought, Action, and Emotion*. Schröder, D. and Demmerling, C., eds. Routledge, New York, pp. 257–276.

Teroni, F. (2007). “Emotions and Formal Objects”. *Dialectica*, (61)3:395–415.

Terry, J. and Calvert, M. (1997). *Processed Lives: Gender and Technology in Everyday Life*. Routledge, London-New York.

Thagard, P. and Kunda, Z. (1987). “Hot Cognition Mechanisms for Motivated Inference”. In: *Proceedings of the 9th Annual Meeting of the Cognitive Science Society, 16-18 July, 1987, Seattle*. Vol. 9.

Thellman, S., Graaf, M. de and Ziemke, T. (2022). "Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings". *Journal of Human-Robot Interaction*, (11)4.

Thepsoonthorn, C., Ogawa, K.-i. and Miyake, Y. (2021). "The Exploration of the Uncanny Valley from the Viewpoint of the Robot's Nonverbal Behaviour". *International Journal of Social Robotics*, (13)6:1443–1455.

Therrien, C. W. (1989). *Decision, Estimation and Classification*. John Wiley & Sons, New York.

Thoits, P. A. (1989). "The Sociology of Emotions". *Annual Review of Sociology*, (15):317–342.

Thormählen, W. (2020). "Music and Dance". In: *A Cultural History Of The Emotions In The Age Of Romanticism, Revolution, And Empire*. Matt, S. J., ed. Bloomsbury Publishing, London, pp. 55–74.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Press, New York.

Tinwell, A. (2014). *The Uncanny Valley in Games and Animation*. A K Peters/CRC Press, New York.

Tobis, S., Piasek-Skupna, J. and Suwalska, A. (2023). "The Godspeed Questionnaire Series in the Assessment of the Social Robot TIAGo by Older Individuals". *Sensors*, (23)16:7251.

Toda, M. (1962). "Design of a Fungus-Eater". *Behavioral Science*, (7):164–183.

— (1982a). "Emotional Fungus-Eaters". In: *Man, Robot and Society: Models and Speculations*. Nijhoff, Boston, pp. 130–153.

— (1982b). *Man, Robot and Society: Models and Speculations*. Nijhoff, Boston.

Tomkins, S. S. (1962). *Affect, Imagery, and Consciousness*. Vol. 1. Springer, New York.

Torricelli, D., Mizanoor, R. S. M., Lippi, V., et al. (2019). "Benchmarking Human Likeness of Bipedal Robot Locomotion: State of the Art and Future Trends". In: *Metrics of Sensory Motor Coordination and Integration in Robots and Animals*. Bonsignorio, F., Messina, E., Pobil, A. P. del, et al., eds. Vol. 36. Cognitive Systems Monographs. Springer, Berlin, pp. 147–166.

—, Mizanoor, R. S. M., Lippi, V., et al. (2020). "Benchmarking Human Likeness of Bipedal Robot Locomotion: State of the Art and Future Trends". In: *BBS 2014: Proceeding of the Third International Conference on Biomimetic and Biohybrid Systems, Living Machine, 30 July - 1 August, 2014, Milan, Italy*. Bonsignorio, F., Messina, E., Del Pobil, A. P., et al., eds. Vol. 36. Springer, Berlin-New York, pp. 147–166.

Toyota, M. C. (2004). *Toyota Announces Overview of "Toyota Partner Robot" Robots to Be Unveiled at Expo 2005, Starting March 2005*. https://global.toyota/en/detail/221580?.

— (2007). *Toyota Unveils Personal-transport, Violin-playing Robots*. https://global.toyota/en/detail/275807?.

Toyota, Motor Corporation. (2007). *Toyota Group Announces Plans for EXPO 2005 AICHI, JAPAN Pavilion Performance*. https://global.toyota/en/detail/245497?.

Trevarthen, C. (1979). "Communication and Cooperation in Early Infancy: A Description of Primary Intersubjectivity". *Before Speech: The Beginning of Interpersonal Communication*, (1):530–571.

Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Journal of Mathematics*, (58)345-363:5.

— (1950). "Computing Machinery and Intelligence". *Mind*, (59)236:433.

Turkle, S. (1984). *The Second Self. Computers and the Human Spirit*. Simon & Schuster, Inc., New York.

— (1995). *Life on the Screen. Identity in the Age of The Internet*. Simon & Schuster, Inc., New York.

Tye, M. (2008). "The Experience of Emotion: An Intentionalist Theory". *Revue internationale de philosophie*, (243)1:25–50.

Università di Bari. (2022). *Rector Decree n. 2280 (Call for Admission to the PhD Program – XXXVIII Cycle)*. July 2022. https://www.uniba.it/dottorati/38-ciclo/call-for-phd-selection-xxxviii-cycle.

Urgen, B. A., Kutas, M. and Saygin, A. P. (2018). "Uncanny Valley as a Window into Predictive Processing in the Social Brain". *Neuropsychologia*, (114):181–185.

Vagelli, M. (2024). *Reconsidering Historical Epistemology: French and Anglophone Styles in History and Philosophy of Science*. Studies in History and Philosophy of Science. Springer International Publishing, Cham.

Vagnetti, R., Camp, N., Story, M., et al. (2024). "Instruments for Measuring Psychological Dimensions in Human-Robot Interaction: Systematic Review of Psychometric Properties". *Journal of Medical Internet Research*, (26):e55597.

Van Gelder, T. (1995). "What Might Cognition Be, If Not Computation?" *Journal of Philosophy*, (92)7:345–381. Smylie, J., ed.

— (1998). "The Dynamical Hypothesis in Cognitive Science". *Behavioral and Brain Sciences*, (21)5:615–628.

Varela, F. J., Thompson, E. and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge (MA).

Vaucanson, J. d. (1738). *Le Mécanisme du flûteur automate*. Eng. tr. by Desaguliers, J.T. (1742). An Account of the Mechanism of an Automaton, or Image Playing on the German-Flute. T. Parker, London. Jacques Guerin, Paris.

Velásquez, J. D. (1996). *Cathexis: A Computational Model for the Generation of Emotions and Their Influence in the Behavior of Autonomous Agents*. PhD thesis, Massachusetts Institute of Technology, Cambridge (MA).

— (1997). "Modeling Emotions and Other Motivations in Synthetic Agents". In: *AAAI 1997: Proceedings of the 14th National Conference on Artificial Intelligence. 27-31 July 1997, Providence*. MIT Press, Cambridge (MA), pp. 10–15.

— and Maes, P. (1997). "Cathexis: A Computational Model of Emotions". In: *AGENTS 1997: Proceedings of the First International Conference on Autonomous Agents*. Johnson, W. L., ed. Association for Computing Machinery, New York, pp. 518–519.

Vinci, L. da. (2008). *Codex Atlanticus*. Facsimile: Pedretti, Carlo (editor). Giunti Editore, Milano.

Von Zitzewitz, J., Boesch, P. M., Wolf, P., et al. (2013). "Quantifying the Human Likeness of a Humanoid Robot". *International Journal of Social Robotics*, (5)2:263–276.

Vonnegut, K. (1952). *Player Piano*. Charles Scribner's Sons, New York (NY).

Voskuhl, A. (2019). *Androids in the Enlightenment: Mechanics, Artisans, and Cultures of the Self*. University of Chicago Press, Chicago.

— and Riskin, J. (2010). "Motions and Passions: Music-Playing Women Automata and the Culture of Affect in Late Eighteenth-Century Germany". In: *Genesis redux: Essays in the History and Philosophy of Artificial Life*. Riskin, J., ed. University of Chicago Press, Chicago.

Vygotsky, L. S. and Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, Cambridge (MA).

