Developing Empathy in Social Robots

by

Molly Graham

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements
for the degree of

Doctor of Philosophy

in

Philosophy

Guelph, Ontario, Canada

Abstract

Developing Empathy in Social Robots

Molly Graham — Advisor:

University of Guelph, 2025 — Don Dedrick

With increasing development and widespread interest in artificial intelligence and robotics over the last several years, novel solutions are being considered for addressing specific problems in a number of domains. In particular, caring for elderly individuals in regions where human labour is insufficient is of special interest. One aspect of this involves an ability to communicate and socialize with people in a familiar and humanlike manner, with an ability to understand human emotion through behavioural cues and language. Furthermore, social robots should be able to act with empathy if we are interested in using them for tasks related to eldercare and healthcare, as empathy involves the adoption of another person's perspective. This work investigates whether social robots are capable of meeting these requirements, and ultimately demonstrates that they cannot exhibit genuine acts of empathy. By investigating one robot in particular named iCub, it becomes apparent that it will not be able to understand human emotions given its cognitive architecture. iCub's framework lacks the kind of affective regulation observed in people and animals, and as such, is unable to experience its own form of emotions. Consequently, iCub is incapable of understanding what emotions are and what they mean to humans. For social robots, an alternative solution which relies on experiences of emotions should instead be pursued. The cognitive architecture developed by Pentti Haikonen is preferable because it is modelled on biological functionality, creating better analogues of living beings than previous efforts in AI. By creating associative networks of neurons which establish connections between stimuli and their effects on the robot's body, it is able to learn what a stimulus means for its continued

functioning. By using signals representing analogues of pain and pleasure, emotional information learned by the robot is grounded on how stimuli impact the robot's body. This architecture provides a foundation for empathy since the robot is able to appeal to its own experiences of pain and pleasure to understand the meanings of behavioural cues exhibited by humans.

Dedication

To Mauricio, for without your inspiration, support, and assistance, this work would not exist.

Acknowledgements

First, I would like to express deep gratitude for my supervisor Dr. Don Dedrick for all of his guidance and assistance as I completed both my Master's and PhD at the University of Guelph. I am very grateful for all the time and energy Don has provided in reading, reviewing, and editing my thesis. The feedback he provided for me over the years has been meaningful and illuminating, and I am deeply appreciative of all of his efforts and input on this project.

I would also like to extend many thanks to Dr. Pentti Haikonon for serving as the external member on my supervisory committee, and for his feedback on all chapters of this work. His perspective and the comments he provided have been very helpful, and I am grateful for the opportunity to have received his advice on how I have presented his contributions to robotics.

For his supervisory role and input, I would also like to thank Dr. Andrew Bailey. His philosophical background provided a unique perspective from which to consider this work, and I am thankful for the comments and ideas he contributed over the years.

I would also like to take the opportunity to thank Dr. Nancy Salay for serving as the external examiner for my defense. Her feedback on my thesis provided significant contributions for which I am very grateful.

Many thanks must also be extended to the staff and faculty in the Department of Philosophy at the University of Guelph. I am particularly thankful for the courses taught by Dr. John Russon, Dr. Stefan Linquist, Dr. John Hacker-Wright, Dr. Mark McCullagh, Dr. Gus Skorburg, and Dr. Omid Payrow Shabani. Not only were these courses enlightening and

enjoyable, but they introduced me to important materials which meaningfully contributed to this work.

I am also very grateful for all of my friends and peers I met at both the University of Guelph and the University of Toronto. Their companionship and support has been extremely valuable, and I wish them all the best as they continue on in their studies, journeys, and endeavours.

Another individual I would like to thank is John Kanwischer, my manager at Western Union Business Solutions where I worked as a junior programmer in 2012. During my time there, I joined the office book club where I was introduced to *Gödel, Escher, Bach* by Douglas Hofstadter. The book itself had a profound impact on my appreciation and understanding of logic and computer science, but it was during one of these meetings when one of my colleagues turned to me and expressed the idea about consciousness arising from recursive functions. This comment resonated with me in a way which few ideas do, and from that moment I understood that consciousness was the one topic I felt a great affinity towards. Years later, while pondering about what I would devote myself to, I thought back to that pivotal moment and knew there was nothing else I was more interested in studying. Without the efforts of John Kanwischer, that significant moment of inspiration would have never occurred.

The completion of this project could not have been possible without the love and support of my family, friends, and all of the teachers I have had along the way. In particular, I would like to thank my parents, my brother Rory, my childhood friend Dr. Corey DeGagne, and my partner Mauricio. Additionally, I would also like to thank Dr. Jim John and Dr. Brian Cantwell Smith at

the University of Toronto for their courses and letters of recommendation for my admittance to the University of Guelph. Many warm thanks must also be extended to Tannis Stallard, Laura Lamb, Norm Joss, and Holly Laing. These teachers were pivotal in instilling within me a love of learning and the arts.

Table of Contents

1 Introduction

Artificial intelligence (AI) has seen an explosion of interest and development over the last several years as a result of widespread internet use and improved computer hardware. Even before the successes of deep learning, optimism about the future of AI and robotics suggested a future where technology rivals human abilities. Media and popular culture throughout the 20th century has showcased future societies both helped and hindered by these technologies, depicting robots in a range of forms from helpful household servants to nefarious entities. Determining whether these fictitious accounts are worthy of concern, however, has been a source of debate since their inception, with many believing that AI can assist in improving the lives of individuals and communities around the world. Optimists view robots and AIs as a means for addressing specific problems, such as caring for the elderly in regions where human labour is insufficient for meeting the needs of its populations. In this case, a specific kind of AI is required, one with an ability to socialize with people in a humanlike way. Moreover, some instances of sociable AIs will require a body to act in a familiar, humanlike manner, learning as children do through experimentation and interacting with their environment. The idea is that if these robots can learn about the world, the meanings of the words and phrases used by humans can be determined through association. For example, learning the spoken word for an object, in conjunction with the sight of the object, can create an association between the two. If successful, learning can be built up over time as observed in childhood, and robots will become able to fulfill the roles once occupied by people.

This work specifically investigates social robots to determine whether the capacities required to communicate and develop social skills are indeed feasible. One such skill is an ability to understand human emotion through behavioural cues and in language, acting appropriately given the situation and person(s) at hand. Furthermore, social robots should be able to act empathically, especially if we are interested in using them for tasks related to eldercare and healthcare. The reason for implementing them in these domains is due to an increasing shortage in human labour to meet the growing demand for care, given recent demographic trends observable worldwide. If we are to address these issues with technological solutions, however, robots must be able to interact with people in a familiar manner. After introducing the issues motivating these novel solutions, the first chapter of this work examines the components of human communication to identify the requirements for social robots. Although empathy is a significant component of healthcare, this capacity appeals to a more fundamental human ability, that of perspective-taking. To know or sense what another person is thinking and feeling is a crucial component of communication, and as such, social robots will require an ability to understand the behavioural cues of human individuals.

For this understanding to occur, the *symbol grounding problem* must be addressed. This problem seeks to uncover how words and concepts become associated with meanings. Because computers only perform rule-based operations, meaning cannot be incorporated into their functionality. While computers and AIs may appear to understand words and concepts, these appearances are illusory. Rather than understanding what certain ideas or terms refer to, they merely track connections established between these ideas which exist as syntactical structures. To demonstrate this point, this work specifically investigates terms related to affect and emotion,

as computerized robots are incapable of subjective experience. Given their inability to feel and understand, our current approaches to AI are incapable of expressing empathy.

To better understand what empathy and perspective-taking involves, the second chapter investigates the philosophical background and scientific basis of this ability. Originating from the domain of aesthetics and phenomenology, the term *empathy* is often associated with simulating within oneself the perceived feelings of others, as detected by behavioural cues. To facilitate this, one's own knowledge of others can be appealed to in cases when simulation is insufficient, suggesting that empathy requires both an appeal to knowledge or theory in addition to acts of internal simulation. Overall, however, these two aspects of empathy are best characterized by a particular interaction unfolding between two or more individuals, where situational context plays an important role in whether an act of empathy is successful or not. Given that individuals are unique beings with specific sociocultural backgrounds, it is to be expected that not all acts of empathy will be successful, and any theory of empathy must account for how and why they succeed or fail. As individuals communicate and interact, their attempts to understand one another are influenced by a number of variables, including their similarities and differences, their relationship, and the setting in which they are in, to name a few.

The third chapter begins with an analysis of the history of AI to highlight the developments in this field since its inception, presenting a story of how current AIs came to be. Included in the evolution of AI agents is one which aims to recreate human development in robots, particularly through learning about the environment through play and experimentation. Rather than providing the robot with specific information about the world, it instead learns through interacting with the world, much like human children. One particular robot named iCub

appears to be successful in its ability to learn the names of objects and how to use them; for example, picking up a mug by its handle upon hearing the word “mug.” Given the successes of iCub, is this robot a good candidate for future social robots? As mentioned previously, it is not, given its inability to understand the meanings of words, and in particular, those related to emotions. Although emotional recognition can be added into its functionality, this approach is insufficient for adequate communication and socialization because iCub is not sentient. The robot’s inability to experience its own version of emotions means it cannot understand what emotions are and what they mean to the person experiencing them. All the information the robot knows about emotions must be provided to it by humans, rather than being generated through its own body. Without this, any recognition exhibited by iCub is necessarily a simulation of understanding.

If iCub is incapable of understanding emotions, then what must be done to overcome this limitation? The fourth chapter introduces a new approach to building robots as developed by Pentti Haikonen. Prior to examining his architecture for neural networks and cognition, the chapter opens with a brief discussion on living beings and a fundamental principle which explains how life emerges and continues to flourish. Through self-organization or *autopoiesis*, physiological processes which constitute the bodies of organisms keep the individual alive, providing a means for restoring diminishing resources and avoiding harms by detecting changes in the environment. Although the history of AI briefly saw these principles adopted into research and development programs, their implementation of autopoiesis did not sufficiently extend to cognition to give rise to behaviours guided by an agent’s own interests. The architecture created by Haikonen is modelled on biological functionality, and as such, creates a better analogue of living beings than previous efforts in AI. By creating associative networks of neurons which

establish connections between stimuli and their effects on the robot's body, it is able to learn what a stimulus *means* to it and its continued functioning. The associations established in the robot's "mind" are thus properly grounded in its body, as the presence of pain and pleasure signals indicate how the stimulus impacts its existence as an organized unity. This architecture provides a foundation for empathy as the robot is able to access to its own sensations of pain and pleasure to understand the meanings of behavioural cues exhibited by humans. Despite requiring an entirely new approach to building robots, the architecture invented by Haikonen provides a promising solution to overcoming the limitations of existing robots and AI agents.

To close, the final chapter summarizes the central premises of this work to illustrate its main argument. Additionally, it offers two considerations for how robot empathy could be generated within a robot which implements the new approach proposed by Haikonen. If the robot were to feel discomfort or pain upon witnessing another in distress, it could feel motivated to assist in alleviating these feelings in order to reduce its own negative sensations. Alternatively, a robot which feels pleasure or happiness when interacting with humans may be more inclined to adopt the perspective of another. This direction aims to recreate the playful interactions involving mimicry witnessed between adults and young children, as these early interactions strengthen emotional bonds and facilitate the infant in its learning of intersubjectivity. By recreating this dynamic between robots and humans, a similar outcome of understanding and reciprocity may develop through repeated interactions.

Overall, the reason for this project and its investigation into robotics is to prevent or reduce potential human harms through an improved understanding of robot capacities. Our tendency to anthropomorphize objects and other natural phenomena introduces a risk of harm

arising from believing that current AIs are capable of caring about humans. These risks involve emotional and psychological harms from false beliefs and unrequited social connections, as an agent which utters sentimental statements only simulates the way humans and animals interact with people. Individuals may come to rely on artificial agents for emotional connections which they are incapable of, as they cannot feel anything at all. To prevent negative outcomes, a robust understanding of what robots are and the extent of their capacities is imperative. As various kinds of AIs and robots become commonplace, it will become increasingly important that our understanding of them is accurate, despite their appearances or claims made by their creators. Although social robots may appear intelligent and emotive, in actuality, this is a simulation. Unlike computerized robots, humans and animals exist as self-organizing unities with cognitive abilities for the sake of maintaining the processes which give rise to this self-organization. Until robots are constructed in an analogous manner to living beings, their behaviours are necessarily a simulation of understanding and empathy.

2 Social Robots for the Future

There is widespread discussion about the implementation of social robots but will they truly be a part of societies of the future? This chapter outlines the existing and growing demand to create robots capable of interacting with humans. I will argue that their presence in human societies is highly likely; however, to be successful, specific components are imperative for their widespread adoption. Social robots must be able to communicate with people by correctly interpreting utterances and behavioural cues, in addition to responding appropriately.

2.1 Examining the Demand for Social Robots

Historically, robots have been created for industrial environments for manufacturing. A new domain for robotic solutions has since emerged, redefining the commonly-held assumptions about what a robot is.[1] The word 'robot' is derived from *robota*, a Czech term for "forced labour" and originally introduced by playwright Karel Čapek in 1920.[2] The term denotes an artificial agent designed to perform dull and dangerous tasks in lieu of human beings. As such, the term has become associated with preprogrammed machines designed to carry out specific tasks in industrial environments. Given this programming, they are unable to act according to their own volition, and are incapable of being invested in the world in the ways in which humans

1. Seibt, 'Towards an Ontology of Simulated Social Interaction', 12.
2. *Merriam-Webster.com Dictionary*, s.v. "robot;" Bhaumik, *From AI to Robotics*, 4.

are. Today, robotics development also includes a branch dedicated to creating automated agents capable of interacting with people in a humanlike manner. To be successful, robots designed for social engagement must be able to use and understand linguistic utterances, as well as be able to detect human emotion to respond appropriately.[3] Because robots are often considered to be automated machines designed to perform repetitive movements regardless of context, a social robot shifts this notion substantially, as it must be able to operate in a nuanced, context-dependent manner. Since social interactions and acts of communication can refer to aspects of the current situation, social robots must be able to follow along and respond appropriately. This idea will be discussed in further detail later in this chapter.

The development of this new branch of robotics began in Japan in the early 2000's in response to the nation's demographic crisis. Japanese birthrates have been below replacement levels since 1980[4] and due to increased life expectancy, the elderly population has increased rapidly over the last fifty years.[5] As a result, there has been an increase in demand for governmental services like pensions, medical treatment, and long-term care, while fewer working-age adults are able to care for family members and contribute to the economy to support these government programs.[6] Today, Japan is the world's most aged country, as it has the lowest mortality rate in the world while also observing low fertility rates.[7] There are many needing care while there are fewer working age adults to support them. With a long history developing

3. Campa, 'The Rise of Social Robots', 107.
4. Suzuki, *Low Fertility and Population Aging in Japan and Eastern Asia*, 4.
5. Suzuki, 46.
6. Suzuki, 70.
7. Suzuki, 2.

industrial robotics, as well as pioneering research into developing humanoid robots, Japan was well-suited to approaching its demographic crisis with robotic solutions.[8] To meet growing demands for elderly care, Japanese researchers began to investigate different ways to support its aging population,[9] leading to promising developments[10] and subsequently becoming an industry leader in this domain.[11] Robots have been designed to assist individuals with sitting, standing, and moving around, in addition to providing support for healthcare workers with lifting and carrying patients.[12] For social support, Japanese researchers began experimenting with a soft companion robot named Paro in 2003 to determine whether it might assist in improving moods and social engagement in elderly individuals.[13] Indeed, this pet-like robot which resembles a baby harp seal has demonstrated positive results in improving emotional, cognitive, and social functioning within nursing home residents.[14] Further research using different types of companion robots has indicated similar results,[15] suggesting overall that these new robotic applications may deliver positive outcomes to elderly individuals. As such, technological solutions for meeting the rising demand of personal care indicates a promising direction for the future.

8. Wright, 'Inside Japan's Long Experiment in Automating Elder Care'.
9. Broekens, Heerink, and Rosendal, 'Assistive Social Robots in Elderly Care', 94; Čaić, Mahr, and Oderkerken-Schröder, 'Value of Social Robots in Services', 464; Pedersen, Reid, and Aspevig, 'Developing Social Robots for Aging Populations', 2.
10. Pedersen, Reid, and Aspevig, 'Developing Social Robots for Aging Populations', 8; Pohl, 'Robotic Systems in Healthcare with Particular Reference to Innovation in the "Fourth Industrial Revolution"', 20.
11. Ishiguro, 'Care Robots in Japanese Elderly Care', 256.
12. Ishiguro, 257–58.
13. Yu et al., 'Use of a Therapeutic, Socially Assistive Pet Robot (PARO) in Improving Mood and Stimulating Social Interaction and Communication for People With Dementia'; Sharkey and Sharkey, 'Granny and the Robots', 35.
14. Šabanović et al., 'PARO Robot Affects Diverse Interaction Modalities in Group Sensory Therapy for Older Adults with Dementia', 5; Shibata, 'Therapeutic Seal Robot as Biofeedback Medical Device', 2527.
15. Scoglio et al., 'Use of Social Robots in Mental Health and Well-Being Research'.

Similar demographic crises can be identified worldwide as well. Western countries like the United Kingdom, the United States, Canada, Australia, Germany, and Italy,[16] to name a few, have witnessed an increase in this trend since the 1960's.[17] Over the latter half of the 20th century and into the 21st, several other countries and regions have followed suit, including China,[18] India,[19] Iran,[20] and several Eastern European ones such as Belarus, Bulgaria, Georgia, Russia, and Ukraine.[21] This phenomenon has been termed the "Silver Tsunami,"[22] a metaphor to encapsulate the larger proportions of populations around the world who will require additional care and assistance, ranging from medical assistance to simple domestic and personal tasks. Given there will be fewer family members and professionals available to provide eldercare services over time,[23] this deviation is expected to put additional pressures on healthcare systems to meet growing demand for services related to aging.[24] Moreover, as women have increasingly entered the workforce over time and around the world, demand for care by paid workers has also increased as, traditionally, the elderly were cared for at home by female family members.[25] While societies have turned to using migrant labour to fill these positions, the number of skilled workers does not meet increasing demand identified worldwide.[26] Consequently, novel solutions

16. Horowitz, 'Who Will Take Care of Italy's Older People?'; Parker, 'Family Support in Graying Societies'.
17. Uhlenberg, *International Handbook of Population Aging*, 1.
18. Feng et al., 'China's Rapidly Aging Population Creates Policy Challenges In Shaping A Viable Long-Term Care System', 2764.
19. Chatterji et al., 'The Health Of Aging Populations In China And India', 1052.
20. Mehri, Messkoub, and Kunkel, 'Trends, Determinants and the Implications of Population Aging in Iran', 327.
21. Chawla, Betcherman, and Banerji, *From Red to Gray*, 1–2.
22. Henderson, Maniam, and Leavell, 'The Silver Tsunami', 153.
23. Czaja and Ceruso, 'The Promise of Artificial Intelligence in Supporting an Aging Population', 183.
24. Meskó, Hetényi, and Győrffy, 'Will Artificial Intelligence Solve the Human Resource Crisis in Healthcare?', 1; Olaronke, Oluwaseun, and Rhoda, 'State Of The Art', 43.
25. Rowland, 'Global Population Aging', 48; Colombo and Mercier, 'Help Wanted!', 3.
26. Chen et al., 'Human Resources for Health', 1985–86; Stone and Harahan, 'Improving The Long-Term Care Workforce Serving Older Adults', 109; Pohl, 'Robotic Systems in Healthcare with Particular Reference to Innovation in the "Fourth Industrial Revolution"', 20.

like the creation and implementation of robots in various roles related to eldercare have been proposed.[27] Though some robotic solutions will not require an ability to verbally communicate or socialize in a humanlike manner, others will require some degree of skill in this domain, indicating a need for the further development of social robots.

Aside from caring for elderly patients, social robots will likely be useful for healthcare in general. Moreover, given reflections upon the COVID-19 pandemic, these robots will also be useful for risky or dangerous healthcare situations.[28] In particular, researchers note a demand for robots to perform jobs like measuring patient temperatures.[29] In hospitals, the use of robots for emotional support, socializing, and information delivery or monitoring also carries the potential to improve healing and recuperation, as research suggests psychological factors like anxiety can delay the duration of post-surgery recovery.[30] Additionally, individuals rehabilitating from a stroke may benefit from physical training accompanied by a social robot, offering support and motivation to individuals required to carry out repetitive tasks.[31] Similarly, social robots may also be useful in pediatric settings, offering emotional support to children and providing a distraction from clinical procedures.[32]

Social robots may also have a positive effect on health outcomes if they can alleviate social isolation, as feelings of loneliness significantly impact individual well-being.[33] Given the

27. Vercelli et al., 'Robots in Elderly Care', 38.
28. González-González, Violant-Holz, and Gil-Iranzo, 'Social Robots in Hospitals'; Aymerich-Franch, 'Why It Is Time to Stop Ostracizing Social Robots', 364.
29. Wirtz, Kunz, and Paluch, 'The Service Revolution, Intelligent Automation and Service Robots', 41.
30. Jamison, Parris, and Maxson, 'Psychological Factors Influencing Recovery from Outpatient Surgery', 36.
31. Polak and Tzedek, 'Social Robot for Rehabilitation', 157.
32. Cifuentes et al., 'Social Robots in Therapy and Care', 67; Jeong et al., 'Huggable', 125.
33. Ghafurian, Ellard, and Dautenhahn, 'Social Companion Robots to Reduce Isolation', 43.

increasing rates of social isolation identified around the world over the last few decades,[34] exacerbated by the COVID-19 pandemic,[35] social robots may provide assistance in this domain as well.[36] Although human social supports are ideal, if robots are able to generate positive outcomes for individuals, these outcomes may be more beneficial over not having a robot at all. Regardless, human-to-human support for individuals, however, ought to remain the norm. That said, specific situations may arise where patients and/or their families can benefit from some degree of support from robotic solutions. Similarly, as the demand for support related to mental health conditions often outpaces the available supply of services, especially in developing countries,[37] social robots may also assist in these domains as well.[38] Overall, however, more research is required to determine the impact of robotic solutions for directly improving mental health outcomes.[39] While it seems likely these robots will provide some degree of social support for individuals in need, it remains to be determined whether these solutions will involve therapeutic approaches which go beyond a friendly conversational partner.

Demand for social robots is also growing in domains related to childcare and education,[40] especially for those with autism.[41] Researchers have noted a unique opportunity for social robots within educational settings for individuals with autism. In particular, the company LuxAI has

34. Snell, 'The Rise of Living Alone and Loneliness in History', 22.
35. Hu, 'A Design of Service Robots in Epidemic Disease Isolation Environment', 12.
36. Ghafurian, Ellard, and Dautenhahn, 'Social Companion Robots to Reduce Isolation', 43.
37. Jacob et al., 'Mental Health Systems in Countries', 1061; Wang et al., 'Use of Mental Health Services for Anxiety, Mood, and Substance Disorders in 17 Countries in the WHO World Mental Health Surveys', 846; Robiner, 'The Mental Health Professions', 601.
38. Guemghar et al., 'Social Robot Interventions in Mental Health Care and Their Outcomes, Barriers, and Facilitators', e36094; Scoglio et al., 'Use of Social Robots in Mental Health and Well-Being Research', 13322.
39. Robinson, Cottier, and Kavanagh, 'Psychosocial Health Interventions by Social Robots', e13203.
40. Belpaeme et al., 'Social Robots for Education', 7; Edwards and Cheok, 'Why Not Robot Teachers', 349; Johal, 'Research Trends in Social Robots for Learning', 81.
41. Dautenhahn, 'Roles and Functions of Robots in Human Society', 446.

developed QTrobot along with an educational curriculum specifically for this purpose.[42] Individuals on the ASD spectrum tend to prefer interacting with objects over people, and generally demonstrate an interest in technology given the predictability of behaviours.[43] Moreover, with robots, social interactions are simplified with fewer behavioural cues to understand and keep track of, potentially reducing feelings of anxiety as a result.[44] Robots are also likely to help educate children with autism as these individuals often require more instruction through repetition than their neurotypical peers.[45] Because robots do not tire nor lose patience with their students,[46] individuals with autism may benefit from interacting with these agents over human counterparts. Indeed, several studies have indicated increased levels of attention and engagement when using social robots over human instructors for learning and developing skills.[47] Given this, however, it is important to mention that autism consists of a spectrum of abilities and challenges, where some individuals may benefit from robot-mediated learning while others might not.[48] Additionally, it remains to be determined whether social skills gained from practice with robots can be successfully applied and incorporated into interactions with humans.[49] Despite the uncertainties, social robots are likely to provide a degree of

42. 'Qtrobot Curriculum for Autism'.
43. Dautenhahn, 'Roles and Functions of Robots in Human Society', 446.
44. Cabibihan et al., 'Why Robots?', 614; Waltz, 'Therapy Robot Teaches Social Skills to Children With Autism'.
45. Huijnen, AS Lexis, and de Witte, 'Robots as New Tools in Therapy and Education for Children with Autism', 3.
46. Singer, 'With Endless Patience and Never Tiring, Robots Are Being Used in Connecticut to Connect with Children with Autism'.
47. Pennisi et al., 'Autism and Social Robotics', 178; Scassellati, Admoni, and Matarić, 'Robots for Use in Autism Research', 289.
48. Diehl et al., 'The Clinical Use of Robots for Individuals with Autism Spectrum Disorders', 255.
49. Scassellati, Admoni, and Matarić, 'Robots for Use in Autism Research', 278.

assistance to some children with autism,[50] indicating an incentive for further research and development in both domestic and childcare settings.

Additionally, social robots are poised to be useful for other services as well, including customer service; such as in airports and train stations,[51] along with restaurants, theme-parks, tourism, and hospitality.[52] Social robots would be able to assist travellers with several services both related and unrelated to travel, such as baggage-claim and customs for international travel. Furthermore, these robots may be particularly useful for travellers requiring translation assistance, as local signage may not include language individuals are familiar with. Individuals would be able to ask the robot a question in their preferred language, rather than having to translate or speak a foreign language to receive assistance. Ideally, these robots should do more than merely giving directions, and instead, act as a friendly face for the airport or station. This indicates an opportunity for features frequently associated with entertainment settings as well, including the ability to tell appropriate jokes, perform a brief song or dance, or ask trivia questions. In these cases, the robot's ability to detect emotions would provide it with feedback as to whether the audience's reactions or experiences engaging with the robot were positive or negative, learning over time as it engages with its audiences.

Emotional sensitivity, however, is not necessary for all social tasks surrounding customer service, as observed in a hospitality robot developed in Italy. Brillo the bartender robot is able to make drinks and engage in conversation with patrons.[53] In this case, a high degree of emotional

50. Cabibihan et al., 'Why Robots?', 615.
51. Thunberg and Ziemke, 'Are People Ready for Social Robots in Public Spaces?', 482.
52. Ivanov and Webster, 'Robots in Tourism', 1065.
53. Rogers, 'BRILLO the Bartending Robot Can Fulfill Your Social Needs While Slinging Cocktails'.

sensitivity may not be required, as the robot's primary task is to efficiently serve drinks in a friendly manner. As such, some social robots may simply appear to be sociable, while others must be able to socialize in a nuanced, intelligent manner, depending on the role they are fulfilling. It seems likely that social robots with a variety of capacities and skill sets will emerge over time, as indicated by market demands. As the breadth of domains in which robots are proven to be useful continues to increase, a greater number of industries will likely find uses for them. If successful implementations continue to indicate a potential to improve client satisfaction, working conditions, and net profits, the adoption and popularity of social robots is likely to grow.

Overall, the presence of global demand, in conjunction with financial incentives, will continue to motivate companies, academic institutions, and other organizations to produce systems capable of fulfilling a variety of roles requiring some degree of communicative and social dexterity. Thus, social robots appear to remain an important technology to develop for various purposes within society. Given the aforementioned demographic needs identifiable worldwide, it seems likely that the development of social robots will continue. Due to shortages in the supply of human labour related to care settings, along with an increasing demand for eldercare services, social robots are poised to become a way of addressing these issues. Though governments and institutions may regulate the ways in which robots can be used or implemented, it appears that demand will continue to drive their development in some fashion for the foreseeable future.

2.2 Requirements for Social Roles

Given this interest in social robots, these agents are likely to become integrated within societies around the world. For them to be deemed acceptable by human populations, however, they will need to have certain capacities to be useful and safe. To determine the scope of these capacities, it is important to understand the necessary elements for social robots. To some, these agents simply consist of an expanded version of regular industrial robots with a *social interface,* in which visual features provide signals for acting in specific ways with these agents.[54] On the other hand, some envision social robots as specialized systems which learn and develop over time, building on previous knowledge to perform complex skills similarly to young humans. This approach can be observed within the domain of *developmental robots* as an outcome of AI research over the 20th century, and will be discussed further in Chapter 3. In general, a variety of architectures or approaches will likely emerge over time, giving rise to different kinds of social robots in the future.

For robots to fulfill roles typically performed by humans, they will require certain capacities to interact with and be accepted by the general public.[55] This can be achieved by recreating aspects of human cognition in computer code. To act in a humanlike manner, robots will need to perceive and express emotions in ways which are appropriate for specific contexts,

54. Hegel et al., 'Understanding Social Robots', 171.
55. Beer et al., 'Affective Human–Robot Interaction', 361; Forgas-Coll et al., 'How Do Consumers' Gender and Rational Thinking Affect the Acceptance of Entertainment Social Robots?', 14.

as well as communicate through language and behavioural cues. Moreover, these robots must be able to recognize, remember, and learn about other agents they interact with.[56]

When polled, individuals rate capacities for empathy, logical thinking, and explanation as some of the most desirable capacities in social robots.[57] In another study investigating variables for the acceptance of social robots, similar themes also emerge as survey results indicated factors like usefulness, sociability, companionship, and behavioural control were rated as important characteristics.[58] Additionally, people have expressed interest in a social robot's ability to work with others in a friendly manner.[59] Experiments with social robots indicate people engaged with emotionally expressive robots are more likely to rate the interaction as highly satisfying, in comparison to interactions with less expressive robots.[60] Because emotion is an important aspect of conversation and socialization, individuals will likely prefer to interact with robots whose behaviours respond appropriately to one's mood or emotional state. Furthermore, robots deployed in public settings must be easy to use and engage with,[61] where visual appearances adequately indicate the range and style of behaviours performed by the agent.[62] Overall, this range of characteristics suggests individuals want social robots to act as humanlike as possible, making their behaviours easier to understand and interact with.

---

56. Hegel et al., 'Understanding Social Robots', 170.
57. Leite et al., 'Empathic Robots for Long-Term Interaction', 329; Malle and Thapa, 'What Kind of Mind Do I Want in My Robot?', 196.
58. de Graaf and Ben Allouch, 'Exploring Influencing Variables for the Acceptance of Social Robots', 1485.
59. Gompei and Umemuro, 'Factors and Development of Cognitive and Affective Trust on Social Robots', 51.
60. Leite et al., 'Empathic Robots for Long-Term Interaction', 330.
61. Leite, Martinho, and Paiva, 'Social Robots for Long-Term Interaction', 297.
62. Leite, Martinho, and Paiva, 300.

If these expectations are not met, users may consider social robots unhelpful or inappropriate.[63] Over time, if disappointment frequently arises from a mismatch in expected ability and actual ability, individuals may begin to lose trust in social robots more broadly. Failures to portray a robot's specific abilities risks the dissolution in human-robot cooperation and relations in the future. This mistrust introduces the potential to negatively impact human populations if robots are to become widely implemented, especially for safety-critical tasks such as search and rescue operations.[64] Moreover, studies on human trust suggests individuals are concerned about how well these robots are attuned to safety and security, and whether these robots are able to independently perform tasks. Therefore, it seems social robots should be required to act in a humanlike manner to facilitate cooperation with people.

Thus, in a variety of social roles, interactions with robots must be as intuitive and familiar as possible. To be successful, social robots are required to perform a wide range of behaviours to serve and work with people. The challenge arises from the nature of these interactions, as social situations are highly dynamic and have the potential to be unpredictable. Furthermore, language relies on words associated with contextual factors to convey certain ideas, in addition to sounds and behaviours which are less linguistic in nature, such as body language. Social robots will likely need the ability to learn about aspects of the world to act appropriately in a reliable manner, drawing on previous experience to determine the best course of action in a given scenario. Therefore, the concept of a "social robot" involves a suite of cognitive capacities which extend beyond the mere production of language, as communication requires knowledge and

63. Jokinen and Wilcock, 'Expectations and First Experience with a Social Robot', 514.
64. Kwon, Jung, and Knepper, 'Human Expectations of Social Robots', 463.

understanding of the world and its elements. Moreover, the robot must be able to incorporate forms of feedback from the environment to adjust their behaviours accordingly, similarly to people and animals. In this way, humanlike behaviours can be created in machines which sufficiently produce capacities and behaviours required for social interactions.

2.3 Human Communication

It is clear that social robots will need to have an understanding of language and social cues if we want them to adopt social roles and interact with humans. A robot's ability to correctly track human utterances and behaviours, however, relies on learning about aspects of its environment or the wider world to know what a human might be referring to.[65] Moreover, while a significant aspect of socialization is following what is being said, an effective robot will need to respond appropriately through a variety of different sounds, words, and bodily expressions.[66] Non-verbal behaviours may consist of gestures, body movements, eye contact, and facial expressions.[67] All together, the acts themselves and the content of linguistic expressions act as cues which provide information about what the robot is conveying or referring to. While language involves the use of syntax and semantics, it also involves *pragmatics*, or the elements of expression not represented

65. Thomaz, Hoffman, and Cakmak, 'Computational Human-Robot Interaction', 122.
66. Pantic and Vinciarelli, 'Social Signal Processing', 144; Esposito and Jain, 'Modeling Social Signals and Contexts in Robotic Socially Believable Behaving Systems', 6; Hegel et al., 'Understanding Social Robots', 172.
67. Pantic and Vinciarelli, 'Social Signal Processing', 148; Thomaz, Hoffman, and Cakmak, 'Computational Human-Robot Interaction', 134.

by the language itself.[68] This implicit information is embedded in an expressive act, such as the use of specific words or sounds, vocal pitch and amplitude, and speech pauses.[69] Implicit acts of communication may also be supported by body language and gestures, particularly hand gestures signifying actions or intentions. These subtle methods of communication appeal to features of situational context, as variables such as the current place and time are often relevant for communicative efforts.[70] Currently, text-based agents like *large language models* (LLMs) do not appeal to information related to situational context, introducing limitations which embodied social robots may be able to overcome.

Another element to consider for social robots is the requirement of joint attention for working with others toward a shared goal.[71] Both robotic and human agents are required to coordinate their intentions or plans in order to execute behaviours appropriate for producing desired outcomes. This involves an ability to infer what the other agent is likely attending to or thinking about, requiring them to align their understanding of the goal and the necessary actions for accomplishing it.[72] Robots are thus not only required to have sufficient background knowledge about the world, but are also required to skillfully reason about the ways in which others see the world.[73] Robots will be required to understand the reasons why individuals perform certain actions, and to make assumptions about what others are directing their attention toward.[74] Therefore, they must have an ability to interpret human behaviours to understand what

68. Portner, *What Is Meaning?*, 94.
69. Esposito and Jain, 'Modeling Social Signals and Contexts in Robotic Socially Believable Behaving Systems', 6; Thomaz, Hoffman, and Cakmak, 'Computational Human-Robot Interaction', 126.
70. Raczaszek-Leonardi, Nomikou, and Rohlfing, 'Young Children's Dialogical Actions', 212–13.
71. Clodic et al., 'Key Elements for Human-Robot Joint Action', 164.
72. Tomasello and Kruger, 'Joint Attention on Actions', 312–13.
73. Clodic et al., 'Key Elements for Human-Robot Joint Action', 168.
74. Clodic et al., 169.

they are focusing on and their larger goals, such that the robot can coordinate its own behaviours to act appropriately. Otherwise, confusion and misunderstandings are likely to emerge, and if user frustration results, it may impede further cooperation with the robot.

Together, this collection of skills and requirements for social robots suggests human communication is more complex than merely combining semantics and syntax, requiring robots to master a variety of verbal and non-verbal abilities. From this analysis on the breadth of human communication, it appears several abilities are needed if these robots are to be useful for our purposes. Given the financial incentives present for developing useful applications, the necessary research and development for generating robotic solutions for a variety of purposes is likely to continue into the foreseeable future.

Even if robots are one day able to satisfy these requirements, there is still an important aspect about human communication to consider. When people interact with one another, they are able to understand the meanings of words and behaviours, to some degree at least. Similarly for written text; words and symbols are associated with meanings that humans learn through experience. The *symbol grounding problem*, namely how meanings or referents become associated with words and concepts, is one which can be answered by research from developmental psychology.[75] For our current approach to AI and robotics, however, the problem remains. The reason being is that meanings are extrinsic to language and symbol systems, existing in the relationship between minds and the world, rather than within the linguistic system itself.[76] For any robot to understand the meaning of a word, the symbol grounding problem must

75. Siegler, DeLoache, and Eisenberg, *How Children Develop*, 216.
76. Harnad, 'The Symbol Grounding Problem', 339.

be resolved. Although some may claim that robotics as an approach to AI provides a solution to this problem, the debate still remains, as will be discussed in later chapters of this work. Without a solution to the symbol grounding problem, any behaviours exhibited by robots which appear to indicate understanding is an illusion.

Moreover, interactions with social robots remain a mere simulation of human interactions, as social situations involve reciprocity and cooperation.[77] Not only do robots not understand the meanings of words and behaviours, any indication of reciprocity and cooperation is also an illusion. According to Johanna Seibt, Founder and Director of the Research Unit for Robophilosophy and Integrative Social Robotics (RISR),[78] a problem arises when we consider the foundations of social interactions more broadly.[79] She claims we cannot treat human-robot interactions as fictionalist analogues of human interactions because the concepts used in these fictionalist accounts do not apply to robots. This is because sociality stems from "joint attention to basic patterns" which begins in childhood, where providing responses and engaging in turn-taking rely on "pre-conscious" abilities intrinsic to human minds.[80] Despite attempts to recreate these capacities in robots, the result is a *model* which imitates the phenomenon identified in human-human interactions. Therefore, human-robot interactions are, at best, "asymmetric simulated social interactions," in which even a perfectly simulated interaction remains ontologically distinct from the human-human interactions the model aims to recreate.[81] It will

77. Hakli and Seibt, '"Sociality and Normativity for Robots"', 2.
78. DIGHUM "Johanna Seibt"; Revsbech, "When Humans and Robots Meet."
79. Seibt, 'Towards an Ontology of Simulated Social Interaction', 14.
80. Seibt, 36.
81. Seibt, 37.

become exceedingly important to keep this idea in mind as robots continue to advance, as humans have a tendency to anthropomorphize objects[82] and ascribe to them abilities that they do not possess. Although they may *simulate* certain abilities, robots do not care about their own existence or existence as a social entity. As a result, a robot is not a *being* at all, as sociality is in part an evolved capacity for the advantage it provides for the organism's own survival and reproduction.[83] Given that a robot is a machine, it cannot demonstrate sociality as it does not care about its own existence or the existence of others.[84] Robot behaviours are generated by a mechanistic sequence of events executed without awareness, a topic which will be discussed in later chapters.

It is important to study the foundations of robot behaviours because our understanding of them impacts the way we think about their actions and abilities. A new and unique risk to human well-being is introduced by social robots when individuals form false beliefs surrounding their actions. These risks involve emotional or psychological harms from the formation of unrequited or one-sided attachments to artificial agents. Today, chatbots and smartphone apps designed as "AI girlfriends" or other romantic interests are designed to appear interested or emotionally invested in the user, mimicking the way humans behave in close relationships. This results in the formation of emotional bonds to agents which do not feel anything at all, deceiving the user as a result. Given that we are witnessing a "loneliness epidemic" which can be observed in many

82. Turkle, *The Second Self*, 57; Waytz et al., 'Making Sense by Making Sentient', 423; Waytz et al., 'Causes and Consequences of Mind Perception', 384.
83. Korb and Heinze, 'Major Hurdles for the Evolution of Sociality', 298.
84. Bickhard, 'Robot Sociality', 62.

countries around the world,[85] the risk of harm only continues to grow. If these apps or technologies were to become more popular over time, an increasing number of individuals will likely fall victim to this trick, potentially leading to disappointment and despair as a result. Because social robots and other AIs are not living beings, they are not invested in their own survival as agents. They are not sentient and therefore, are not motivated by needs and desires, including the ability to develop feelings for people and a desire to form relationships. Moreover, given our tendency to anthropomorphize machines,[86] the risk still remains for social robots which are not explicitly designed to mimic romantic interest. Individuals may become overly emotionally attached or dependent on artificial agents as a result of loneliness and anthropomorphization, projecting capacities and abilities onto them which do not exist in reality.

This chapter has argued for an increasing likelihood of the widespread implementation of social robots, given identifiable needs for them to serve in roles typically fulfilled by humans. Additionally, it has argued that at least some social robots will be required to understand the meanings of words and behaviours used by people to communicate with one another. Until the symbol grounding problem has been solved, this understanding will remain illusory and instead, social robots will merely track or follow what is being said or implied. Social robots will also require some amount of background information of the world and its features to act as a form of knowledge, as an understanding of the world is a significant aspect of communication. Furthermore, these agents will likely need to be able to interpret various non-verbal behaviours

85. Depounti, Saukko, and Natale, 'Ideal Technologies, Ideal Women', 724; Brandtzaeg and Følstad, 'Chatbots', 43; Jacobs, 'Digital Loneliness—Changes of Social Recognition through AI Companions', 2; Surkalim et al., 'The Prevalence of Loneliness across 113 Countries', 7.
86. Turkle, *The Second Self*, 57.

as well. For example, an emotion can be communicated through the use of language but also through visual cues like facial expressions.[87] In addition to being able to follow human communication, social robots will also need to produce language and behaviours which people can understand and deem appropriate. This will involve choosing the correct words for conveying a specific meaning but also an ability to perform supplementary behavioural cues which further support or contribute to the message or idea communicated. Together, an ability to understand and communicate a variety of social cues and linguistic elements constitute the core capacities required for social robots. If they fail to meet these requirements, human individuals may not enjoy interacting with them, impacting their overall acceptance as a result. In certain contexts, inadequacies in social robot performance may negatively impact human life; for example, patients in healthcare settings may have their dignity infringed upon or safety threatened if a robot does not treat them distinctly from an inanimate object. Therefore, the stakes for social robots are rather high as simulated forms of social interactions have the potential to deeply affect emotional, psychological, and physical well-being.