Vyzas, E. and Picard, R. W. (1998). "Affective Pattern Classification". *AAAI Fall Symposium Series: Emotional and Intelligent: The Tangled Knot of Cognition, October 23-25, Orlando, Florida. TR 473.*, (176182).

— and Picard, R. W. (1999). "Offine and Online Recognition of Emotion Expression from Physiological Data". In: *AGENTS 1999: Proceedings of the Third International Conference on Autonomous Agents, Workshop on Emotion-Based Archetectures, 1-5 May, 1999, Seattle, WA*. Etzioni, O., Müller, J. P. and Bradshaw, J. M., eds. ACM, New York.

Wallace, J. G. (1983). "Motives And Emotions in a General Learning System". In: *IJCAI 1983: Proceedings of 8th International Joint Conference on Artificial Intelligence 8-12 August, 1983, Karlsruhe, West Germany*. Publishers, M. K., ed., San Francisco, pp. 84–86.

Wallén, J. (2008). *The History of the Industrial Robot*. Linköping University Electronic Press, Linköping.

Walter, W. G. (1952). *Featured: Cora and the Long Low Whistle. The Sunday Herald*. 1951, January 28, page 11.

Walter, W. G. (1950a). "An Electro-Mechanical "Animal"". *Dialectica*, (4)3-4:206–213.

— (1950b). "An Imitation of Life". *Scientific American*, (182)5:42–45.

— (1951). "A Machine That Learns". *Scientific American*, (185)8:60–63.

— (1953). *The Living Brain*. W. W. Norton & Company, New York.

Wang, S., Lilienfeld, S. O. and Rochat, P. (2015). "The Uncanny Valley: Existence And Explanations". *Review of General Psychology*, (19)4:393–407.

Wang, Y., Song, W., Tao, W., et al. (2022). "A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances". *Information Fusion*, (83):19–52.

Ward, A. A. (1948). "The Cingular Gyrus: Area 24". *Journal of Neurophysiology*, (11)1:13–23.

Warner, L. R. and Shields, S. A. (2009). "Judgements of Others' Emotional Appropriateness Are Multidimensional". *Cognition & Emotion*, (23)5:876.

Wason, P. C. (1966). "Reasoning". In: *New Horizons in Psychology*. Foss, B. M., ed. Penguin, Harmondsworth, pp. 131–151.

Watson, J. B. (1914). *Behavior: An Introduction to Comparative Psychology*. Henry Holt and Co., New York.

Weber, E. H. (1834). *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*. Koehler, Leipzig.

— (1851). *Die Lehre vom Tastsinne und Gemeingefühle auf Versuche gegründet*. Friedrich Vieweg und Sohn, Braunschweig.

Wegman, C. (1977). "A Computer Simulation of Freud's Counterwill Theory". *Behavioral Science*, (22)3:218–233.

Weinberg, G., Bretan, M., Hoffman, G., et al. (2020). "Platforms—Georgia Tech's Robotic Musicians". In: *Robotic Musicianship: Embodied Artificial Creativity and Mechatronic Musical Expression*. Weinberg, G., Bretan, M., Hoffman, G., et al., eds. Springer International Publishing, Cham, pp. 25–61.

— and Driscoll, S. (2006a). "Robot-Human Interaction with an Anthropomorphic Percussionist". In: *SIGCHI 2006: Proceedings of the Conference on Human Factors in Computing Systems, 22-27 April, 2006, Montréal, Canada*. Grinter, R., Rodden, T., Aoki, P., et al., eds. ACM, Montréal, pp. 1229–1232.

— and Driscoll, S. (2006b). "Toward Robotic Musicianship". *Computer Music Journal*, (30)4:28–45.

—, Driscoll, S., and Parry, M. (2005). "Musical Interactions with a Perceptual Robotic Percussionist". In: *RO-MAN 2005: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, 13-15 August, 2005, Nashville, TN, USA*. IEEE, ed. IEEE, New York, pp. 456–461.