Considering that empathy is rated as a highly desirable capacity in social robots,[88] especially in healthcare settings,[89] it is important to examine what empathy involves and whether this is a tenable goal. Given the pressing need for healthcare workers, social robots are likely to be found in roles related to care and human interaction in the future, making empathy a

87. Bartneck et al., *Human-Robot Interaction*, 119.
88. Bagheri et al., 'A Reinforcement Learning Based Cognitive Empathy Framework for Social Robots', 1079; Leite et al., 'Empathic Robots for Long-Term Interaction', 329; Malle and Thapa, 'What Kind of Mind Do I Want in My Robot?', 196; Park and Whang, 'Empathy in Human–Robot Interaction', 16; Quick, 'Empathizing and Sympathizing With Robots'.
89. James, Watson, and MacDonald, 'Artificial Empathy in Social Robots', 632; Pepito et al., 'Intelligent Humanoid Robots Expressing Artificial Humanlike Empathy in Nursing Situations', 1; Vallverdú and Casacuberta, 'Ethical and Technical Aspects of Emotions to Create Empathy in Medical Machines', 341.

significant component of social robot behaviour. Although robots may one day act as if they can understand how an individual is feeling, given our tendency to anthropomorphize robots, we should carefully consider whether those projections are accurate. It will be important to ensure our beliefs about a robot's abilities are accurate so as not to attribute these machines with capacities they do not possess. This can be accomplished by the robot's form and appearance, as it provides an indication to users the kinds of behaviours or abilities it is likely to perform.[90] For example, a humanoid robot indicates humanlike abilities, where the inclusion of a smiling mouth suggests a friendly, communicative agent. A mismatch between human expectation and robot functionality has the capacity to disappoint or confuse users, potentially leading to negative experiences which decrease the likelihood of subsequent use or interaction. Moreover, significant harms can be introduced if humans rely on robots for behaviours or tasks which they cannot fulfill. If robots are to care for humans in a respectful manner, patients, staff, visitors, and other personnel must be aware of social robot capabilities and limitations. I will argue that existing robots exhibiting emotions involve a simulation as the robot is unable to truly understand what emotions are. This is in virtue of the relationship between the robot's body and its "mind," as I aim to explain in this dissertation. In regards to empathy, the question about whether these outcomes are even possible must be raised. To answer this question, however, a separate inquiry must be undertaken prior to this investigation: what is empathy and how does it manifest? The next chapter explores and provides the philosophical history of this capacity, in addition to an investigation into the scientific literature on theory of mind to determine whether social robots will one day be capable of acting in such a manner.

---

90. Bartneck et al., *Human-Robot Interaction*, 46.

3 Analyzing Empathy

Within the philosophical and scientific literature on empathy, two views on *theory of mind* have historically characterized how acts of empathy occur. Referred to as *theory theory of mind* and *simulation theory of mind*, each attempt to explain different facets of perspective-taking required for empathy. Both theories of mind, however, are insufficient for fully encapsulating what empathy involves, despite their significant contributions. After identifying the short-comings of each theory, I discuss a preferable third option called *interaction theory* which accounts for contextual variables impacting the accuracy of an act of empathy.

3.1 Empathy as a Philosophical Concept

The philosophical literature on empathy begins in the late 19th century and involves a few related but distinct ideas. Various positions and theories about the mind and human behaviour have influenced the way philosophers have characterized empathy throughout history. With scientific advancements enhancing our capabilities to study the brain, *folk psychology* became insufficient for explaining empathy. Eventually, the discovery of mirror neurons gave new life to the study of empathy, as brain activity could now be used to account for ideas or theories generated from philosophy. Although controversy surrounding the existence of special mirror neurons has arisen since their discovery, the activity these cells generate provides insight into how imitation, mimicry, and empathy are accomplished. That said, these empirical findings and the resulting

simulation theory [*of mind*] remains insufficient for a complete understanding of what empathy is and how it is generated within individuals. The reason being is that it does not provide an account for how and why acts of empathy may be correct or incorrect. Though our understanding of empathy has improved by appealing to scientific evidence, simulation theory alone remains insufficient for explaining what empathy involves. Theorizing in the form of top-down reasoning about the internal states of others is still required at times, though the conceptual framework offered by *simulation theory* provides a better understanding of how empathy occurs. As such, acts of empathy involve both cognitive and affective elements.[91] While *theory theory* [*of mind*] seems to adequately articulate many of the cognitive aspects of empathy, and *simulation theory* is able to explain many affective elements,[92] together, these accounts still do not provide a sufficient explanation for empathy. As such, an alternative perspective is required for a comprehensive theory of this remarkable capacity.

The term 'empathy' was introduced in 1909 by English psychologist Edward Titchener, a student of Wilhelm Wundt,[93] as a translation of the German word *Einfühlung* or "feeling into."[94] *Einfühlung* was introduced as a technical term within the field of Aesthetics by Robert Vischer in 1873, and eventually popularized by Theodor Lipps to describe a process of projection.[95] Lipps describes aesthetic experiences as involving the use of a specific perceptual ability which causes us to imitate movements and expressions as perceived from elements of our physical

91. Cuff et al., 'Empathy', 147; Schertz, 'Empathy as Intersubjectivity', 165.
92. Shamay-Tsoory, 'Empathic Processing', 216.
93. Angell, 'Titchener at Leipzig', 195–96.
94. Agosta, *Empathy in the Context of Philosophy*, 6; Stueber, 'Empathy', sec. 1.
95. Montag, Gallinat, and Heinz, 'Theodor Lipps and the Concept of Empathy', 1261; Stueber, *Rediscovering Empathy*, 6.

environment.[96] According to Lipps, we project a part of ourselves into or onto an object, and how we feel in this projection informs our attitudes or thoughts about the object itself.[97] We tend to feel good about an object if it possesses a form which, when aligned with our own human body, contributes or generates a kind of energy or "vitality" within us.[98] Because this projection occurs subconsciously, we ascribe the associated feelings onto the object we are observing, attributing them to the object rather than to ourselves. Thus, when viewing another person, Lipps describes an innate tendency to experience oneself in their perceived state, where we experience the other's feelings as our own due to this projection.[99] By using our own mind as an analogy for understanding another's mind, we are able to know about the inner experiences of others.[100]

It has been speculated that Lipps may have been building off of David Hume's work given his familiarity with *A Treatise of Human Nature* after translating it into German.[101] In it, Hume discusses ideas associated with mental projection: "The minds of men are mirrors to one another."[102] Hume describes a transmission of emotion between one person and another in which the observer notices a distinction between his own feelings and the feelings of another.[103] The term he uses, however, is not 'empathy' but 'sympathy' to describe a process in which the observation of behaviours provides an idea about the affective state of another. To Hume, *ideas* generally follow from *impressions,* in which impressions refer to perceptions and sensations,

96. Coplan and Goldie, 'Introduction', XII; Stueber, *Rediscovering Empathy*, 7.
97. Zahavi, 'Empathy and Other-Directed Intentionality', 130.
98. Stueber, *Rediscovering Empathy*, 7–8.
99. Coplan and Goldie, 'Introduction', XII.
100. Stueber, *Rediscovering Empathy*, 9.
101. Coplan and Goldie, 'Introduction', XII; Montag, Gallinat, and Heinz, 'Theodor Lipps and the Concept of Empathy', 1261.
102. Hume, *A Treatise of Human Nature*, 236.
103. Coplan and Goldie, 'Introduction', X; Schertz, 'Empathy as Intersubjectivity', 167.

while ideas consist of "faint images" of perceptions for thinking and reasoning.[104] During instances of empathy, Hume specifies that this relation is reversed, in which impressions follow from ideas:

> When any affection is infus'd by sympathy, it is at first known only by its effects, and by those external signs in the countenance and conversation, which convey an idea of it. This idea is presently converted into an impression, and acquires such a degree of force and vivacity, as to become the very passion itself, and produce an equal emotion, as any original affection. However instantaneous this change of the idea into an impression may be, it proceeds from certain views and reflections, which will not escape the strict scrutiny of a philosopher, tho' they may the person himself, who makes them.[105]

These ideas of empathy as 'sympathy' seem to have also influenced the perspective of Adam Smith as reflected in his work *The Theory of Moral Sentiments*.[106] For both Hume and Smith, 'sympathy' provides a way to explain why individuals driven by self-interest would be concerned for others and the society they live in.[107] While Hume's view focuses on lower-level abilities which reflect the affective states of others, Smith's view involves a higher-level or more cognitive approach, articulating an ability to imagine oneself in another's situation.[108] To know how another feels, according to Smith, we project ourselves into their position or point of view to reason about their actions or behaviours. While this reasoning process relies on introspection, it is performed with full awareness to generate a top-down process, differing from Hume's understanding of empathy as a subconscious, bottom-up process.[109] Thus, the debate between whether our capacity for empathy is more deliberative or affective in nature is present in the

---

104. Hume, *A Treatise of Human Nature*, 7.
105. Hume, 206.
106. Schertz, 'Empathy as Intersubjectivity', 168.
107. Stueber, *Rediscovering Empathy*, 31.
108. Schertz, 'Empathy as Intersubjectivity', 169; Coplan and Goldie, 'Introduction', XI.
109. Schertz, 'Empathy as Intersubjectivity', 170.

origins of the identification of the capacity itself, where the term used to describe this ability is not the one we use today.

To avoid confusion, however, it is important to briefly articulate the contemporary usage of 'sympathy' and 'empathy'. This is not an easy distinction to draw, especially considering historical uses of 'sympathy' for describing abilities which have since been considered as empathic in nature. It has been suggested that sympathy is associated with feelings of sorrow or concern for another,[110] however, 'sympathy' has also been associated with emotional contagion, in which the perceived states of others are adopted as one's own.[111] Empathy, on the other hand, is an *other-oriented* emotional response in which the perspective of another person is adopted by an observer, aiming to experience thoughts and feelings as if they were the other individual.[112] It can be characterized as "feeling *as*" rather than "feeling *for*" another person.[113] While this distinction between sympathy and empathy may oversimplify a richer understanding of this conceptual relationship, for the purposes of this discussion, we can think of the difference as residing in to whom these feelings belong. If an observer is adopting the feelings of others to understand or experience the feelings of another, they are demonstrating empathy. Alternatively, if the observer attributes the feelings as belonging to another, and instead modifies their own emotional state in response to the experiences of another person, it can be characterized as sympathy.

---

110. Darwall, 'Empathy, Sympathy, Care', 273; Eisenberg and Eggum, 'Empathic Responding', 71; Cuff et al., 'Empathy', 145.
111. Burns, 'Empathy, Simulation, and Neuroscience', 214; Hein and Singer, 'I Feel How You Feel but Not Always', 154.
112. Coplan and Goldie, 'Introduction', 10.
113. Hein and Singer, 'I Feel How You Feel but Not Always', 157.

Lipps's work on *Einfühlung* would go on to influence the work of others, including Titchener in his study of psychology,[114] but also in the writings of phenomenologists including Martin Heidegger,[115] Edmund Husserl,[116] Edith Stein,[117] and Max Scheler.[118] Criticism of Lipps's work included a dissatisfaction with his appeal to instinct and a lack of empirical evidence. Additionally, given that individuals differ from one another, Lipps's ideas were criticized for their inability to explain how one's own experience generates an understanding of how another person may experience that same stimulus.[119] Similarly, Scheler notes that Lipps's account does not consider the accuracy of this projection, and whether our understanding of another can be justified given its origins in subjective experience.[120] Phenomenological accounts of empathy have contributed to further identifying features of empathy and how it functions, however, over time, they would fall from favour given their independence from empirical evidence.[121] As the 20th century saw a rise in the development of scientific methodologies for studying the mind, simply theorizing about empathy was no longer sufficient for establishing an explanation of this ability.[122] Although these theoretical accounts were foundational for developing a philosophical understanding of empathy, the concepts it relied upon were derived from an outdated methodology and considered to be akin to conjecture.[123] Despite this,

---

114. Debes, 'From Einfühlung to Empathy', 290; Wispé, 'History of the Concept of Empathy', 20.
115. Agosta, *Empathy in the Context of Philosophy*, 5.
116. Agosta, 113.
117. Stein, *On the Problem of Empathy*, 11; Coplan and Goldie, 'Introduction', XIII.
118. Zahavi, 'Empathy and Other-Directed Intentionality', 133.
119. Zahavi, 131.
120. Zahavi, 132.
121. Stueber, 'Empathy', sec. 2.
122. Debes, 'From Einfühlung to Empathy', 318–19; Stueber, *Rediscovering Empathy*, 17; Wispé, 'History of the Concept of Empathy', 23–24.
123. Stueber, *Rediscovering Empathy*, 2.

philosophical theories still provide an essential perspective for analyzing and interpreting scientific data of behaviour and neural activity, and as such, are still required for guiding discussions on what empathy involves and how it is achieved.[124]

Upon noticing a brain region in macaque monkeys containing cells which fire while observing the actions of others and also when performing an action,[125] *simulation theory* gained a renewed interest in the late 1980's when evidence for the existence of mirror neurons was discovered.[126] This is due to the theory's position which argues that empathy is accomplished in the act of imagining how another person is experiencing some stimulus or situation. We are able to do so given the general structural and functional similarities between the minds and brains of human individuals.[127] Consequently, it has been described as a "process-driven" rather than a "theory-driven" procedure, as it requires individuals to perform cognitive tasks to know what another is feeling.[128] To contemplate how or what another person is feeling, we use our own perspective as a model for how another may see the world, using the same causal mechanisms when expressing our own feelings. Although Lipps's views were introduced prior to any empirical evidence on the matter, his views eventually received empirical support. The neural activity discovered by researchers seemed to indicate a "common representation format" which enables individuals to mimic or imitate the actions of others.[129] It was thought these cells

124. Stueber, 20.
125. Iacoboni, 'Imitation, Empathy, and Mirror Neurons', 659.
126. Goldman, 'Two Routes to Empathy', 33; Oberman and Ramachandran, 'Reflections on the Mirror Neuron System', 40.
127. Stueber, *Rediscovering Empathy*, 111.
128. Goldman, 'Interpretation Psychologized', 173; Pfeifer and Dapretto, '"Mirror, Mirror, in My Mind"', 186.
129. Iacoboni, 'Imitation, Empathy, and Mirror Neurons', 657; Decety and Jackson, 'The Functional Architecture of Human Empathy', 80; van Baaren et al., 'Being Imitated', 31.

accomplish this through genetic encoding, in which natural selection gave rise to unique cells which enable individuals to understand the goal the other's actions aim to achieve.[130] By coding for actions by the effects they produce, individuals are thus able to recall or imagine an action by anticipating its effects.[131] As such, these neurons generate an automatic and relatively simple mechanism for mimicking others, providing an ability to imitate actions and learn new behaviours through observation.[132]

More recently, a debate surrounding these cells has emerged to indicate an alternative perspective. This view argues that mirror neurons are not a unique form of neuron but a straightforward associative neuron, such as the ones used in classical and operant conditioning.[133] Consequently, when individuals observe the actions of others, they learn to respond with particular motor programs accordingly, generating an association between the two.[134] For example, experimental evidence indicates that behaviours can be retrained in response to observed actions, as individuals can be trained to unintentionally move their little finger after observing the movement of an index finger.[135] Furthermore, when a stimulus is associated with two or more actions, the context in which behaviours are observed determines which action is subsequently performed.[136] Moreover, additional research on mirror neuron functionality indicates that rather than representing high-level information such as the intended

130. Cook et al., 'Mirror Neurons', 180; Hickok, 'Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans', 1230; Heyes and Catmur, 'What Happened to Mirror Neurons?', 161.
131. Decety and Meyer, 'From Emotion Resonance to Empathic Understanding', 1058–59.
132. Iacoboni, 'Within Each Other', 49.
133. Hickok, *The Myth of Mirror Neurons*, 111–12.
134. Cook et al., 'Mirror Neurons', 181.
135. Hickok, 'Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans', 1236.
136. Cook et al., 'Mirror Neurons', 184.

effects of actions or the goals they aim to produce, mirror neurons instead respond to observations of simple actions such as the style of grip observed by the hand of another.[137] Therefore, from an array of experimental evidence which contradicts or complicates initial assumptions on mirror neurons, our current understanding of these cells has shifted from genetically-determined neurons exhibiting unique behaviours, to common associative ones identifiable throughout the brain. Though researchers and publications may still refer to "mirror neurons," the cells they are referring to are not functionally distinct, behaving in ways which no other neurons do, but rather consist of associative neurons found elsewhere in the brain.

In addition to deliberate, intentional mimicry and imitation, the ability to imitate others occurs on subconscious levels as well. Evidence indicates a physiologically-grounded tendency to mimic others during interactions, where we align our facial, vocal, and postural expressions to those demonstrated by others.[138] Moreover, research suggests mimicry can be established on deeper physiological levels, in which *autonomic* mimicry arises between individuals to synchronize bodily processes like heart rate and breathing rhythms.[139] This synchronization suggests emotional contagion serves as a foundation for empathy and other prosocial behaviours, as mimicry contributes to feelings of similarity and closeness, influencing how we judge or perceive others.[140]

---

137. Heyes and Catmur, 'What Happened to Mirror Neurons?', 161.
138. Hatfield, Rapson, and Le, 'Emotional Contagion and Empathy', 19; van Baaren et al., 'Being Imitated', 32.
139. Carter, Harri, and Porges, 'Neural and Evolutionary Perspectives on Empathy', 170; Prochazkova and Kret, 'Connecting Minds and Sharing Emotions through Mimicry', 100.
140. van Baaren et al., 'Being Imitated', 34–35; Decety and Meyer, 'From Emotion Resonance to Empathic Understanding', 1054.

Given these sorts of brain activity, *simulation theory* seems to identify a physiological foundation for explaining how empathy is generated. Mimicry and imitation provides insight into other's feelings, as we use our own emotional responses as an analogy for the feelings of others.[141] Although some mirroring processes are automatic, others can be intentionally undertaken to recreate an experience in an effort to align oneself with another.[142] That said, there remains the question of how appealing to one's own simulation can fully account for successful acts of empathy when individual differences are apparent. Is it the case that empathy can be fully explained by *simulation theory*? The third section of this chapter demonstrates why this is not exactly the case, and that additional considerations are required for a full explanation of empathy. Before doing so, however, it is necessary to discuss the connection between *simulation theory* and conclusions drawn from evolutionary biology. This relation indicates a continuity between human and animal behaviour to suggest the social significance of this perceptual capacity.

3.2 Empathy as an Evolutionary Outcome

Although initially discovered in macaque monkeys, mirror neuron activity has been subsequently identified in humans, indicating we are not unique in our ability to mimic and imitate others.

141. de Waal and Preston, 'Mammalian Empathy', 503; Iacoboni, 'Imitation, Empathy, and Mirror Neurons', 666–67.
142. Chong and Mattingley, 'Automatic and Controlled Processing within the Mirror Neuron System', 216; Goldman, 'Two Routes to Empathy', 38.

While affective communication and emotional contagion have been identified in birds and fish as well,[143] empathy as an other-oriented perspective appears to exist in several mammalian species.[144] In a model proposed by de Waal and Preston, empathy is conceived as involving three layers, building on emotional contagion and motor mimicry to generate an ability which adopts the perspective of another and potentially provide comfort.[145] These higher forms of empathy have been identified in other apes like chimpanzees and bonobos, but also in elephants, mice, and canines.[146] The evolution of these abilities is likely due to their ability to promote cooperative behaviours by generating a shared perspective for ensuring safety and executing goal-directed behaviours.[147] Therefore, empathy is an advantageous trait which evolved in many social species since it helps individuals learn about emotions of others, as well as a means for organizing and securing group cohesion among individuals.

In their paper titled 'Mammalian empathy: behavioural manifestations and neural basis', de Waal and Preston articulate a process which describes how empathy emerges from its foundation as emotional contagion.[148] Given that I will be relying on this view to demonstrate the merits of an alternative explanation for empathy, I will articulate the steps it proposes. The process begins with emotional contagion, in which an observer adopts the emotion as perceived in another, creating a shared affect which serves as a foundation for empathy to emerge. At this stage, for the observer to consider how another feels, they must control their own feelings

143. Plutchik, 'Evolutionary Bases of Empathy', 44.
144. de Waal, 'Empathy in Primates and Other Mammals', 100.
145. de Waal and Preston, 'Mammalian Empathy', 499.
146. de Waal and Preston, 500.
147. Carter, Harri, and Porges, 'Neural and Evolutionary Perspectives on Empathy', 173; de Waal and Preston, 'Mammalian Empathy', 507.
148. de Waal and Preston, 'Mammalian Empathy', 502.

through emotional regulation, shifting their attention from their own feelings to the perspective of another instead. Without inhibiting or ignoring their own feelings of distress through self-regulation, individuals may become too distracted to provide assistance or respond to another.[149] The understanding derived from this other-oriented consideration can be used to perform actions which reduce feelings of distress, in which the subsequent relief is then shared by the observer through a second occurrence of emotional contagion. As such, the initial distress picked up from another and felt by the observer is eventually reduced, creating an intrinsic reward for acting altruistically. Moreover, evidence from experiments involving chimpanzees and human children suggests this intrinsic reward operates independently from extrinsic rewards, in which the act of providing individuals with rewards for assistance does not influence their helping behaviours.[150] Interestingly, young children seem to demonstrate a reduction in helping behaviours when extrinsically rewarded, though this effect does not apply to verbal rewards like praise, which instead seems to increase the inclination to help others.[151] Thus, intrinsic rewards serve as a motivator for performing actions which assist others; however, to do so, individuals must successfully turn their attention away from their own internal states to consider the emotions of another. In doing so, individuals are able to understand how another is feeling, in which emotional contagion assists with empathy but does not guarantee empathy will occur.

From this discussion of empathy as a process, it becomes apparent that while the discovery of so-called mirror neurons has significantly contributed to our understanding of empathy, *simulation theory* alone is not sufficient for explaining or describing what empathy

149. de Waal and Preston, 501.
150. de Waal and Preston, 502.
151. Warneken and Tomasello, 'Extrinsic Rewards Undermine Altruistic Tendencies in 20-Month-Olds', 1787.

involves. The reason being is that the act of simulating within oneself how another might be feeling does not guarantee that these interpretations are accurate. Although we may interpret behavioural cues as meaning one thing, in reality, they mean something else entirely. This may be due to interpersonal unfamiliarity or due to differences in sociocultural norms. Consequently, there will always remain the question of how well our intuitions or projections accurately represent the inner states of another. Although we are able to imitate behavioural cues and simulate within ourselves the perceived states of others, we may still be incorrect in our inferences and assumptions.[152] There may be instances where we may find ourselves uncovering limitations in our ability to project or infer in certain instances when empathy is required, and appeals to one's own perspective results in an accurate understanding of the other.[153] For example, if the observer has never experienced the death of a loved one, they may find it difficult to know exactly how another is feeling in that circumstance. Other limitations may arise from sociocultural differences, as the meanings of others' behavioural cues can differ from expectations and assumptions, impacting the way these cues are interpreted.[154] In particular, coping with or reacting to a death may be perceived or experienced in a variety of differing ways, based on factors like sociocultural norms, beliefs, and values. For death especially, religious beliefs may play a role in how death is experienced in the loss of a loved one. Therefore, to know how another feels in this situation, the observer must appeal to their knowledge of the other person to know how these feelings are experienced, independently of their own simulations, interpretations, or reactions. Overall, while associative neuron activity, or

152. Ickes, 'Empathic Accuracy', 57.
153. Debes, 'Which Empathy?', 234; Nickerson, Butler, and Carlin, 'Empathy and Knowledge Projection', 51; Stueber, *Rediscovering Empathy*, 136.
154. Burns, 'Empathy, Simulation, and Neuroscience', 214; Stueber, *Rediscovering Empathy*, 196.

so-called "mirror neuron" activity, indicates a mechanism for responding to the emotional states of others, a degree of theory or knowledge is still required to accurately perceive behavioural cues and adopt emotional states accordingly. *Simulation theory* is therefore insufficient for fully explaining empathy, as one's own simulation may not provide enough information to know how another is feeling.

Despite this limitation, the process-view behind *simulation theory* is vital for developing the ability to generate theoretical knowledge of other minds.[155] The domain of developmental psychology is able to account for how associative "mirror" neurons provide a foundation for understanding the minds of others. Experimental evidence suggests that infants as young as 2 weeks are able to imitate specific facial movements observed in adults,[156] demonstrating an innate ability to generate new behaviours through observation. Internal representations of actions and emotions are created by associating the actions of others with how these actions are subjectively experienced.[157] At around 6 weeks of age, infants begin engaging in "proto-conversations" with others through smiling and mimicking facial movements, in which these exchanges continue to develop associations between perceived actions and subjective feelings.[158] From these emotional exchanges, infants develop an ability for *joint attention* where they are able to shift their focus to an object being looked at by another.[159] This ability provides a

155. Decety and Meyer, 'From Emotion Resonance to Empathic Understanding', 1056; Heyes, 'Empathy Is Not in Our Genes', 500.
156. Decety and Meltzoff, 'Empathy, Imitation, and the Social Brain', 59.
157. Decety and Jackson, 'The Functional Architecture of Human Empathy', 75; Decety and Meltzoff, 'Empathy, Imitation, and the Social Brain', 62; Siegler, DeLoache, and Eisenberg, *How Children Develop*, 270.
158. Decety and Meyer, 'From Emotion Resonance to Empathic Understanding', 1057; Zahavi and Rochat, 'Empathy≠sharing', 547.
159. Carpenter et al., 'Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age', 1; Siegler, DeLoache, and Eisenberg, *How Children Develop*, 159.

foundation for intersubjectivity and a shared understanding of the environment and its features. Along with social learning, these emotional exchanges generate emotional bonds between children and caregivers which facilitates further learning as children continue to grow and explore. At around age 1, children are capable of demonstrating a concern for others, and around a year and a half, can offer consoling behaviours to people in distress, indicating empathy is a capacity which emerges early in life.[160] With an acquisition of language and conceptual knowledge, children expand their understanding of how other minds operate in different situations, especially as they continue to socialize with individuals outside of their own family. At about age 3 or 4, children develop an ability to differentiate between their own perspective and the perspectives of other people, as per experiments using *false-belief tasks*.[161] This experimental paradigm involves a child and another person in possession of an object they have placed in a specific location. When the child witnesses the experimenter move the object to a new location, while the other person is not watching, they are then asked where they think the other person believes the object is, in the previous location or the new location.[162] Children who think the other person believes the object is in the new location are thought to *not* possess a theory of mind, as they conflate their own knowledge with that of another. It is debated, however, whether this test is suitable for drawing conclusions on theory of mind.[163] Regardless, it seems our ability to understand the minds of others, to some degree, begins early in life and

160. Siegler, DeLoache, and Eisenberg, *How Children Develop*, 269; Warneken and Tomasello, 'Altruistic Helping in Human Infants and Young Chimpanzees', 1301.
161. Bloom and German, 'Two Reasons to Abandon the False Belief Task as a Test of Theory of Mind', B27; Helming, Strickland, and Jacob, 'Making Sense of Early False-Belief Understanding', 169; Siegler, DeLoache, and Eisenberg, *How Children Develop*, 271.
162. Schidelko et al., 'Why Do Children Who Solve False Belief Tasks Begin to Find True Belief Control Tasks Difficult?', 1–2.
163. Bloom and German, 'Two Reasons to Abandon the False Belief Task as a Test of Theory of Mind', B30.

continues to develop over childhood as individuals encounter new situations, experiences, and types of people. Interacting and socializing with others informs individuals of the ways in which the perspectives of others can differ from their own, especially in those from diverse sociocultural backgrounds. Given that the frontal regions of the brain do not fully mature until late adolescence, it has been suggested a more robust form of empathy takes years to develop.[164] This suggestion is the result of neural-imaging studies which suggest the brain regions used differ between adults and adolescents. Therefore, developing an ability to empathize with others involves many years of learning and practice, beginning with imitation and joint attention to establish a cognitive foundation required to support higher-cognitive functions. Included in this set of capacities is a repertoire of social knowledge established through interacting with other people of varying personalities and backgrounds. Together, it seems as though *simulation theory* and *theory theory* can provide a sufficient account of empathy, since one expands upon the other from a lifetime of learning.

To summarize the main debate in the philosophical and empirical literature on empathy, while *simulation theory* explains how we are able align our own feelings with the perceived emotions of others, *theory theory* describes an approach which considers the way the observer's knowledge facilitates or clarifies our understanding of others' behaviour. Together, these two theories reflect the relationship between conscious and subconscious processes, in which neuronal activity occurs subconsciously and the application of knowledge or theory requires conscious effort. Furthermore, knowledge and theory are acquired from years of experience and practice, arising from interacting socially with others to develop these cognitive abilities.

---

164. Tremblay et al., 'Functional Connectivity Patterns of Trait Empathy Are Associated with Age', 105859.

Activity from neuronal processes, on the other hand, enables an ability to imitate the actions of others like facial reactions and behavioural cues. These subconscious processes facilitate acts of empathy by aligning emotional states, where individuals may also appeal to theory or knowledge to augment perceptual information to know how another is feeling.

Although *simulation theory* and *theory theory* contribute important perspectives to establishing a better understanding of what empathy involves, there is still the question of why or how an act of empathy may be accurate or inaccurate. An adequate theory of empathy must also account for cultural or individual variables, since the ways that individuals see and experience the world plays a role in their emotional and cognitive reactions to it, including the states of other people. To accomplish this, an explanation is required to understand how an act of empathy fails or succeeds. Explanations of this sort can be generated by investigating the specific, contextual interaction between two individuals.

3.3 Empathy as Context-Dependent

This section outlines *interaction theory* as an alternative to *theory theory* or *simulation theory* and argues that context serves as a significant variable for determining how or when acts of empathy are accurate. Given the process-model of empathy offered by researchers de Waal and Preston outlined in the previous section, it follows that we should conceive of empathy as a process involving an interaction between two or more individuals. If an individual is inaccurate in their ability to know what another is feeling, it may be due to a number of reasons, one of

which is a lack of familiarity or understanding of the other. By closely examining the contextual components involved, the significance of the factors involved becomes more apparent. The literature on empathy often mentions "the situation" or "the feeling," in which these details are considered in light of the individuals involved, the setting or location, as well as other factors such as initial stimuli, related information, and additional behavioural cues. To know how to respond in a given context, cognitive skills must be developed through repeated experiences and ongoing practice between a variety of different individuals and perspectives. Therefore, context plays a significant role in determining how and when acts of empathy are accurate or inaccurate.

As proposed by Shaun Gallagher, *interaction theory* provides an alternative view on how we perceive and encounter others. He describes it as adopting a second-person perspective rather than a first-person perspective, an internal representation, or a third-person perspective, an observation.[165] Instead, the second-person perspective is a subjective point of view oriented outward rather than inward and toward another person in an attempt to understand, communicate, or interact with them.[166] Rather than focusing on one's own experiences in a first-person point of view, a second-person perspective is directed outward and towards another to establish an intersubjective understanding of the world for the sake of communication.[167] In contrast, the third-person perspective aims to see beyond a subjective point of view to understand or describe events in a manner independent of how they appear within the first-person perspective.

---

165. Gallagher, 'Inference or Interaction', 164; Gallagher, 'Neurons, Neonates and Narrative', 173–74.
166. Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 90–91.
167. Gallagher, 'Neurons, Neonates and Narrative', 174; Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 99.

To better illustrate the differences between the first-person, second-person, and third-person perspectives, an example regarding the truth-values of a sentence can be examined. The fictional scenario involves two individuals living in Toronto who encounter a book written in Simplified Chinese. The statement, "this text is illegible" may be true for one person while simultaneously being false to another, provided they understand the language. From the first-person perspective, the truth of the statement depends on the individual uttering it. For the Torontonian who only speaks English, in this case named Bill, the statement will be true. Bill has a friend named Li who recently moved from Beijing. When Bill considers this statement from Li's perspective, the second-person perspective, the statement is false as Li can read and write Simplified Chinese. From a third-person perspective, this statement is false, given that individuals exist in the world who can read and write Simplified Chinese. Overall, the truth of the statement depends on the person is uttering it, however, from a third-person perspective, its truth-value depends on states of affairs which are independent of the speakers asserting it.

Returning now to Gallagher's *interaction theory*, it provides a preferable alternative precisely because it is oriented in the second-person perspective, inherently taking into account the viewpoint of another person rather than one's own. When perspectives differ, individuals are still able to understand aspects of the world from an alternative point of view because they have oriented themselves outward and toward the other. This enables them to acknowledge or consider how someone else experiences a particular situation even when individuals differ from one another.

Although we may appeal to simulations or theories to understand others, Gallagher believes this rarely occurs because, in most cases, we are able to able to understand others

without having to simulate their internal states or appeal to a theory.[168] Instead, as we engage with others in social or pragmatic activities, we are normally able to immediately perceive the meanings or intentions behind the behaviours of others, a capacity which is developed over childhood.[169] Gallagher suggests that mirror neuron activity is a part of a broad network of neuronal processes which underlie a form of "intersubjective enactive perception."[170] By this he means a "non-articulated immediate perception of the other person's intentional actions" which is active and embodied, rather than a passive processing of sensory input data detected in the environment.[171] Rather than thinking or reasoning about the intentions of others, we directly perceive and understand them through another's behaviours and actions.[172] The bodily movements, facial expressions, gestures, and postures of others signal their inner states which we immediately comprehend, an ability which manifests early in life.[173] Situational context also plays a significant role in understanding the intentions of others, as the circumstances and setting of the interaction structures an individual's actions.[174] Furthermore, one's perceptions are also informed by previous interactions with others, cultural norms and social practices, and one's own "habitual ways of understanding."[175] Within this second-person perspective, a shared world emerges as individuals interact, where both parties respond intuitively from an immediate comprehension of the other.

---

168. Gallagher, 'Neurons, Neonates and Narrative', 174.
169. Gallagher, 'Inference or Interaction', 165; Gallagher, 'Neurons, Neonates and Narrative', 175; Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 86.
170. Gallagher, 'Neurons, Neonates and Narrative', 181.
171. Gallagher, 181.
172. Gallagher, 'Direct Perception in the Intersubjective Context', 538.
173. Gallagher, 'Neurons, Neonates and Narrative', 181–82.
174. Gallagher, 'Direct Perception in the Intersubjective Context', 540.
175. Gallagher, 540.

Gallagher appeals to evidence from developmental psychology to explain how we are able to directly perceive the intentions of others. Experimental evidence indicates that infants are able to follow the gaze of another, suggesting an inherent recognition and understanding of this behaviour as meaningful or goal directed.[176] Infants usually demonstrate a keen interest in activity or actions arising from humans over non-living things, as well as an inherent interest in faces over objects.[177] As such, it suggests they can identify subjects and beings like themselves, and possess a pre-reflective sense of self as a subject.[178] Their interest in others facilitates learning about the meanings of various aspects of the world, the same world in which they share and interact in with others.[179] As children mature, their understanding of the world expands, along with the types of interactions they have with others, introducing them to new situations and unfamiliar individuals. To correctly interpret behavioural cues exhibited by another requires individuals to draw on a "communicative and narrative competency" or conceptual framework which is learned through childhood within a particular sociocultural context and used to explain the meanings of behaviours.[180] This competency shapes the ways in which we understand the behaviours of others, as it provides contextual information that influences our perceptions of others.

Interaction theory is thus the preferable theory for explaining empathy because it accounts for contextual variables. It describes an intersubjective dynamic situated within a particular setting with individuals who are oriented in the second-person perspective. Here,

176. Gallagher, 'Neurons, Neonates and Narrative', 182; Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 88–89.
177. Gallagher, 'Inference or Interaction', 165; Siegler, DeLoache, and Eisenberg, *How Children Develop*, 268–69.
178. Gallagher, 'Neurons, Neonates and Narrative', 182.
179. Gallagher, 183.
180. Gallagher, 186.

understanding and communication is facilitated by a perceptual ability to immediately comprehend the intentions and feelings of another. In general, one's direct perception is often sufficient for understanding others, however, in cases when ambiguity does arise, other aspects of the situation at hand can provide information or insight. This occurs without the need for explicit reasoning or simulating within oneself what another is experiencing. *Interaction theory* describes a rapid and holistic awareness which is not intentionally undertaken by individuals. Rather, the exchange unfolds in an intuitive manner, as individuals seek to understand one another and respond appropriately. To Gallagher, "what we call social cognition is often nothing more than social interaction,"[181] a position which differs from *theory theory* and *simulation theory*. For these two theories, understanding the minds of others involves either an active and deliberate consideration, or a passive contagion which affects an individual's internal state. These theories frame empathy in terms of individuals rather than a dynamic between people, and as such, cannot account for how or why acts of empathy succeed or fail. *Interaction theory* circumvents this issue by remaining situated within a particular context and within the second-person perspective. Although our perceptual abilities are generally sufficient for understanding the feelings and intentions of others, in cases in which they are not, individuals can appeal to contextual information to improve their understanding. It must be noted that this process occurs automatically and intuitively, and is not a form of deliberate reasoning. The adjustments made during social interactions occur subconsciously, similar to regaining one's balance upon slipping on ice. As social beings, we are inherently invested in understanding the other person within

181. Gallagher, 'Direct Perception in the Intersubjective Context', 540.

these interactions, and though errors can be made, we are able to make adjustments according to a variety of contextual cues.

In conclusion, empathy emerges as individuals respond to their direct perceptions of others as they occur within social interactions. While this can involve a degree of emotional contagion, framing empathy solely in terms of *simulation theory* reduces empathy to experiential states existing within an individual. Instead, we ought to think of empathy as emergent responses to the behaviours and expressions of others during dynamic social interactions. Our ability to comprehend and respond appropriately to these behaviours may be in part due to associative "mirror" neurons, however, reducing the capacity for empathy to the activity of these cells mischaracterizes how it emerges. Rather, empathy manifests as a response from within a second-person perspective, one which responds according to intersubjective dynamics. Individuals are situated in wider social and environmental contexts which include factors that influence the interaction as it unfolds, and to account for how or why an act of empathy succeeds or fails, these factors must be considered.

Consequently, *interaction theory* provides a beneficial framework for considering the requirements we have placed upon social robots, as contextual factors significantly impact interpretations of human behaviour. It will be important for social robots to establish joint attention and reciprocity or turn-taking when communicating with humans for it to feel as if one were interacting with another human and not an artificial agent. To act with empathy, robots will need to correctly interpret the behaviours of people and the affective states behind behavioural cues, subsequently adopting these feelings in order to classify these acts as empathetic rather than sympathetic. For a robot to accomplish this, however, it must be able to appropriately

represent human emotions in its own cognitive architecture. To best model the behaviours and capacities of humans, it must be able to use information available within a particular social interaction to adjust its behaviours in an appropriate manner. The next chapter challenges this notion by providing a brief history of artificial intelligence and robotics to determine the current state of research and development, and whether this trajectory will one day produce robot empathy. The chapter concludes by arguing that even if robots were to mimic empathy-related behaviours, social robots are not capable of real empathy as it manifests in humans and some mammalian species. In virtue of their design, current robotic solutions are unable to adopt a human point of view because it is unable to understand the emotional cues and language of others, since it unable to experience emotional or affective states from its own lived perspective. Given the discrepancy between how information on affect manifests, empathy-like behaviour exhibited by social robots is merely an illusion.

4 The Limitations of Developmental Robotics

This chapter outlines the history of artificial intelligence (AI) to provide context for the development of social robots as inspired by biological processes. Given the successful outcomes generated by neural network architectures, researchers have designed robots which replicate human development. One robot in particular named iCub is discussed in the second part of this chapter to demonstrate how sensorimotor learning scaffolds language acquisition. While iCub appears to understand what words mean, the final section investigates emotion to argue why this is not the case. Though embodied, iCub is not adequately modelled on human physiology and neurology to support an understanding of concepts related to affect.

4.1 Progress in Artificial Intelligence

Though the term was coined in 1956, 'artificial intelligence' as a concept dates back to ancient Greece from Homer's *The Iliad*.[182] Philosophers since then, including Aristotle and Descartes, have discussed the possibility of behaviours arising from mechanical agents.[183] Similarly, the computer is an idea dating back to calculators crafted by Leibniz and Pascal and the Analytic

182. Adami, 'A Brief History of Artificial Intelligence Research', 133; McCorduck, *Machines Who Think*, 4; Crevier, *AI*, 2.
183. Ekmekci and Arda, 'History of Artificial Intelligence', 1–2; Buchanan, 'A (Very) Brief History of Artificial Intelligence', 53.

Engine by Babbage.[184] There is a long history to the philosophical concept of AI which will not be discussed here, as it is beyond the scope of our discussion of social robots. Instead, the start of our investigation into AI's history will begin in the early to mid 20th century with the invention of the computer. From the beginning of the field of *artificial intelligence*, there were two approaches to designing sophisticated machines, consisting of *symbolic reasoning* and *connectionism*.[185] While symbolic reasoning relies on explicit instructions and concepts, connectionist approaches use artificial neural networks. Subsequent progress in AI has been described in terms of cycles consisting of "winters" and "springs" or "Golden Ages" which are characterized by stops and starts in research and development.[186] Though many have remained optimistic about AI development and its potential capacities, my aim is to demonstrate significant limitations in our attempts at replicating human intelligence. To accomplish this, a brief historical account of AI is required. During the 1950's and 1960's, the dominant approach to building AIs involved symbolic reasoning, however, in the 1980's, a renewed interest in neural networks shifted research and development in AI towards connectionism. By the early 1990's, a growing interest in robotics and *embodiment* aimed to ground meaning by replicating human learning. Robots designed to learn about the world from experience are theoretically able to subsequently learn how these elements are represented in symbols and language. As AI

184. Lungarella et al., 'AI in the 21st Century – With Historical Reflections', 1; Ekmekci and Arda, 'History of Artificial Intelligence', 3.
185. Adami, 'A Brief History of Artificial Intelligence Research', 133; Steels, 'Fifty Years of AI', 23.
186. Haenlein and Kaplan, 'A Brief History of Artificial Intelligence', 6; Lee, *Artificial Intelligence in Daily Life*, 25–26.

developed over history, a trend toward the modelling of biological systems to generate intelligent behaviour emerged, in which *embodiment* introduced a phenomenological perspective.

The precursors to the idea of "artificial intelligence" can be identified in a number of technological developments in the 1930's and 1940's. In 1936, Alan Turing created the Turing Machine, a theoretical machine capable of performing calculations based on a set of rules by reading and writing information to a tape of infinite length.[187] That same year, the first programmable computer was created by the German engineer Konrad Zuse.[188] His machine performed calculations using binary numbers by reading instructions from a punched tape, controlled by a computing unit and capable of storing information.[189] Six years later, Isaac Asimov published *Runaround* with its "Three Laws of Robotics" in 1942, stipulating that robots must not harm humans, nor allow them to be harmed, along with protecting its own existence provided it does not conflict with the previous laws.[190] The following year, the first mathematical model of the biological neuron was presented by Warren McCulloch and Walter Pitts,[191] demonstrating an all-or-nothing firing pattern which can then be used to generate logical operations.[192] In 1945, John von Neumann released a report on a computer system called EDVAC, originally developed by Presper Eckert and John Mauchly at the Moore School of Electrical Engineering at the University of Pennsylvania.[193] Despite being created by Eckert and

187. Haikonen, *The Cognitive Approach to Conscious Machines*, 15–16.
188. Rojas, 'Konrad Zuse's Legacy', 5.
189. Rojas, 5–6.
190. Haenlein and Kaplan, 'A Brief History of Artificial Intelligence', 6.
191. Ekmekci and Arda, 'History of Artificial Intelligence', 5; Lee, *Artificial Intelligence in Daily Life*, 22; Franklin, 'History, Motivations, and Core Themes', 16.
192. Ekmekci and Arda, 'History of Artificial Intelligence', 5; McCorduck, *Machines Who Think*, 56.
193. Williams, 'The Origins, Uses, and Fate of the EDVAC', 22.