—, Raman, A., and Mallikarjuna, T. (2009). "Interactive Jamming with Shimon: A Social Robotic Musician". In: *HRI 2009: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, 9-13 March, 2009, La Jolla (CA)*. Scheutz, M. and Michaud, F., eds. ACM, New York, pp. 233–234.

Weiner, B. (1980). "A Cognitive (attribution)-Emotion-Action Model of Motivated Behavior: An Analysis of Judgments of Help-Giving". *Journal of Personality and Social Psychology*, (39)2:186–200.

— (1985). "An Attributional Theory of Achievement Motivation and Emotion". *Psychological Review*, (92)4:548–573.

Weiskrantz, L. (1990). *Blindsight: A Case Study and Implications*. Oxford University Press, Oxford.

Weiss, A. and Bartneck, C. (2015). "Meta Analysis of the Usage of the Godspeed Questionnaire Series". In: *RO-MAN 2015: Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication, 31 August-1 September, 2015, Kobe, Japan*. IEEE, ed. IEEE, New York, pp. 381–388.

Weizenbaum, J. (1966). "Eliza—a Computer Program for the Study of Natural Language Communication Between Man and Machine". *Communications of the ACM*, (9)1:36–45.

Wells, F. L. (1925). "Reactions to Visual Stimuli in Affective Settings". *Journal of Experimental Psychology*, (8):50–80.

— and Forbes, A. (1911). "On Certain Electrical Processes in the Human Body and Their Relations to Emotional Reactions". *Archives of Psychology*, (2)16:1–26.

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Committee on Applied Mathematics. Harvard University, Cambridge (MA).

— (1982). "Applications of Advances in Nonlinear Sensitivity Analysis". In: *IFIP 1982: Proceedings of the 10th IFIP Conference, System Modeling and Optimization, 31 August-4 September, New York*. Drenick, R. and Kozin, F., eds. Springer, New York, pp. 762–777.

— (1994). “The Brain as a Neurocontroller. New Hypotheses and New Experimental Possibilities”. In: *Origins: Brain and Self Organization*. Pribram, L., ed. Erlbaum, Mahwah (NJ), pp. 680–706.

Wertheimer, M. (1961). “Psychomotor Coordination of Auditory and Visual Space at Birth”. *Science*, (134):1692.

Westermann, B. (2016). “The Biomorphic Automata of the 18th Century. Mechanical Artworks as Objects of Technical Fascination and Epistemological Exhibition”. *Figurationen*, (17)2:123–137.

Wheeler, M. (2008). “Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem”. *International journal of philosophical studies*, (16)3:323–349.

Whiting, D. (2011). “The Feeling Theory of Emotion and the Object-Directed Emotions”. *European Journal of Philosophy*, (19)2:281–303.

Widrow, B. and Hoff, M. E. (1960). “Adaptive Switching Circuits”. In: *IRE 1960: Proceedings of the 7th Region Conference, Wisconsin Convention Report, 23-26 August, 1960,* CESTA, ed. Vol. 4B, Paris, pp. 96–104.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. Hermann & Cie, Paris.

— (1950). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin, Boston.

Wierzbicka, A. (1999). *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge university press, Cambridge.

Williamson, M. M. (1998a). “Exploiting Natural Dynamics in Robot Control”. In: *EMCSR 1998: Proceedings of the Fourteenth European Meeting on Cybernetics and Systems Research, 14-17 April, 1998, University of Vienna*. Trappl, R., ed. Austrian Society for Cybernetic Studies, Vienna, pp. 160–170.

— (1998b). “Rhythmic Robot Control Using Oscillators”. In: *IEEE/RSJ 1998: Proceedings of the International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications, 13-17 October, 1998, Victoria (CA)*. Wong, A. K. C., ed. IEEE, Piscataway, pp. 77–83.

Wilson, M. and Knoblich, G. (2005). “The Case for Motor Involvement in Perceiving Conspecifics”. *Psychological Bulletin*, (131)3:460–473.