Mauchly, von Neumann would later receive credit for the computer's invention, as his was the only name listed on the report detailing EDVAC's functionality.[194] The so-called "von Neumann architecture" is still used in computers today,[195] consisting of a control unit, an arithmetic/logic unit, a memory unit, as well as input and output devices such as a keyboard and screen.[196]

During this period of time, computers were very large and expensive, often referred to as "giant brains."[197] Their use of vacuum tubes rather than transistors and microprocessors, as these would not be invented until later that century, meant these machines were much bigger and costlier than modern personal computers.[198] In 1948, the mathematician Norbert Wiener's book *Cybernetics: Or Control and Communication in the Animal and the Machine* launched the "cybernetics movement" with Warren McCulloch, Walter Pitts, and John Von Neuman.[199] The 1949 publication *Giant Brains: Or Machines That Think* by computer scientist Edmund Berkeley compares machines to human brains by describing them as "hardware and wire instead of flesh and nerves"[200] and because they could process information similarly to a brain, "a machine, therefore, can think."[201] In 1950, Alan Turing's paper *Computing Machinery and Intelligence* asked whether computers think, responding by stating that a computer can be considered to think if it can pass the "Imitation Game."[202] The same year, Claude Shannon, the father of information

194. Williams, 23.
195. McCorduck, *Machines Who Think*, 78; Rojas and Hashagen, '"Nothing New Since von Neumann"', 195.
196. Haikonen, *The Cognitive Approach to Conscious Machines*, 17; Lee, *Artificial Intelligence in Daily Life*, 43.
197. Buchanan, 'A (Very) Brief History of Artificial Intelligence', 54; McCorduck, *Machines Who Think*, 78.
198. O'Regan, *A Brief History of Computing*, 27–28.
199. Adami, 'A Brief History of Artificial Intelligence Research', 133; Russell and Norvig, *Artificial Intelligence*, 15.
200. Berkeley, *Giant Brains, or Machines That Think*, 1.
201. Berkeley, 5.
202. Franklin, 'History, Motivations, and Core Themes', 17; Turing, 'Computing Machinery and Intelligence.', 434.

theory, published *Programming a Computer for Playing Chess*,[203] and Marvin Minsky created the first neural network of a rat running a maze.[204] Although interest in thinking machines was increasing, the term "artificial intelligence" would not be created for another four years.

A series of conferences would lead to the creation of the term and field of *artificial intelligence*. In 1948, the conference *Hixon Symposium on Cerebral Mechanisms in Behaviour* held in Pasadena, California, featured papers by Warren McCulloch, John von Neumann, and psychologist Karl Lashley whose presentation would lay the foundation for cognitive science.[205] These presentations investigated how the nervous system and brain compares to computer systems,[206] and in attendance was computer scientist John McCarthy,[207] who would coin the term *artificial intelligence* eight years later. The 1955 conference *Session on Machine Learning* held in Los Angeles, California, featured papers on neural networks by Belmont Farley and Wesley Clark, computational image processing by Gerald Dinneen and Oliver Selfridge, and a program for playing chess presented by Allen Newell.[208] The following year, John McCarthy would include the term *artificial intelligence* in a proposal written for the *Dartmouth Summer Research Project*, a workshop held in Hanover, New Hampshire.[209] The proposal for funding described a study to be conducted of simulating learning and intelligence in machines, with coauthors

203. Ekmekci and Arda, 'History of Artificial Intelligence', 6; Shannon, 'Programming a Computer for Playing Chess', 256.
204. Franklin, 'History, Motivations, and Core Themes', 19; Russell and Norvig, *Artificial Intelligence*, 16.
205. Nilsson, *The Quest for Artificial Intelligence*, 49–50.
206. Nilsson, 49.
207. Nilsson, 52.
208. Nilsson, 50–51; Pitts, 'Comments on Session on Learning Machines', 110.
209. McCorduck et al., 'History of Artificial Intelligence', 953; Moor, 'The Dartmouth College Artificial Intelligence Conference', 87.

including Nathaniel Rochester, Marvin Minsky, and Claude Shannon.[210] The proposed "study" became a workshop which, in addition to the proposal's authors, included Arthur Samuel, Oliver Selfridge, Allen Newell, Herbert Simon and Cliff Shaw.[211] Here, Newell, Simon, and Shaw presented the Logic Theorist program capable of proving 38 of 52 theorems from Whitehead and Russell's *Principia Mathematica*.[212] In 1958, the conference *Mechanisation of Thought Processes* was held in Teddington, Middlesex, England, and featured papers by Minsky, McCarthy, Selfridge, and McCulloch.[213] Here, Selfridge presented a paper on symbol-manipulation for pattern recognition and McCulloch presented a paper on neural network signal switching.[214] These conferences would stir up global interest in the idea of developing systems capable of replicating or surpassing human mental abilities.[215]

From the outset, two approaches to AI architecture emerged: *symbolic reasoning* and *connectionism*. In the beginning, the dominant approach was symbolic AI, especially heuristic search and knowledge representation.[216] Heuristic search involves the use of a general rule to guide the system toward producing a solution despite not guaranteeing a solution, such as "protecting the queen" in a game of chess.[217] To avoid becoming mired by irrelevant solutions from the use of heuristics, researchers developed programs which aimed at representing human knowledge to assist with producing outputs. Data structures used by the system would reflect or

210. Ekmekci and Arda, 'History of Artificial Intelligence', 7; McCorduck et al., 'History of Artificial Intelligence', 953.
211. Lee, *Artificial Intelligence in Daily Life*, 23; Nilsson, *The Quest for Artificial Intelligence*, 53.
212. Gugerty, 'Newell and Simon's Logic Theorist', 880; Nilsson, *The Quest for Artificial Intelligence*, 54.
213. Flasiński, 'History of Artificial Intelligence', 4; Nilsson, *The Quest for Artificial Intelligence*, 56.
214. Davies, 'Mechanization of Thought Processes', 225; Nilsson, *The Quest for Artificial Intelligence*, 57.
215. Lungarella et al., 'AI in the 21st Century – With Historical Reflections', 2; Steels, 'Fifty Years of AI', 20..
216. Franklin, 'History, Motivations, and Core Themes', 24–25; Steels, 'Fifty Years of AI', 22.
217. Boden, 'GOFAI', 90.

replicate human knowledge and algorithms would use these representations to perform inferences to produce an output.[218] Notable examples include McCarthy's Advice Taker which, given a set of propositions about the world, would deduce conclusions about actions to be performed.[219] Similarly, the General Problem Solver created by Newell, Simon, and Shaw in 1957, used heuristics from the Logic Theorist to generate a system capable of breaking problems down into sub-problems.[220] From there, each sub-problem could be represented symbolically and compared to the representations stored in the system, in which the difference between these representations indicated an operation to minimize this difference.[221] Alternatively, researchers like McCulloch and Selfridge were interested in replicating neural networks to mimic how the brain works.[222] In 1958, Frank Rosenblatt introduced the *perceptron*, a single-layer neural network based on the mathematical model created by McCulloch and Pitts in 1943.[223] In a neural network, a representation is generated from the activated patterns of neurons, in which processing is achieved by propagating these activations to the nodes of the network through their interconnections. This propagation is governed by *weights* or numerical values between pairs of nodes, and learning occurs through the change of the value of these weights based on the accuracy of the output.[224] Rather than using human-based knowledge and representations, neural

218. Frankish and Ramsey, *The Cambridge Handbook of Artificial Intelligence*, 339; Flasiński, 'History of Artificial Intelligence', 5; Franklin, 'History, Motivations, and Core Themes', 20.
219. Davies, 'Mechanization of Thought Processes', 225; McCorduck, *Machines Who Think*, 254; Russell and Norvig, *Artificial Intelligence*, 19.
220. Dreyfus, *What Computers Can't Do*, xxiv; Lee, *Artificial Intelligence in Daily Life*, 23; Nilsson, *The Quest for Artificial Intelligence*, 88.
221. McCorduck, *Machines Who Think*, 247; Nilsson, *The Quest for Artificial Intelligence*, 88–89.
222. Davies, 'Mechanization of Thought Processes', 225; Nilsson, *The Quest for Artificial Intelligence*, 64; Sejnowski, *The Deep Learning Revolution*, 2018, 39.
223. Lefkowitz, 'Professor's Perceptron Paved the Way for AI – 60 Years Too Soon'; Sun, 'Connectionism and Neural Networks', 110.
224. Sun, 'Connectionism and Neural Networks', 109.

networks are sub-symbolic and can learn rapidly about incoming information by updating how neurons in the network fire.[225] While the two main branches of AI research, namely the symbolic and connectionist approaches, view problems quite differently, both are still used today and have even been combined to create hybrid systems.[226]

The mid 1950's through to the early 1970's is characterized as the "First Golden Age" of AI given the successes of early programs and systems to perform sophisticated tasks.[227] With large sums of funding sourced from private companies and governmental agencies in the United States like the Defence Department's Advanced Research Projects Agency, subsequently known as DARPA.[228] Contributions provided by individuals from the Massachusetts Institute of Technology (MIT), Carnegie Mellon University, and Stanford assisted to the flourishing of AI over this period of time.[229] In 1958, Minsky and McCarthy established the Artificial Intelligence Laboratory at MIT, where McCarthy would create the programming language LISP.[230] Improvements in computing hardware would also contribute to AI research, as increases in processing speeds and memory capacity improved efficiency.[231] Early projects in robotics began during this period as well,[232] with work on the Stanford Cart project beginning in 1960. This remote-controlled robot was operated by a computer using television camera signals and was

225. Ekmekci and Arda, 'History of Artificial Intelligence', 10; Flasiński, 'History of Artificial Intelligence', 10.
226. Adami, 'A Brief History of Artificial Intelligence Research', 133; Franklin, 'History, Motivations, and Core Themes', 31; Sun, 'Connectionism and Neural Networks', 119.
227. Lee, *Artificial Intelligence in Daily Life*, 22.
228. Crevier, *AI*, 65; Lee, *Artificial Intelligence in Daily Life*, 24; McCorduck, *Machines Who Think*, 131.
229. Buchanan, 'A (Very) Brief History of Artificial Intelligence', 59; Ekmekci and Arda, 'History of Artificial Intelligence', 10; McCorduck, *Machines Who Think*, 131.
230. Kaynak, 'The Golden Age of Artificial Intelligence', 2; Russell and Norvig, *Artificial Intelligence*, 19.
231. Buchanan, 'A (Very) Brief History of Artificial Intelligence', 56.
232. McCorduck, *Machines Who Think*, 261.

able to navigate around obstacles.[233] In 1961, General Motors introduced the Unimate robot to its factories to unload car parts from die-casting machines capable of operating under high-temperatures.[234] Development of the first mobile robot for performing tasks began in 1966, named Shakey from its tendency to wobble when coming to an abrupt stop.[235] The first chatbots emerged during this period as well, as in 1965, the chat program ELIZA by Joseph Weizenbaum is introduced, capturing the imaginations of its users to the surprise of its developers.[236] Modelled on Rogerian psychotherapy, the inquisitive nature of ELIZA generated in its users a sense of interpersonal connection as a consequence of its style of questioning.[237] Because the system did not understand the words it was using when responding to human input, it would ask users to clarify ideas or concepts as a way of overcoming limitations in interpretation and production.[238] As such, ELIZA generated an illusion of knowledge and understanding from the user's background assumptions and inferences from the interpretation of its output messages.[239] ELIZA would become famous for its ability to engage users; however, this accomplishment was more a product of human psychology rather than technical achievement.[240]

233. Moravec, *Robot*, 15–16; Bhaumik, *From AI to Robotics*, 12.
234. Hudson, *The Robot Revolution*, 22; Moravec, *Mind Children*, 10–11.
235. Husbands, 'Robotics', 273; Nilsson, *The Quest for Artificial Intelligence*, 164.
236. Weizenbaum, 'ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine', 36; Weizenbaum, *Computer Power and Human Reason*, 6.
237. Crevier, *AI*, 134; Weizenbaum, 'ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine', 42.
238. Weizenbaum, 'ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine', 43.
239. McCorduck, *Machines Who Think*, 296; Weizenbaum, 'ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine', 42–43.
240. Crevier, *AI*, 133; Wooldridge, *A Brief History of Artificial Intelligence*, 25.

The abundance of innovation and excitement generated a lot of optimism during this time.[241] In 1957, Herbert Simon predicted AI would beat a world champion at chess in the next 10 years.[242] In actuality, it would not be until 1997, four decades later, when World Champion and Grandmaster Gary Kasparov was defeated by Deep Blue.[243] In 1967, Marvin Minsky stated that "Within a generation... the problems of creating 'artificial intelligence' will be substantially solved."[244] Although much progress has been made since then, this prediction has yet to be realized.

Enthusiasm dwindled as expectations could not be fulfilled, and the "First AI Winter" would arrive in 1969 when Marvin Minsky and Seymour Papert published their book *Perceptrons* which demonstrated the limitations of Rosenblatt's single-layer neural network.[245] One limitation included the *exclusive-or problem*, in which the neural network could not recognize an entity which contains one of two properties but not both.[246] It was concluded that neural networks were not a promising direction for research, and as a result, funding for researching AI in general would cease for about a decade.[247] Furthermore, in 1972, Hubert Dreyfus published *What Computers Can't Do* which criticized the prospect of symbolic AI to replicate human reasoning. Symbolic-reasoning can only process facts provided by people, while humans are beings who create themselves and "the world of facts in the process of living in the

241. Wooldridge, *A Brief History of Artificial Intelligence*, 35.
242. Dreyfus, *What Computers Can't Do*, xxix; Russell and Norvig, *Artificial Intelligence*, 21.
243. Kaynak, 'The Golden Age of Artificial Intelligence', 3.
244. Minsky, *Computation*, 2.
245. Flasiński, 'History of Artificial Intelligence', 10; Olazaran, 'A Sociological Study of the Official History of the Perceptrons Controversy', 629; Sejnowski, *The Deep Learning Revolution*, 2018, 47.
246. Olazaran, 'A Sociological Study of the Official History of the Perceptrons Controversy', 625–26; Haikonen, *Consciousness and Robot Sentience*, 85.
247. O'Connor, 'Undercover Algorithm', 10; Lee, *Artificial Intelligence in Daily Life*, 24.

world."[248] As such, "there is no reason to suppose that a world organized in terms of these fundamental human capacities should be accessible by any other means."[249] Additional criticism against symbolic AI emerged in 1973 in a report by Prof. Sir James Lighthill which stated that while impressive, research outcomes had failed to live up to early optimistic expectations.[250] This report suggested that the largest issue with expert systems was their failure to acknowledge the *combinatorial explosion* of symbolic approaches, where the act of programming the required concepts or knowledge generates a quantity of code which cannot be supported by existing hardware.[251] Moreover, DARPA's Speech Understanding Research project had not seen the successes researchers had hoped for,[252] causing the agency to cut funding for this project in 1974.[253] Together, these criticisms and obstacles negatively impacted public support and funding for AI projects, causing research to stall until the 1980s.

While AI in the United States was witnessing a stagnation, organizations around the world were still working on different AI systems and making progress.[254] In 1972, Teuvo Kohonen at the Helsinki University of Technology in Finland published *Correlation Matrix Memories*[255] and would continue to develop approaches to building neural networks throughout the decade before focusing on *self-organizing maps* (SOMs) in the 1980's.[256] The year 1972 also

248. Dreyfus, *What Computers Can't Do*, 202–3.
249. Dreyfus, 203.
250. Lee, *Artificial Intelligence in Daily Life*, 24; Lighthill, 'Artificial Intelligence: A General Survey'; Russell and Norvig, *Artificial Intelligence*, 22.
251. Wooldridge, *A Brief History of Artificial Intelligence*, 60.
252. Crevier, *AI*, 116; Wooldridge, *A Brief History of Artificial Intelligence*, 62.
253. Crevier, *AI*, 117.
254. Franklin, 'History, Motivations, and Core Themes', 20.
255. Flasiński, 'History of Artificial Intelligence', 10; Kohonen, 'Correlation Matrix Memories', 353.
256. Rosenfeld, 'Teuvo Kohonen', 154.

witnessed the first anthropomorphic robot, named WABOT-1, completed in Japan at Waseda University, where it could walk, carry objects, and communicate with humans.[257] Work on self-organizing neural networks would continue to be pursued by Christoph von der Malsburg in Germany,[258] and in 1973, he would publish the article *Self-Organization of Orientation Sensitive Cells in the Striate Cortex.*[259] Inspired by Hubel and Wiesel's discovery of neuronal sensitivity to specific line orientations in the cortices of higher vertebrates, this work presented a method for training neural networks in which single neurons become capable of responding to specific orientations of line segments, such as vertical or horizontal lines.[260] In 1975, Japanese researcher Kunihiko Fukushima created a multi-layered neural network called Cognitron which was capable of self-organizing without the need for feedback or supervision.[261] So although AI research and development seemed to stagnate during the 1970's in the United States, researchers elsewhere continued to contribute to the growing body of research.

Renewed interest in AI research began in 1980 and spirits remained high for the next 7 years. The period of time known as the "Second Golden Age" was characterized by advances in the design of neural networks.[262] John Hopfield created a new type of learning neural network in 1982, one which recurrently passes information repeatedly through the same network node for further refining.[263] Geoffrey Hinton and Terry Sejnowski published a paper on a learning

257. Hudson, *The Robot Revolution*, 23; Lee, *Artificial Intelligence in Daily Life*, 25.
258. Franklin, 'History, Motivations, and Core Themes', 20.
259. von der Malsburg, 'Self-Organization of Orientation Sensitive Cells in the Striate Cortex', 85.
260. von der Malsburg, 85.
261. Flasiński, 'History of Artificial Intelligence', 10; Fukushima, 'Cognitron', 122–23.
262. Ekmekci and Arda, 'History of Artificial Intelligence', 13; Lee, *Artificial Intelligence in Daily Life*, 25.
263. Lee, *Artificial Intelligence in Daily Life*, 25.

algorithm for multi-layered neural networks in 1985,[264] and David Rumelhart and James McClelland published *Parallel Distributing Processing* in 1986.[265] The latter presented a method for overcoming the *exclusive-or problem* identified decades prior, in which multi-layered neural networks are trained by the technique called *back propagation*.[266] There were also large successful projects in symbolic AI during this period, with expert systems created for locating molybdenum mineral ore deposits, forecasting thunderstorms, and identifying points of trouble in telephone networks and recommending corrective measures.[267] The Japanese government also contributed to this Second Golden Age by launching the Fifth Generation computer project which aimed to generate humanlike communication and reasoning.[268]

By this point, three main approaches to creating artificial neural networks could be identified: supervised, unsupervised, and reinforcement learning.[269] All of these approaches are modelled on *associative learning*, in which a connection between two stimuli is established when presented in short succession. Associative learning will be discussed in further detail in the next chapter; however, introducing this idea here provides further background for the functionality of neural networks. In associative learning, neurons which "fire together, wire together"[270] to establish an association between two stimuli, such as an image and a word. In supervised learning, the association generated relies on feedback regarding the correct answer or the degree

264. Sejnowski, *The Deep Learning Revolution*, 2018, 79.
265. Franklin, 'History, Motivations, and Core Themes', 21; Russell and Norvig, *Artificial Intelligence*, 24.
266. Franklin, 'History, Motivations, and Core Themes', 21; Wooldridge, *A Brief History of Artificial Intelligence*, 117–18.
267. Crevier, *AI*, 199.
268. Kaynak, 'The Golden Age of Artificial Intelligence', 3; Lee, *Artificial Intelligence in Daily Life*, 26.
269. Russell and Norvig, *Artificial Intelligence*, 694–95.
270. Haikonen, *Consciousness and Robot Sentience*, 90.

of error between the program's output and the ideal or goal state, altering the functionality in response.[271] Without this feedback, unsupervised learning uncovers commonalities or structures from incoming data based on intrinsic features and patterns.[272] Feedback can also be provided through the use of rewards, and reinforcement learning is an approach to machine learning which provides a reward in response to correct actions taken, thus increasing the likelihood this action occurs in the future.[273] While supervised learning mirrors human pedagogy,[274] unsupervised learning reflects the functioning of perceptual capacities in humans and animals,[275] and reinforcement learning simulates behaviour training in a wide range of animals.[276] From neural networks to learning paradigms, approaches to artificial intelligence modelling biological processes illustrate promising directions for generating a range of key capacities and behaviours.

The "Second AI Winter" occurred from 1987 to 1993.[277] In 1987, there was a collapse of the market for specialized AI computer hardware.[278] Desktop computers were catching up with these specialized machines and as such, were no longer in demand. Moreover, a few revelations about expert systems like XCON, which was created to much acclaim during the previous Golden Age,[279] demonstrated these systems were expensive to maintain, difficult to update, unable to learn, and prone to mistakes.[280] Additionally, results from Japan's Fifth Generation

---

271. Hastie, Tibshirani, and Friedman, 'Unsupervised Learning', 485; Russell and Norvig, *Artificial Intelligence*, 757.
272. Dike et al., 'Unsupervised Learning Based On Artificial Neural Network', 324; Watson, 'On the Philosophy of Unsupervised Learning', 3.
273. Russell and Norvig, *Artificial Intelligence*, 830.
274. Hastie, Tibshirani, and Friedman, 'Unsupervised Learning', 485.
275. Russell and Norvig, *Artificial Intelligence*, 985.
276. Russell and Norvig, 830.
277. Lee, *Artificial Intelligence in Daily Life*, 26; McCorduck, *Machines Who Think*, 532.
278. Crevier, *AI*, 210.
279. Crevier, 197; Wooldridge, *A Brief History of Artificial Intelligence*, 69.
280. Crevier, *AI*, 204; Lee, *Artificial Intelligence in Daily Life*, 26.

project had been underwhelming as expectations were higher than what had actually been achieved.[281] As a result, funding was once again withdrawn by governmental agencies in the United States and Japan.[282] It was during this Winter, in 1988, that Hans Moravec introduced his paradox: he noted that some problems are easy for computers and difficult for humans, like proving theorems, while others appear to be rather difficult for AIs but easy for humans, like recognizing a face.[283] Arguably, this fact arises from the physical and functional differences between animals and computers, a topic which will be explored further in subsequent chapters.

Despite these setbacks, this period marked the beginning of the application of embodied cognition to computers and AIs. Some early proposals for machines which could interact with their environments were suggested in the mid to late 1980s,[284] as had been originally suggested in the cybernetics movement of the 1940's.[285] Robotics introduced a new approach to AI because behaviour could emerge from learning generated by neural networks.[286] Robots that can navigate the environment and interact with its features can build up knowledge and reasoning from experience interacting with the world.[287] The focus shifted from building computer programs as components of intelligence to building agents which generate knowledge and intelligence from experience.[288] In 1990, Rodney Brooks published the article 'Elephants Don't Play Chess' which argues that symbolic AI is "fundamentally flawed" given the symbols it uses are "ungrounded

281. Crevier, *AI*, 211; Lee, *Artificial Intelligence in Daily Life*, 26.
282. Lee, *Artificial Intelligence in Daily Life*, 27.
283. Baldwin, *The Globotics Upheaval*, 104.
284. Beer, 'Dynamical Systems and Embedded Cognition', 128; Lee, *Artificial Intelligence in Daily Life*, 27.
285. Husbands, 'Robotics', 272; McCorduck, *Machines Who Think*, 442.
286. Steels, 'Fifty Years of AI', 23.
287. Arkoudas and Bringsjord, 'Philosophical Foundations', 55.
288. Wooldridge, *A Brief History of Artificial Intelligence*, 93.

representations."[289] While these symbols aim to represent entities in the world, their meaning is unknown by the system and only the human user is capable of connecting the symbols to their meaning.[290] Moreover, symbolic systems work with discrete concepts and relations, generating an inflexibility or rigidity to the system itself.[291] Instead, Brooks proposed intelligent systems ought to be *embodied*, or robotic, to ground these symbols or representations on environmental information captured by sensors.[292] The benefit to this approach is that agents would be able to react to dynamic, noisy environments by having smaller subsystems generate simple behaviours, where complex behaviour emerges from the interaction of these underlying components.[293] By creating autonomous entities, learning occurs by interacting with the world and generating learned associations from experience rather than being programmed into the system directly.[294] Although Brooks was on the right track with respect to producing responsive and context-sensitive behaviours, embodiment is insufficient for symbol grounding, an idea which will be explained further in this chapter and in subsequent chapters.

This new approach to AI brought about the "Third Golden Age" in the early 1990's which continues to this day.[295] In 1993, Rodney Brooks began work on Cog the Humanoid Robot with a team of researchers and graduate students which included Cynthia Breazeal,[296] who would later become the Dean of Digital Learning at MIT.[297] Cog is a torso bolted to a bench with

289. Brooks, 'Elephants Don't Play Chess', 3.
290. Brooks, 4.
291. Brooks, 4–5; Dreyfus, *What Computers Can't Do*, 61.
292. Brooks, 'Elephants Don't Play Chess', 5.
293. Brooks, 14.
294. Alonso, 'Actions and Agents', 235.
295. Ekmekci and Arda, 'History of Artificial Intelligence', 13; Lee, *Artificial Intelligence in Daily Life*, 27.
296. Brooks et al., 'The Cog Project', 52; McCorduck, *Machines Who Think*, 463.
297. MIT, 'People Overview Cynthia Breazeal'.

arms and a head capable of machine-vision.[298] As previously stated, 1997 witnessed the defeat of the world champion chess player Gary Kasparov in a match against Deep Blue, a symbolic AI system developed by IBM.[299] Two years later, the robotic dog AIBO by Sony was introduced to the public after years of development in Japan,[300] and in 2000, Kismet, a robot that could recognize and simulate facial expressions and emotions, was created by Cynthia Breazeal.[301] Despite appearing to represent emotional states, Kismet's expressions were superficial and unaccompanied by any underlying experiences or changes to its internal motivations, unlike humans and animals. As discussed in depth in a later section of this chapter, emotional expressions reflect internal, affective states which are subjectively experienced by the individual exhibiting them, often in response to particular stimuli or circumstances.

The year 2000 was also the launch of *developmental robotics*, an approach to symbol grounding based on a model of human learning in childhood.[302] By modelling the stages of childhood development, researchers can create architectures which generate increasingly sophisticated behaviours with practice.[303] The emerging behaviours which manifest are the result of interactions with its environment,[304] incorporating new information about its features and objects, rather than having these aspects programmed within the robot's code.[305] Instead, these

298. Husbands, 'Robotics', 278.
299. Franklin, 'History, Motivations, and Core Themes', 23; Haenlein and Kaplan, 'A Brief History of Artificial Intelligence', 8; Lungarella et al., 'AI in the 21st Century – With Historical Reflections', 3.
300. Lee, *How to Grow a Robot*, 29.
301. Husbands, 'Robotics', 278–79.
302. Cangelosi and Schlesinger, *Developmental Robotics*, 5.
303. Adami, 'A Brief History of Artificial Intelligence Research', 132; Cangelosi and Schlesinger, 'From Babies to Robots', 184; Lee, *How to Grow a Robot*, 232.
304. Clark, *Being There*, 113; Thompson, 'Sensorimotor Subjectivity and the Enactive Approach to Experience', 13; Ziemke, 'The Body of Knowledge', 9.
305. Lungarella et al., 'AI in the 21st Century – With Historical Reflections', 3; Vernon and Furlong, 'Philosophical Foundations of AI', 60.

robots use incoming sensory data to learn about objects, actions, and environmental features to generate appropriate responses over time through experience.[306] By combining and refining simple actions and mechanisms, larger and more complex abilities and functionality can emerge over time, as capacities are generated from repeated experiences.[307] Learning the names of objects provides knowledge of the world for the robot, and this knowledge may be used in the generation of linguistic expressions and verbal communication.[308] By following a model of human development, a potential solution to the symbol grounding problem had been identified, further corroborating that biological models provide a promising direction for generating solutions to problems within AI.

AI has since exploded in its ability to perform a number of sophisticated tasks, especially thanks to *Big Data*,[309] a phenomenon which roughly includes the collection, processing, and sharing of data generated by users of devices and services.[310] Also contributing to the rise of neural networks were further improvements in computing hardware and the construction of global internet infrastructure.[311] Moreover, language generation continued to improve, launching new consumer devices to execute tasks through voice commands. In 2010, Apple released Siri, a virtual assistant on iOS, adapting to create unique profiles to individualize the experience.[312] The following year, the natural language system Watson created by IBM defeated two former

306. Bhaumik, *From AI to Robotics*, 34.
307. Law et al., 'Infants and iCubs', 273.
308. Lungarella et al., 'AI in the 21st Century – With Historical Reflections', 5; Hoffmann and Pfeifer, 'Robots as Powerful Allies for the Study of Embodied Cognition from the Bottom Up', 850.
309. Haenlein and Kaplan, 'A Brief History of Artificial Intelligence', 5.
310. Boyd and Crawford, 'Critical Questions for Big Data', 663–64; Leonelli, 'Scientific Research and Big Data'.
311. Sejnowski, *The Deep Learning Revolution*, 2018, 25; Wooldridge, *A Brief History of Artificial Intelligence*, 139.
312. Wooldridge, *A Brief History of Artificial Intelligence*, 97–98.

Jeopardy! champions Ken Jennings and Brad Rutter.[313] The first social robot for families named Jibo, developed by Cynthia Breazeal, launched in 2014 and was intended to serve as a personal assistant for the home, capable of tasks such as placing phone calls, setting reminders, and taking photos.[314] This was also the year Google acquired the company DeepMind with its neural network system which learns to play games through trial and error only.[315] In 2016, a DeepMind system named AlphaGo beat world champion Go player Lee Sedol,[316] further demonstrating the power of deep neural networks for learning. Today, ChatGPT has been responsible for producing much excitement around *large language models* (LLM) as it provides detailed responses to questions asked by users.[317] In 2023, Bing released their text-to-image generation system, producing detailed renderings incorporating the elements specified by user input.[318] Text-to-video generation would follow about a year later with the introduction of a model called Sora by OpenAI, the creator of ChatGPT.[319] Thus, AI seems on track to continue meeting and potentially exceeding human ability in some ways, however, the question remains about whether it will meet our expectations with respect to complex social interaction and empathy.

By investigating the history of artificial intelligence, it becomes apparent that programs inspired by cognition and biological processes running on digital computers have become the dominant approach to recreating human intelligence. As scientific progress and technological development expanded over the 20th century, our approaches to AI and software development

313. Franklin, 'History, Motivations, and Core Themes', 23.
314. Hodson, 'The First Family Robot', 21.
315. Wooldridge, *A Brief History of Artificial Intelligence*, 123–24.
316. Wooldridge, 127.
317. Kasneci et al., 'ChatGPT for Good?', 1.
318. Ribas, 'Building the New Bing'.
319. OpenAI, 'Sora'.

increasingly turned to biology for inspiration. Neural networks, especially when paired with large datasets, became increasingly capable of generating impressive results, proposing a new approach to recreating human behaviour. Moreover, when a neural network is embodied in a robot, it can learn about the world in a manner which is similar to human development. The successes produced by this approach to AI continued to reinforce the idea that modelling biological processes can generate fruitful results, as it became increasingly apparent that the body and exterior environment influence cognition and behaviour. The following section examines particular outcomes from an industry-leading robot[320] named iCub to demonstrate the current abilities of developmental robots.

4.2 Introducing iCub

This section explores a particular developmental robot named iCub to demonstrate a specific implementation of a model of human development. A description of its architecture is provided to give context to its abilities, illustrating how behaviours result from computer code. Currently, iCub is able to learn the names of objects and converse with humans. Examining iCub provides insight into the current status of robot development with some ability to predict how these machines might develop during the next few years and into the future. Although it may seem like iCub has an understanding of the world, I ultimately argue that there is no genuine perspective in this machine. The discussion of its functionality presented here provides the explanatory

320. Metta et al., 'The iCub Humanoid Robot', 1133.

foundation for the argument presented in the next section; specifically, why iCub cannot adopt a human perspective.

iCub is the result of a collaborative research initiative called RobotCub, funded by the European Commission beginning in 2004 to create an embedded cognitive system with artificial intelligence.[321] Created by Giulio Sandini, Giorgio Metta, and David Vernon, this humanoid robotic platform for researching embodied cognition consists of open-source hardware and software, allowing others to build upon existing frameworks to contribute to the development of iCub.[322] Modelled after a three year old child, iCub stands at 104 centimetres and weighs 22 kilograms with 53 *degrees of freedom* (DOF), most of which are located in the upper body.[323] Its hands have been carefully designed, with 9 DOF, independent thumbs, index, and middle fingers, and fully driven by tendons to operate like a human hand.[324] Moreover, the fingertips, palm, and forearm contain pressure sensors.[325] The reason is because humans learn about the world through the manipulation, interaction, and feedback of objects in an environment, making this a necessary aspect of iCub's functionality.[326] The robot contains two digital cameras for binocular vision, microphones for auditory processing, gyroscopes for detecting orientation, accelerometers for measuring the rate of change of velocity, and torque sensors collecting force data used for governing motor signals.[327] The main computational unit is located in the head,[328]

---

321. Istituto Italiano di Tecnologia, 'iCub History'; Sandini, Metta, and Vernon, 'RobotCub', 13.
322. Metta et al., 'iCub', 1; Parmiggiani et al., 'The Design of the iCub Humanoid Robot', 3.
323. Metta et al., 'iCub', 1–2.
324. Metta et al., 2–3.
325. Parmiggiani et al., 'The Design of the iCub Humanoid Robot', 15.
326. Metta et al., 'The iCub Humanoid Robot', 1127.
327. Sandini, Metta, and Vernon, 'The iCub Cognitive Humanoid Robot', 361.
328. Metta et al., 'iCub', 3–4.

however, further processing of iCub's sensorimotor data is processed externally in a separate computer,[329] where this information can be transmitted wirelessly or through an Ethernet cable.[330]

The iCub architecture consists of 3 layers: hardware, firmware, and software, in which the firmware serves as a functional bridge between the hardware and the software.[331] Data produced by hardware sensors must be preprocessed before being used by the software, and as such, relies on firmware for proper formatting. The firmware often used in iCub is called YARP which stands for Yet Another Robotic Platform. However, iCub can run on other middleware architectures as well, such as the Robot Operating System (ROS) and OROCOS or Open Robot Control Software.[332] YARP runs on dedicated CPUs throughout the body, processing sensorimotor data when communication between boards uses protocols.[333] It achieves this through *ports* which are assigned with names such as "camera/right" to access information produced by hardware.[334] These ports can send and receive data to any other port, and connections between ports are easily added or removed, using a variety of different protocols or transports.[335] Together, they create an inter-communicative network where data from sensors can be accessed throughout the network.[336] Moreover, the internal state of iCub can be viewed by software on the external machine to assist developers with troubleshooting and testing.[337] These

329. Parmiggiani et al., 'The Design of the iCub Humanoid Robot', 20.
330. Natale et al., 'The iCub Software Architecture', 2.
331. Sandini, Metta, and Vernon, 'The iCub Cognitive Humanoid Robot', 363.
332. Natale et al., 'The iCub Software Architecture', 4; 'The Orocos Project'.
333. Natale et al., 'The iCub Software Architecture', 2.
334. Metta et al., 'The iCub Humanoid Robot', 1130.
335. Metta et al., 1130.
336. Metta et al., 'iCub', 5.
337. Natale et al., 'The iCub Software Architecture', 16.

software systems use libraries of computer code which provide instructions for iCub to perform behaviours. One example is the iCub-HRI library which supports perception, object manipulation, social interaction, including face and action recognition, babbling, and learning the names of objects.[338]

The cognitive abilities of iCub are generated from sensorimotor mapping and coordination from basic reflexes which serve as the foundation for learning more complex behaviours.[339] It first learns to improve control and coordination of its eyes and neck by attending to objects or features of the environment.[340] With further development, iCub gains better control over movement in its shoulders, elbows, and torso, generating representations of its own body. It learns to crawl before standing and walking,[341] and with its hands free to reach for objects, iCub's manual dexterity will continue to improve.[342] Similarly to human children, this also facilitates object manipulation as the robot can pick up and inspect objects. As its ability to grasp and handle objects improves, iCub builds up other abilities like object recognition and word learning.[343] By interacting with humans, iCub can learn the names of objects and features of the world, following the gaze of others to attend to specific items or stimuli.[344] By using internal models of its body, iCub is eventually able to navigate its environment to learn more about its features and dynamics, both physical and social.[345] This "ontogenetic training program"

338. Fischer et al., 'iCub-HRI', 3–4.
339. Vernon, Metta, and Sandini, 'The iCub Cognitive Architecture', 124.
340. Law et al., 'Infants and iCubs', 273.
341. Vernon, Metta, and Sandini, 'The iCub Cognitive Architecture', 124.
342. Sandini, Metta, and Vernon, 'The iCub Cognitive Humanoid Robot', 365–66.
343. Sandini, Metta, and Vernon, 364.
344. Vernon, Metta, and Sandini, 'The iCub Cognitive Architecture', 124.
345. Shaw, Law, and Lee, 'Representations of Body Schemas for Infant Robot Development', 127.

follows human development as skill acquisition and improvement enables the robot to build up more complex behaviours by mastering and combining simple actions.[346]

This improved control and dexterity allows iCub to learn object *affordances* which are "action possibilities" for a given item, like grasping the handle of a mug.[347] The theory of affordances was first introduced by psychologist James J. Gibson and suggests a "complementarity of the animal and environment" given the interaction between the individual's body and elements of the physical world.[348] The environment offers or *affords* actions which can be performed as a result of the body's morphology. As such, it suggests that humans and animals have a perceptual ability which can identify the utility or purpose of an item or environmental feature given its specific physical composition.[349] iCub learns affordances using Bayesean networks as a model of the discrepancies between actions, objects, and their effects based on information from sensors in the hands and eyes.[350] From these variables and discrepancies, it is able to learn the relations between actions, objects, and effects where these relationships are updated as iCub continues to gain experience. Just like human children, iCub learns how to use tools and interact with objects by investigating their features through play and experimentation.[351]

346. Sandini, Metta, and Vernon, 'The iCub Cognitive Humanoid Robot', 365.
347. Metta et al., 'The iCub Humanoid Robot', 1131.
348. Gibson, *The Ecological Approach to Visual Perception*, 127.
349. Gibson, 128.
350. Metta et al., 'The iCub Humanoid Robot', 1131–32.
351. Mar et al., 'Self-Supervised Learning of Grasp Dependent Tool Affordances on the iCub Humanoid Robot'.

It is through these interactions that iCub is thought to ground language in sensorimotor information by creating associations between spoken language and objects.[352] As a separate research project, also funded by the European Commission, the iTalk Project aims to create a cognitive platform to support language learning for the iCub robot.[353] By combining auditory signals from speech with visual information, iCub creates an inner representation which it uses to determine which objects are being referred to in order to reach for the ascribed objects in the environment, guided by feedback from its hands to better direct its movement.[354] Similarly, the Epigenetic Robotics Architecture (ERA) can also generate word learning using self-organizing maps (SOMs) trained to recognize features, body position, and speech data, in which associations between mapped representations follow from the frequency of the association.[355] The SOMs are first trained to recognize objects and their features before learning to represent more complex objects and relations.[356]

A framework for emotions has also been developed, allowing iCub to mimic expressions detected from scanning human faces, where positive reinforcement from humans functions as a reward for a SOM to learn whether its identification is correct.[357] iCub can recognize and imitate seven "universal" emotions: anger, disgust, sadness, fear, surprise, happiness, in addition to a neutral face. Furthermore, an architecture of emotion has been developed which allows iCub to tailor its behaviour to reflect the detected mood or personality of the individual its interacting

352. Marocco et al., 'Grounding Action Words in the Sensorimotor Interaction with the World', 11.
353. Tikhanoff et al., 'An Open-Source Simulator for Cognitive Robotics Research'.
354. Cangelosi et al., 'The iTalk Project', 12; Tikhanoff, Cangelosi, and Metta, 'Integration of Speech and Action in Humanoid Robots', 27.
355. Cangelosi and Schlesinger, 'From Babies to Robots', 2.
356. Morse et al., 'Epigenetic Robotics Architecture (ERA)', 335.
357. Churamani et al., 'iCub', 1.

with, based on how engaged they are with the robot.[358] If iCub detects a person uninterested in interacting, it will continue to look down and play with its toys, however, if individuals take an interest in iCub, it will make eye contact, direct its gaze at its toys, and point to them, encouraging the participant to interact.[359] This architecture also allows for emotion recognition and mimicry, choosing actions which are predicted to increase the human participant's happiness during interactions.[360]

iCub can also learn to count to ten using its fingers. An experiment by Di Nuovo et al. demonstrates an ability for iCub to map auditory information of spoken numbers to its hands, moving the appropriate finger joints in response to the numbers one through ten when spoken in sequence.[361] As is the case with human children, finger counting aims to assist in the development of internal representations of numbers, establishing a foundation for developing a broader understanding of mathematical concepts in future research.[362]

This glimpse at iCub and its capacities indicates that AI embodiment appears to be a promising line of research for developing agents capable of socializing with humans. It seems plausible that with more research and development, iCub and similar robots will one day be able to interact and communicate with others in a humanlike manner. Because iCub is able to learn to recognize and respond appropriately to words for items in its environment, it appears this ability to learn from practice and experience is sufficient for establishing the meanings of words. In the

358. Tanevska et al., 'A Cognitive Architecture for Socially Adaptable Robots', 200.
359. Tanevska et al., 199.
360. Tanevska et al., 'Towards an Affective Cognitive Architecture for Human-Robot Interaction for the iCub Robot', 4.
361. Di Nuovo et al., 'The iCub Learns Numbers', 693.
362. Di Nuovo et al., 698.

following section, an examination of emotions and affective states demonstrates why this is not the case.