Wimmer, H. and Perner, J. (1983). “Beliefs About Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception”. *Cognition*, (13)1:103–128.

Winters Fisher, A. (2005). “Perceptions of Body Posture and Emotion: A Question of Methodology”. *The New School Psychology Bulletin*, (3).

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Kegan Paul, Trench, Trubner & Co., London.

Wood, D., Bruner, J. S. and Ross, G. (1976). “The Role of Tutoring in Problem-Solving”. *Journal of Child Psychology and Psychiatry*, (17):89–100.

Woodcroft, B. (1851). “The Pneumatics of Hero of Alexandria”. In: Taylor Walton and Maberly, London.

Wooldridge, M. (2022). *Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*. Flatiron Books, New York.

Wright, T. (1604). *The Passions of the Minde in Generall*. Miles Flesher, London.

Wu, C.-F. J. (1983). “On the Convergence Properties of the EM Algorithm”. *The Annals of Statistics*, (11)1:95–103.

Wu, Y. and Schulz, L. E. (2018). “Inferring Beliefs and Desires From Emotional Reactions to Anticipated and Observed Events”. *Child Development*, (89)2:649–662.

Wundt, W. (1874). *Grundzüge der physiologischen Psychologie*. Wilhelm Engelmann, Leipzig.

— (1896). *Grundriss der Psychologie*. Wilhelm Engelmann, Leipzig.

Yacoob, Y. and Davis, L. S. (1996). “Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (18)6:636–642.

Yan, Z., Song, Y., Zhou, R., et al. (2024). “Facial Expression Realization of Humanoid Robot Head and Strain-Based Anthropomorphic Evaluation of Robot Facial Expressions”. *Biomimetics*, (9)3:122.

Yates, D. (2008). "Response-Dependence". *Philosophical Books*, (49)4:344–354.

Yohanan, S. and MacLean, K. E. (2012). "The Role of Affective Touch in Human-Robot Interaction: Human Intent and Expectations in Touching the Haptic Creature". *International Journal of Social Robotics*, (4)2:163–180.

Young, T. (1807). *Lectures on Natural Philosophy and the Mechanical Arts*. Joseph Johnson, London.

Zagzebski, L. (2003). "Emotion and Moral Judgment". *Philosophy and Phenomenological Research*, (66)1:104–124.

Zanella, A., Harrison, C. M., Lenzi, S., et al. (2022). "Sonification and Sound Design for Astronomy Research, Education and Public Engagement". *Nature Astronomy*, (6)11:1241–1248.

Zanzarella, I. (2024). "Outlining a Protohistory of Artificial Intelligence and Music: From Antiquity to Nineteenth Century". *Rivista di Studi Politici*, (1):99–141.

— (2025a). *Familiarity, Anthropomorphism and Emotion. Toward a Third Dimension of the Uncanny Valley Model from a Historical-Scientific Perspective*. Workshop "Antropomorfismo. L'umano nell'interazione uomo-robot", 10 april, 2025, Università di Milano-Bicocca, Milan (Italy).

— (2025b). "The Problem of Musical Creativity and its Relevance for Ethical and Legal Decisions towards Musical AI". *Filosofia Morale/Moral Philosophy*, (2):101–113.

Zhang, J., Yin, Z., Chen, P., et al. (2020a). "Emotion Recognition Using Multi-Modal Data and Machine Learning Techniques: A Tutorial and Review". *Information Fusion*, (59):103–126.

Zhang, J., Li, S., Zhang, J.-Y., et al. (2020b). "A Literature Review of the Research on the Uncanny Valley". In: *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments: Proceedings of the CCD 2020 12th International Conference on Cross-Cultural Design, held as Part of the HCII 2020 22nd International Conference on Human-Robot Interaction, 19-24 July, 2020, Copenhagen*. Rau, P.-L. P., ed. Springer, New York, pp. 255–268.