4.3 iCub's Failure to Empathize

The goal to be met by iCub and developmental robotics in general is a solution to the symbol grounding problem, which, as introduced in Chapter 1, seeks to explain how words and concepts become associated with meanings. A response proposed to this problem was presented by Stevan Harnad in his paper 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem' where he expands upon Alan Turing's Imitation Game. Harnad recommends that instead of testing AI intelligence through conversing using text-based chat, AI should be embodied in a robot, such that they can "recognize and identify and manipulate and describe real objects, events and states of affairs in the world."[363] Harnad calls this modified intelligence test the *Total Turing Test* (TTT) explaining that "the candidate must be able to do, in the real world of objects and people, *everything* that real people do, in a way that is indistinguishable (to a person) from the way real people do it."[364] By building AI systems with integrated sensorimotor functionality, the meanings of words can be learned from bottom-up functionality as the agent engages with its environment.[365] Harnad recommends this should be implemented in a *"*non-symbolic" manner,[366] one which does not follow the approach called *symbolic reasoning* which

363. Harnad, 'Other Bodies, Other Minds', 44.
364. Harnad, 44.
365. Harnad, 50–51.
366. Harnad, 51.

was common in early AI research, as discussed in the first section of this chapter. Alternatively, robotic systems should detect sensory features and elements within the environment from a bottom-up approach, similar to the way neural networks operate today. This way, these features and elements can be combined and associated together to create artificial neural representations which supposedly grounds the meanings of words. These representations, as patterns of activations in neural networks, would be used to support language comprehension and production, allowing the robot to recognize which object or feature is being referred to. Thus, according to Harnad, an AI which passes the TTT will act sufficiently humanlike and as a result, will be treated as if it has a mind of its own. Specifically, "if our only basis for judging that other people have minds is that they behave indistinguishably from ourselves, then there's no ground for withholding the same benefit of the doubt from robots."[367]

From the previous section's exploration of iCub and its current abilities, sensory projections exist as patterns of activations within neural networks and are used to organize incoming information from the external environment. Further, iCub learns about object affordances by "playing" with them, exploring the relationship between its body and aspects of the physical environment. When spoken words are presented as iCub is engaged with an object, it is able to create an association between item and its label. This extends to emotions as well, as iCub can identify the expression of basic emotional states in humans. Though this information influences iCub's behaviour, the robot does not feel or experience the emotional information its body uses to generate behaviours. As will be discussed shortly, *experiences* of emotions and

367. Harnad, 46.

affective states are required to ground the meanings of words and concepts related to emotions. In fact, experience is required to ground *all* meaning, because meaning is derived from the perspective of the subject[368] as a consequence of its physiological structure. The mind directs the body to behave in appropriate ways in response to environmental stimuli to further or improve its chances of survival. Sentience, or the ability to sense and respond to environmental stimuli,[369] is required for understanding what a symbol, word, or idea means. Computers are not sentient because they are not structured in a self-organizing manner, an idea which is explained in the following chapter. Because of this, it cannot understand the meanings of words and symbols. The behaviour of a computers emerges in a distinct fashion than those in humans and animals, emerging from syntactical structures rather than direct references to features of the environment. Since iCub is an embodied computer, it has no access to meaning of any kind, a fact which becomes apparent when investigating affect and emotion in particular.

As mentioned previously, iCub can learn to recognize human emotions like happiness, sadness, anger, among others, based on the movement of muscles in the face. These categories of affect have been described as *basic emotions* by the psychologist Paul Ekman from his study of facial expressions in the 1960's and onward.[370] Today, Ekman's position is known as Basic Emotion Theory (BET), however, it is one of two prominent theories of emotion which have been the source of debate. His work builds on Charles Darwin's theory of emotion as articulated in *The Expression of the Emotions in Man and Animals*. Ekman's research primarily investigated

368. Harnad, 'The Symbol Grounding Problem', 339.
369. *Merriam-Webster.com Dictionary*, s.v. "sentient."
370. Ekman, 'Body Position, Facial Expression, and Verbal Behavior during Interviews', 300; Ekman and Friesen, 'Head and Body Cues in the Judgment of Emotion', 718.

the measurement of behavioural cues, particularly facial expressions, which are consistently or universally associated with emotions.[371] Particular movements of facial muscles are associated with emotions, and these can be grouped into six "basic" emotions: happiness, sadness, fear, anger, disgust, and surprise.[372] Moreover, basic emotions are associated with nine specific characteristics.[373] The first characteristic is that basic emotions exhibit distinctive universal signals, involving a set of configurations of facial muscle movements. Though similarities between individuals exist, differences can also be observed. Anger, Ekman notes, can involve more than sixty particular expressions depending on the situation and causal factors, in which subtle differences may involve the lips tightly pressed together or slightly open and "in a square shape."[374] These variations within categories of emotion are due to an array of influences, including individual physiological differences, previous experiences, and cultural norms.[375] As such, the boundaries between groups or categories can be blurry and ill-defined, however, exist nonetheless.[376] Other characteristics of basic emotions listed by Ekman include their existence in other primates besides humans, and consisting of particular physiological changes within the body.

Ekman appeals to changes in the *autonomic nervous system*, responsible for bodily functions such as heart rate and breathing, to demonstrate the adaptive qualities of emotions, and thus their universal qualities.[377] It has been suggested that these automatic responses serve an

371. Ekman, 'Darwin's Contributions to Our Understanding of Emotional Expressions', 3449.
372. Ekman, 'An Argument for Basic Emotions', 170.
373. Ekman, 175.
374. Ekman, 172.
375. Ekman, 'Universal Facial Expressions of Emotion', 152–53.
376. Ekman, 'An Argument for Basic Emotions', 173.
377. Ekman, 'Are There Basic Emotions?', 552.

adaptive purpose which is not limited to humans. For example, a facial expression indicating disgust suggests to conspecifics to avoid eating a particular item,[378] while fearful expressions indicate the presence of a threat which can be observed by others. Moreover, basic emotions can be reliably initiated by certain events or stimuli, resulting in emotions which are shared by most individuals.[379] The loss of a loved one, for example, regularly invokes sadness in the majority of individuals from a number of cultural backgrounds. Furthermore, basic emotions are characterized by a quick onset and a brief duration, involving an automatic appraisal of the stimulus or situation which is not intentionally induced by the individual. Basic emotions also indicate a "coherence among emotional response," in which changes in facial expressions are usually accompanied by physiological changes.[380] Finally, basic emotions involve an "unbidden occurrence" which, as the result of automatic, physiological changes and a rapid onset, happen *to* an individual rather than being chosen *by* them. Ekman succinctly states "One cannot simply elect when to have which emotion."[381] Thus, basic emotions involve a number of universal characteristics, and while these characteristics involve a degree of variation, are said to be present in all humans, regardless of how affective states are subsequently expressed or regulated.

BET considers emotions to be *natural kinds* in which members of a category of emotion share one or more common features, namely a unique pattern of specific outputs in the form of behaviours.[382] These categories are thought to exist ontologically, external to human minds and

---

378. Ekman, 'An Argument for Basic Emotions', 177.
379. Ekman, 184.
380. Ekman, 184–85.
381. Ekman, 189.
382. Barrett, 'Are Emotions Natural Kinds?', 29.

out in the world waiting to be discovered.[383] Prior to Darwin's theory of natural selection, categories of emotions were considered to be akin to Platonic types, or ontologically-independent aspects of psychological functioning which gives rise to behaviours once instantiated within particular individuals.[384] Distinguished by signatures in the body, each emotion's physical *essence* is the way it generates changes in the peripheral nervous system, facial muscles, or the structure or function in the mammalian brain. As causal agents, basic emotions have been selected for by evolutionary processes, motivating individuals to behave in adaptive ways as indicated by stimuli detected in the environment, stemming directly from the detection of a stimulus.[385] As such, they *cause* individuals to act in certain ways, and because of their adaptive benefit, have evolved from their utility for responding to environmental changes.[386]

As mentioned above, there exists a debate surrounding the categories of emotion. Do these categories exist as ontological, causal agents in humans and animals, or are these categories instead *effects* of other causal factors and organized as such by human minds?[387]

An alternative view to BET considers the categories of emotion to be epistemological rather than ontological, in which emotional responses are categorized by the way they feel or appear. This is the view of Social Constructionist Theory (SCT) which claims there are no preestablished causal categories of emotions that have emerged from evolutionary processes.

---

383. Scarantino and de Sousa, 'Emotion', sec. 8.2; Barrett, 'Are Emotions Natural Kinds?', 29.
384. Barrett, 'Psychological Construction', 379.
385. Barrett, 380.
386. Barrett, 'Are Emotions Natural Kinds?', 31.
387. TenHouten, 'Basic Emotion Theory, Social Constructionism, and the Universal Ethogram', 613.

Instead, emotions are outcomes of the same physiological systems which give rise to other cognitive processes.[388] These affective responses are subsequently labelled as belonging to a particular category of emotion based on similarities and patterns which arise in a regular or predictable way.[389] This view has origins in the early psychology of William James and Wilhelm Wundt, as emotion words were considered misleading as they inspire thinking in an *essentialist* manner.[390] Since then, this theory has grown in popularity as no evidence has indicated a one-to-one correspondence between bodily states and emotional categories.[391] Furthermore, brain imaging studies have failed to uncover localized brain regions or centres for basic emotions.[392] Given this, categories of emotion and their terms like 'sadness' and 'anger' are considered to be "technical terms" and are used to categorize outcomes of physiological and cognitive processes.[393] Words for emotion categories do not refer to causal mechanisms which produce these states, but instead, mental states which arise from attending to thoughts and feelings related to affective states.[394] Thus, attention and language contribute significantly to the way we think about emotions, as to label an affective state, it must be in the focus of attention.[395]

In SCT, the foundation which generates affective states and emotions is called *core affect* which involves interactions between two dimensions: pain/pleasure or *valence*, and

388. Barrett, 'Psychological Construction', 380.
389. Lindquist, 'Emotions Emerge from More Basic Psychological Ingredients', 357.
390. Barrett, 'Psychological Construction', 386; Lindquist, 'Emotions Emerge from More Basic Psychological Ingredients', 357.
391. Barrett, 'Psychological Construction', 383.
392. TenHouten, 'Basic Emotion Theory, Social Constructionism, and the Universal Ethogram', 614.
393. Barrett et al., 'The Experience of Emotion', 390; Russell, 'Core Affect and the Psychological Construction of Emotion', 146.
394. Barrett et al., 'The Experience of Emotion', 388–89.
395. Barrett, 'Psychological Construction', 383.

excitement/relaxation or *arousal*.[396] These dimensions are generated by the body from specific physiological mechanisms which have been selected for by evolutionary processes. For example, physical pain is caused by the activation of nociceptors, a type of nerve cell found in the skin and tissues of many animals.[397] The physical pain experienced by a particular stimulus, like the thorn of a plant or stinger on a bee or wasp, signals the presence of an entity capable of causing harm to the individual. Any affective reactions or emotions arising from the processing of sensory information creates an association between causal factors and their effects within the body, providing the individual with a source of information about elements of the environment.[398] Individuals stung by a wasp may develop a fear of these insects in an effort to avoid interacting with them in the future. Alternatively, this individual may not feel a sense of fear upon seeing wasp, if they were indoors and behind a pane of glass. As such, though the body and its physiological processes give rise to core affect, the emotions and subjective experiences they produce also depend on the context in which they occur.

Concepts related to emotion and affective states thus emerge from noticing physiological regularities and labelling subjective experiences.[399] According to SCT, categories of emotion like *fear* and *anger* emerge from noticing the ways in which objects or situations generate affective

396. Barrett and Lindquist, 'The Embodiment of Emotion', 250; Lindquist et al., 'The Brain Basis of Emotion', 125; Russell, 'Emotion, Core Affect, and Psychological Construction', 1264; Schimmack and Grob, 'Dimensional Models of Core Affect', 327.
397. Dubin and Patapoutian, 'Nociceptors', 3760; Kavaliers, 'Evolutionary and Comparative Aspects of Nociception', 929.
398. Barrett and Lindquist, 'The Embodiment of Emotion', 251; Barrett et al., 'The Experience of Emotion', 382; Phelps, 'Emotion and Cognition', 29.
399. Barrett and Lindquist, 'The Embodiment of Emotion', 252; Barsalou and Wiemer-Hastings, 'Situating Abstract Concepts', 129.

responses and motivate actions.[400] These concepts are useful for emotional self-regulation, working toward achieving a goal,[401] as well as considering the ways in which others may experience certain types of situations.[402] In humans, emotions become articulated in concepts and language, in which these ideas and the ways they are expressed are influenced by background knowledge and social norms.[403] These variables influence the way emotions manifest and are experienced at any given time, and as such, one's conception of an emotional event is highly *heterogeneous*.[404] For example, one instance of fear may feel or be outwardly expressed distinctly on different occasions, despite both being appropriately experienced and labelled as 'fear'. While both instances of fear involve the expectation of painful or negative stimuli,[405] the way an individual's feelings are outwardly expressed may differ due to circumstance. Moreover, the manner in which individuals describe their feelings can also vary in *granularity*. Terms may be described with *high-granularity*, discrete and specific, noting slight differences between states, or *low-granularity*, in which affective states are described vaguely through general terms, such as pleasure or displeasure.[406] Together, these considerations indicate a number of variables

400. Barrett and Lindquist, 'The Embodiment of Emotion', 251; Barsalou et al., 'Grounding Conceptual Knowledge in Modality-Specific Systems', 84.
401. Barrett and Lindquist, 'The Embodiment of Emotion', 252; Benita et al., 'Emotion Regulation during Personal Goal Pursuit', 84; Barrett and Gross, 'Emotional Intelligence', 286.
402. Barrett and Lindquist, 'The Embodiment of Emotion', 252; Keltner and Haidt, 'Social Functions of Emotions at Four Levels of Analysis', 506; Levenson et al., 'Emotion', 448.
403. Barrett and Lindquist, 'The Embodiment of Emotion', 252–53; Carr, Kever, and Winkielman, 'Embodiment of Emotion and Its Situated Nature', 537–38; Barrett, Lindquist, and Gendron, 'Language as Context for the Perception of Emotion', 328.
404. Barrett and Lindquist, 'The Embodiment of Emotion', 253; Winkielman, Davis, and Coulson, 'Moving Thoughts', 1532.
405. Haikonen, *The Cognitive Approach to Conscious Machines*, 111–12.
406. Barrett et al., 'The Experience of Emotion', 388.

present in the process of conceptualization and interpretation of affective states, in which labels for emotions are ascribed based on contextual or situational factors and sociocultural norms.

Overall, the debate on BET or SCT as a better theory continues today, and attempts to reconcile these two perspectives into one theory has also emerged.[407] Though the debate lingers, adherents to each perspective do not deny the significance of the other theory. Proponents of BET agree that social norms shape the way emotions are expressed, and proponents of SCT acknowledge the significance of biological patterns which are responsible for shaping and characterizing emotion categories. Each position considers the other theoretical stance to be theoretically relevant; however, to what degree is part of the debate.[408] Moreover, there are also variations on BET, in which some are interested in maintaining the idea of basic emotions while refuting that they exist as natural kinds.[409] Overall, though the debate continues, it remains uncontroversial to claim that the brain and its cognitive faculties, along with the body and its physiological processes, work together to give rise to affective states which typically motivate individuals to perform certain actions.[410]

To demonstrate the relationship between the brain and body, consider the example of *fear*, an arguably paradigmatic emotional experience. Fear involves a particular physiological response profile for organisms to generate adaptive behaviours for avoiding harm. Upon detecting a stimulus which indicates a potential threat, a region of the brain called the *amygdala*

---

407. TenHouten, 'Basic Emotion Theory, Social Constructionism, and the Universal Ethogram', 611.
408. TenHouten, 612.
409. Scarantino and Griffiths, 'Don't Give Up on Basic Emotions', 450.
410. Scarantino and de Sousa, 'Emotion', sec. 11.

sends a cascade of changes throughout the brain and body. While the amygdala is involved in these responses, it also responds to the detection of novelty and rewards, indicating its role involves more than generating fear responses.[411] Moreover, not every instance of fear activates the amygdala.[412] It is, however, associated with the activation of the hypothalamic-pitutary-adrenocortical axis (HPA axis) in response to a perceived threat.[413] Upon the recognition of a stimulus, signals from the brain travel down the spinal cord to release hormones from a region known as the *adrenal cortex* located above the kidney.[414] These hormones, epinephrine and cortisol, generate a range of multifaceted physiological effects, such as an increase in heart rate, blood pressure, respiration, glucose production, and the suppression of immune system functions.[415] In addition, the relaxing and expanding arteries allow more blood to flow throughout the body, delivering oxygen and nutrients to muscle tissue to facilitate specific adaptive strategies. These processes all contribute to the subjective experiences of an emotional state in response to a stressful stimulus.[416] From the subject's point of view, a racing heart and increase in sweat production feels like heat in the body and an abundance of energy. Moreover, epinephrine and cortisol bind to receptors in the brain, including amygdala and hippocampus, which regulate the activity along the HPA axis.[417] If the stimulus indicating a threat diminishes or changes, activity in the HPA axis adjusts accordingly, slowing the release of epinephrine and cortisol and allowing the body to return to a state of rest. This example of fear clearly

411. Lindquist et al., 'The Brain Basis of Emotion', 130; LeDoux, 'The Amygdala', 2007, R873.
412. LeDoux, 'The Amygdala', 2007, R873.
413. Tsigos and Chrousos, 'Hypothalamic–Pituitary–Adrenal Axis, Neuroendocrine Factors and Stress', 866–67.
414. Fulford and Harbuz, 'An Introduction to the HPA Axis', 44.
415. Mueller, Figueroa, and Robinson-Papp, 'Structural and Functional Connections Between the Autonomic Nervous System, Hypothalamic–Pituitary–Adrenal Axis, and the Immune System', 952–53.
416. Mueller, Figueroa, and Robinson-Papp, 953.
417. LeDoux, 'The Amygdala', 1 August 1994, 234.

demonstrates the way the body and brain coordinate to give rise to physiological and affective states which facilitate adaptive responses to perceived threats. Moreover, the body and its physiological processes consist of an interactive system of feedback and modulation, where brain regions regulate physiological activity and contribute to the conceptualization of feelings.

From the internal feelings of fear, humans can control how they respond to subjective feelings, to a degree. While it can be difficult to control a racing heart, some outward expressions can be regulated or suppressed, such as the motivation to flee from a situation. Social factors influence how emotions are felt and expressed, in addition to the ways in which emotions are labelled and recognized. Different cultures regulate emotional expressions in different ways, depending on social norms. In certain Asian countries including China, Korea, and Japan, it is considered improper to display negative emotions like anger and contempt in public,[418] while other cultures may deem these expressions to be more permissible. Expressions of fear, on the other hand, may be consciously controlled in instances of combat and war, where soldiers exhibiting fear may instill confidence in the enemy. When dealing with certain wild animals, like cougars and black bears, fear-based responses such as running away may entice the animal to pursue an attack, as the individual can be perceived as a non-threatening target.[419] It is recommended that when dealing with these animals, acts of aggression indicate the presence of a threat or a difficult pursuit. In this way, regulating emotions depends on contextual factors, from

418. Porter and Samovar, 'Cultural Influences on Emotional Expression', 457.
419. Pester, 'Humans Are Practically Defenseless. Why Don't Wild Animals Attack Us More?'; Maron, 'What to Do If You're Attacked by a Bear—or Any of These Other Wild Animals'.

the situation at hand to sociocultural norms, subsequently shifting how emotions are experienced and conceptualized.

From this brief analysis of a significant debate in the theory of emotions, it is reasonable to conclude emotions involve a set of behaviours and inner experiences aimed at responding appropriately to the detection of environmental stimuli. Subjective experiences associated with emotions are caused by or associated with physiological events or processes, such as a racing heart. As such, the body and mind interact with one another to give rise to affective states which can be ascribed a label, as per the effects of cognitive and physiological processes. Therefore, it seems that the preferable approach for adding emotional processing to iCub would be one which follows SCT. This is due to the generative flexibility offered from the two elements of core affect, namely valence and arousal, supporting a wider array of affective states than if BET were to be implemented instead. Since the categories identified by BET can be created from core affect, this approach would avoid the need to preprogram categories of emotions into the robot while retaining the ability to recognize and respond to the six categories identified by BET.

Even if a version of SCT were to be implemented into iCub, however, the outcomes produced from this approach are insufficient for grounding the meanings of words related to emotions. The reason for this is due to how information about emotions is generated. All information about emotions and affective states remains *exogenous* rather than *endogenous*, arising from human expressions and concepts rather than the robot's own body and subjective experiences. In biological organisms, the body and its physiological processes generate information about stimuli detected within the environment to better their chances at survival and reproduction. Consequently, any information a person knows about emotions is predicated on

their own subject feelings, and because iCub lacks feelings and phenomenal experiences, it cannot appeal to its own knowledge generated from bodily functions. For iCub, all knowledge of affective states comes from human concepts, whether gained through interactive experience, or written in the robot's code. Therefore, the robot cannot understand concepts related to affect, but also to other concepts and words relating to external objects and actions. Since meaning is generated internally from the lived perspective of an individual, rather than an inherent aspect of systems of representation, the referent of words is inaccessible thus leaving the word ungrounded.

The relationship between iCub's body and its simulated cognitive faculties is reversed in comparison to humans and other animals. Rather than its body generating robot "cognition" and behaviour, behaviour is intentionally produced from rules and instructions and exist independently from the body. Although a robot may act like it has a mind, it does not, as cognition arises from a specific structure of physiological processes.

The reason for this outcome is arguably due to the perspective adopted by robotics engineers, as a machine's parts are ideally assembled in ways which are modular and decoupled from one another. This modularity enables engineers and developers to easily replace and test a particular component separately, gaining specific knowledge about the part in question. This relationship between functionality and embodiment is distinct from animals and their physiological processes, since in biological agents, there is a high degree of coupling or dependency is due to evolutionary processes. Physical materials and their biochemical interactions give rise to an interconnected network of functionality which constitutes physiology. In robotics engineering, this degree of interconnection between functionality and its physical

implementation is avoided for practical reasons. Given this, however, robots like iCub do not sufficiently model the biological processes which are required for developing analogues of emotions, empathy, and sophisticated social behaviours. To produce these outcomes, a new approach to robotics engineering is required. One which is designed as an integrated, organized unity rather than a collection of decoupled functional modules built for the sake of generating specific behaviours. Alternatively, robot behaviour should be motivated by its experiences of environmental stimuli, involving learned associations between stimuli and sensory signals to avoid harmful situations.

In addition to grounding words and concepts related to emotion, affect should be incorporated into social robots given the significance of affective processing for cognition in general. The significant contributions of psychologist Antonio Damasio and his *somatic marker hypothesis* indicates affective processes contribute to cognitive abilities which are unrelated to emotion. Experiments conducted during the 1990's indicated decision-making abilities were negatively impacted in individuals with damage to frontal regions of the brain associated with emotional expression.[420] From this, Damasio suggests that rather than existing as separate from "rational" cognitive processes, emotions and emotional experiences significantly contribute to reasoning tasks.[421] Given the evidence of a high degree of coupling between emotions and reasoning, it is not enough to program iCub and other social robots with a theory of emotions, and instead, they must *feel* emotions as generated from the bodily processes. This approach

420. Damasio, 'The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex', 1413.
421. Damasio, 'Emotion and the Human Brain', 102.

would likely also improve the ways in which the robot operates and performs behaviours, appearing to reason in a manner more similar to a human than a computer.

Without emotion, iCub will likely produce behaviours which appear to be inappropriate or insufficient, indeed acting *robotically* during instances where affective sensitivity is required. One of these cases is during acts of empathy. As discussed, empathy requires the recognition and adoption of affective states as perceived in another. It relies on experiencing for oneself how another person is feeling within a particular interaction, where the observer, as a living being, is able to experience subjective states for themselves to understand their meaning. These experiences and subjective feelings are responsible for grounding behavioural cues and acts of empathy, as without an understanding of what emotional expressions refer to, any act of empathy is a mere simulation. Since iCub cannot appeal to its own knowledge of 'sadness' or 'pain', it must rely on humans to provide it with representations of these feelings. These representations could be implemented in code or by learning the behavioural cues related to specific emotions as exhibited by people. Any perceived understanding of concepts related to emotion would thus remain a very narrow representation of the ways in which affect manifests in animals including humans. For example, a smiling face is yet another symbol or representation of a real affective state, one which iCub does not have access to. Experience is therefore required for emotion, and emotional experience is required for empathy, as it requires one to adopt the perspective of another. Since this is missing in iCub, any act of perceived robot empathy remains a simile of a true instance of perspective-taking. It is important to note that this principle will still apply when robots of the future become more sophisticated at detecting and responding to human behaviour. Though these similes may one day appear to be convincing and impressive, any instances of empathy or emotion will remain a simulation.

These limitations indicate that a new approach to creating robots is required if they are to understand human emotions. The following chapter provides an alternative approach to building robots in a manner which is analogous to human and animal physiology to better model the processes which generate affect and emotion. This approach, and its corresponding robot prototype, is a better version of embodiment than iCub because the robot's hardware gives rise to its functionality, analogous to the way cognition and behaviours manifests in humans and animals. To further motivate this new approach, a brief discussion of the underlying principles involved in both living organisms and cybernetics will be considered. By appealing to the self-organizing processes and feedback systems in both domains, an improved model of living beings for building social robots can be identified.

In conclusion, this chapter examines the history of artificial intelligence to better understand the direction taken for social robots created today. From a history of computers, robot behaviours emerge in a distinct manner from the way behaviours emerge in living organisms, despite apparent similarities. Though outward appearances may indicate functional similarities, the behaviours of computers and social robots are produced by rules and instructions, rather than its experiences of its body as it lives as part of a wider world. Although iCub's body may react to stimuli and establish connections between sensory information, the robot has no experience of this occurring. Though theories of emotions can be programmed into robots like iCub, the requisite experiential perspective for grounding symbols, words, and concepts is absent. As such, a new approach to building social robots is required, otherwise, their inherent limitations will become even more apparent, especially during acts of empathy. Embodiment and emotion, for humans, is not ancillary nor separate from cognition, but is instead foundational for behaviour. The current approach taken by social robots demonstrates the opposite: the robot's body remains

a shell for its computer-generated "mind" to govern. Historically, the development of social robots and AIs over the 20th and into the 21st century witnessed a prioritization of behaviours and simulated cognitive capacities, only to be followed by the development of a robotic body. Because biological organisms develop in the opposite manner, in which cognitive capacities arise from bodily functions, a new approach is recommended if we are interested in overcoming inherent limitations identified in social robots. Affective states and the physiological processes which give rise to them are fundamental for behaviour, whether related to affective processing or not, suggesting iCub and other robots are insufficiently embodied for acquiring the requisite capacities for socialization.

5 Embodiment for Robot Experience

Words related to emotions are grounded in subjective experience because they refer to bodily states involving pain and pleasure. For a robot to understand what human behavioural cues mean, it must experience the world in an analogous manner to the way living beings do. To accomplish this, a brief investigation of self-organization in biology is necessary as background for incorporating experiences to robots. The specific form of embodiment required for subjective experience is one which includes associative learning and signals representing pain and pleasure. Experiences thus arise from a robot learning how various stimuli impact its body, grounding the meaning of emotions through the associations between stimuli and sensations that it learns.

5.1 Autopoiesis for Experience in Living Beings

This section presents an overview of a proposed explanation of life as a phenomenon in the natural world. As such, we can identify the requirements for embodiment, a precondition necessary for its role in generating experiences. In Chapter 3, it was demonstrated that embodied computers are not embodied in the way living organisms are, as the body of a robot is secondary for the primary purpose of creating minds and behaviour. Below, we will broadly examine life and embodiment to determine how experiences and the mind arise from living bodies; a strategy inverse to prevailing approaches to social robots. To start, let's investigate the foundational principle involved, *autopoiesis*, or the capacity for self-organization. This involves a collection

of physiological mechanisms interacting with one another to maintain a variety of functions required for the system. If variations arise within these processes, they can be offset through adjustments in other mechanisms within the system. Through evolutionary processes, organisms became better able to respond and adapt to changes, resulting in experiential learning. Additionally, subjective experiences which accompany the detection of stimuli and environmental changes assisted this process by informing individuals about what a particular stimulus means. These associations lead to responses which regulate interactions with the environment, thus increasing the likelihood the individual remains alive and will potentially reproduce. Organisms are inherently invested in the effects caused by environmental changes to their bodies. As such, self-organization can be considered a foundation for the grounding of meanings. For example, if a stimulus negatively impacts an aspect of the organism's functionality, sensory receptor signals present a feeling of harm which motivates avoidance. The body and its processes present signals about various elements of the environment, in which the meaning of these signals is derived from the ways in which bodies adapted for survival. Self-organization is the underlying mechanism of bodies of living beings, establishing the meaning of stimuli and guiding behaviour based on experiences of various sensory mechanisms.

The general principle supporting living things is an ability to self-organize on a number of different levels of systems, from cells to organs to more complex physiological constructs like the circulatory system. The concept of autopoiesis was first developed by Humberto Maturana and Fransisco Varela in Chile in their book *De Maquinas y Seres Vivos* in 1972.[422] Their work,

422. Maturana and Varela, *Autopoiesis and Cognition*, xvii; Stano et al., 'Autopoiesis', 105008.

however, would not receive wide-spread attention in English speaking countries until a paper was published in the journal *BioSystems* in 1974.[423] *Autopoiesis and Cognition*, an English version of their 1972 book, was published in 1980[424] and has since been slow to be adopted into mainstream science.[425] Some claim that this is due to biology's tendency to study life from the perspective of populations,[426] reproduction, and evolution[427] instead of the underlying mechanisms which lead to these outcomes. Rather, autopoiesis offers an explanation for a more fundamental level of biology by investigating the organization of bodily processes.

The explanatory strength of autopoiesis is due to its ability to distinguish living things from non-living things, providing an explanation distinct from other *descriptive* accounts for articulating what life is. Descriptive accounts often provide lists of properties characteristic of living organisms, like metabolism, reproduction, self-maintenance, containing genetic material like DNA, and evolving via mutations and natural selection.[428] Maturana claims that simply describing the characteristics of living systems cannot sufficiently articulate a definition of life as a physical phenomena.[429] Moreover, he claims that enumerating the properties of living organisms is an impossible task.[430] Due to these short-comings, living systems should instead be considered as a unity, "an entity distinct from a background"[431] and created by a physical boundary. In fact, the idea of a boundary is essential for autopoiesis, where the delineation

423. Stano et al., 'Autopoiesis', 105008.
424. Maturana and Varela, *Autopoiesis and Cognition*, iv.
425. Luisi, 'Autopoiesis', 58.
426. Luisi, 52.
427. Varela, Maturana, and Uribe, 'Autopoiesis', 187.
428. Thompson, *Mind in Life*, 96.
429. Maturana and Varela, *Autopoiesis and Cognition*, 49.
430. Maturana and Varela, 5.
431. Maturana and Varela, xix.

between interior and exterior is what separates an individual from other elements within the environment.[432] In a cell, for example, the cell membrane or outer layer creates a boundary which separates the inside of the cell from the outside.

So rather than focusing on physical properties of living organisms, as in, their qualities or attributes, Maturana and Varela are instead interested in organization and relations between components.[433] As such, these systems can be realized or instantiated in different mediums,[434] like software as observed in the domain of *artificial life.* The concept of interest is the unity which arises from the self-organization which emerges from a network of physical interactions. The necessary properties of the components are not of interest, but rather the necessary *organization* for systems to become a unity is of primary concern. This circular organization ensures the production and maintenance of the relations which give rise to the same organization which produces these relations.[435] As a result, the entity is *autonomous*, able to maintain its identity through the "active compensation of deformations"[436] or perturbations due to and despite physical interactions within the environment.[437] The regeneration required to maintain a network of interactions occurs inside the boundary, and this demarcation is made and sustained by these processes.[438] It is also semipermeable to let in physical materials,[439] as these self-sustaining systems still require incoming nutrients and energy, and as such, are operationally open

---

432. Luisi, 'Autopoiesis', 50.
433. Varela, Maturana, and Uribe, 'Autopoiesis', 188.
434. Maturana and Varela, *Autopoiesis and Cognition*, 77.
435. Maturana and Varela, 48.
436. Maturana and Varela, 73.
437. Maturana and Varela, 79.
438. Luisi, 'Autopoiesis', 51.
439. Luisi, 52.

systems.[440] An operationally closed system, in contrast, would neither require nor allow the incorporation of physical inputs or materials. There is, then, a dependence on the environment and external mediums in which these systems reside.[441] The fact that autopoietic systems are operationally open means the external environment shapes the internal structure of the organism, where changes within the external environment give rise to and promote internal changes.[442] Provided these changes do not disintegrate the relations which are necessary for their organizational capacity, the system is able to compensate for alterations which follow from environmental variation.[443] The relations generated by a network of processes which produce components of the autopoietic network are thus the most significant aspect of autopoietic systems.

In contrast, *allopoietic systems* are ones whose organization is not autonomous or self-sustaining but are created by means external to the system itself.[444] A car, for example,[445] as a unity is produced by processes independent from the organization and operation of the car itself, namely a factory with assembly workers and other machines. Computers and robots are also allopoietic systems, requiring creation and maintenance from humans or other allopoietic machines, as they are not autonomous and instead are reliant on some degree of human intervention for their continued operation. The distinction between autopoietic and allopoietic systems is significant, as it indicates a key difference between animals, including humans, and manmade artifacts like robots. Since artifacts like computers and robots exist as objects and

---

440. Luisi, 51; Stano et al., 'Autopoiesis', 105008.
441. Luisi, 'Autopoiesis', 54.
442. Luisi, 54; Varela, Maturana, and Uribe, 'Autopoiesis', 188.
443. Maturana and Varela, *Autopoiesis and Cognition*, 81; Varela, Maturana, and Uribe, 'Autopoiesis', 188.
444. Varela, Maturana, and Uribe, 'Autopoiesis', 189.
445. Maturana and Varela, *Autopoiesis and Cognition*, 79.

cannot function without some degree of intervention from human action, they are a qualitatively different kind of system than living organisms. This difference has important implications for how we should think of social robots as they continue to develop over time.

Autopoiesis can also account for genetic reproduction and variation between parents and offspring. During reproduction, some components might be slightly different due to mutations within genetic code, but if other systems are modified as a result, the individual may still survive. Provided its physiology can maintain self-sustainability, organizational variations can be compensated for by other aspects of the system.[446] From these self-organizing processes, individuals can survive and reproduce, and as mutations occur within the genetic code of offspring, populations may be able to compensate for certain alterations.

Thus, biological processes interacting with one another to support autopoiesis enable organisms to remain alive and reproduce. These processes need to be renewed and maintained, and the individual requires mechanisms to facilitate the fulfillment of these needs. At the level of the cell, one of these mechanisms involves the intake of nutrients to transform them into components which restores depleted resources. Similarly, multicellular organisms require the intake of food where digestion breaks down substances into components to be used by the body and its physiological processes. Thus, any damage or disruption caused by the depletion of components can be restored, provided individuals are able to consume the necessary resources which facilitate this restoration. Though the biological cell was the inspiration for autopoiesis,

446. Varela, Maturana, and Uribe, ‘Autopoiesis’, 189.

the concept transcends many domains, from individuals to social dynamics,[447] although this remains contested.[448] Overall, any system which is characterized by a set of relations which give rise to self-organization and autonomous functionality is said to be autopoietic.

To acquire nourishment, sensory mechanisms enable it to detect environmental changes associated with nutrients, and later in the phylogenetic tree, an ability move around the environment and seek food.[449] Early lifeforms without an ability to move, such as sea sponges, were still able to detect the presence of food. Organisms without sensory mechanisms to detect food were left relying on chance encounters to acquire the necessary resources to survive. Organisms without these mechanisms risk a higher degree of failure to meet their needs, increasing the likelihood of dying as a result. Even if the individual managed to reproduce, its offspring would likely share these missing capacities and thus would not fare as well as individuals inheriting mechanisms for the detection of food.[450] Presumably, an ability to detect environmental changes associated with acquiring food and other necessities to sustain physiological processes was selected for by evolutionary processes. Organisms lacking sensory mechanisms were more likely to die off prior to reproduction and go extinct as a species. Thus, these sensory mechanisms contributed to autopoiesis, as an ability to detect food assists in an organism's ability to continue to self-organize. As a result, these mechanisms were selected for by evolutionary processes.

447. Luisi, 'Autopoiesis', 52; Maturana and Varela, *Autopoiesis and Cognition*, xxiv.
448. Luisi, 'Autopoiesis', 57.
449. Feinberg and Mallatt, 'Phenomenal Consciousness and Emergence', 5.
450. Feinberg and Mallatt, 6.

As such, a degree of environmental awareness is required for optimizing the ability to find nutrients and other necessary resources like mates and shelter.[451] Hence, species with sensorimotor mechanisms became capable of traversing their environment, thus maximizing the chances of satisfying their requirements. As such, and over time, a rudimentary environmental awareness emerged, pertaining to both internal bodily states and external awareness of environmental changes.

To facilitate the detection of environmental changes, a specialized cell evolved to respond rapidly to stimuli.[452] The neuron and interconnected neural networks evolved to generate reflexes, automatic responses to stimuli which rapidly transmit information to generate basic motor patterns.[453] The firing of one neuron would transmit a signal to nearby neurons which caused parts of the body to move about, a response which facilitates the acquisition of resources or escaping dangerous situations. At first, these networks were distributed throughout the body without a central structure like a brain, however, over time central nervous systems would develop.[454] As the bodies and sensory mechanisms of organisms developed from evolutionary processes, additional neuronal networks and structures were required to process the information being transmitted throughout the body.[455] The Cambrian explosion, occurring around 560 million years ago,[456] saw the diversification of animal species with advanced central nervous systems and sensorimotor programs for detecting and adapting to environmental changes.[457] From

451. Feinberg and Mallatt, 5.
452. Feinberg and Mallatt, 'The Nature of Primary Consciousness. A New Synthesis', 118.
453. Feinberg and Mallatt, *The Ancient Origins of Consciousness*, 25.
454. Feinberg and Mallatt, 'Phenomenal Consciousness and Emergence', 7.
455. Feinberg and Mallatt, 7.
456. Feinberg and Mallatt, *The Ancient Origins of Consciousness*, 51.
457. Feinberg and Mallatt, 64.

invertebrate species such as sponges and worms, vertebrates with central nervous systems evolved to better detect and capture prey animals,[458] culminating in ever improving sensory mechanisms related to vision, hearing, touch, and chemoreceptors for smell and taste.[459] While nociceptors for detecting pain likely developed prior to the Cambrian explosion to detect noxious or harmful stimuli, the mechanisms for pain and pleasure would also develop further during this period of time.[460] Additionally, cognitive capacities like memory and attention would continue to develop over this same period, eventually resulting in internal presentations of stimuli to better respond to elements of the environment.[461]

It is from this prodigious and diverse evolutionary history that humans and their sophisticated nervous systems emerged. Brains with large neocortices, neuronal structures integrating lower brain regions like the evolutionarily older *limbic system*, enabled the development of sophisticated cognitive abilities. Moreover, advanced neocortices introduced an improved capacity to create and use tools and think abstractly. Humans also benefited from our evolved social tendencies, allowing us to work together towards different shared goals. To maintain social harmony, however, more complex norms and practices were increasingly required, and as such, our communication skills and linguistic capacities also needed to evolve. The information generated from these practices, when combined with our ability to communicate, resulted in cultural information or ideas which could be transmitted to others. One

458. Feinberg and Mallatt, 'Phenomenal Consciousness and Emergence', 7.
459. Feinberg and Mallatt, 'The Nature of Primary Consciousness. A New Synthesis', 118; Feinberg and Mallatt, *The Ancient Origins of Consciousness*, 97.
460. Feinberg and Mallatt, *The Ancient Origins of Consciousness*, 79.
461. Feinberg and Mallatt, 'The Nature of Primary Consciousness. A New Synthesis', 118.

way to describe this phenomenon has been through the use of the word *memes*, as coined by biologist Richard Dawkins in his book *The Selfish Gene*.[462] Dawkins postulated that as units of information, memes mutate over time when passed between individuals and groups, evolving similarly to genetic material through replication. Since the introduction of the term, however, a debate has emerged over the details of memetics,[463] such as the nature of their replication.[464] Nonetheless, the general notion of the spreading of ideas as a significant factor for human cultures and societies helps explain the development of humans and their societies. Together, the evolution of physical bodies and spread of cultural information further developed capacities and innovations like tool creation and social practices, resulting in developments like agriculture, legal systems, art, and advanced technology.

In 1991, Varela published *The Embodied Mind* with coauthors Evan Thompson and Eleanor Rosch, generating a new conception of minds and behaviour by establishing connections to autopoiesis. Inspired by the writings of French phenomenologist Maurice Merleau-Ponty, their discussion of minds and behaviour is one which considers both biology and phenomenology.[465] Doing so provides a perspective which considers minds as *enactive*, in which cognition arises from the interaction of the body and the world in which it lives.[466] To motivate this stance, the first chapter provides a quote from Merleau-Ponty which provides a glimpse into his contributions to phenomenology:

---

462. Dawkins, *The Selfish Gene*, 192.
463. Mesoudi, 'Cultural Evolution', 489.
464. Henrich, Boyd, and Richerson, 'Five Misunderstandings About Cultural Evolution', 124.
465. Varela, Rosch, and Thompson, *The Embodied Mind*, xv.
466. Varela, Rosch, and Thompson, 9.

> Perception is not a science of the world, it is not even an act, a deliberate taking up of a position; it is the background from which all acts stand out, and is presupposed by them: The world is not an object such that I have in my possession of the law of its making; it is the natural setting of, and field for, all my thoughts and all my explicit perceptions.[467]

Cognition is an evolved capacity from autonomous, living organisms and their interest in the world, given that the environment impacts the self-organizing processes which constitute their being. Experiences of the world are a product of an awareness of the ways in which the body and its physiological systems detect and respond to changes in the environment.[468] The mind and its contents result from this primordial awareness, in which acts of reflection further allow individuals to better understand the qualities of these experiences. As such, introspection assists individuals in understanding the relationship between their bodies and mental contents.[469] By focusing one's attention on what the mind is doing as it is occurring, greater insight on the human experience can be gained.[470] As a discipline, phenomenology analyzes and describes the contents of experience as they are perceived from the first-person perspective.

Rather than conceiving of a mind as an input-output machine, it should instead be understood as emerging from a network of physiological processes, all of which self-organize for the sake of responding to environmental demands. Self-organization occurs on many biological levels,[471] from cells to organs to systems of organs working together, in which this nesting produces a multilayered, complex system which operates as a unity or integrated whole. In animals, cognition is produced by the nervous system and aims to maintain this multi-level self-

---

467. Varela, Rosch, and Thompson, 3.
468. Varela, Rosch, and Thompson, 174.
469. Varela, Rosch, and Thompson, 28–29.
470. Varela, Rosch, and Thompson, 23.
471. Varela, Rosch, and Thompson, 196.

organization through the avoidance of harms, along with an attraction toward beneficial resources and activities. For example, feelings of hunger motivate the intake of food to replenish the nutrients and resources required to maintain autopoiesis. As a result, we see a *structural coupling* or a close and inseparable relationship between the functional structure of living organisms and regularities within the environment.[472] As organisms evolve, their physiology shifts to meet environmental demands, and as such, these outcomes are closely linked to elements of the environment.

A tension has long existed between two modes of study: those oriented from a first-person perspective and ones adopting a third-person perspective. By considering both domains, the idea of enactive minds describes the connection between inner and outer, subjective and objective. Verifiability is a key component in a third-person perspective, as aimed at by scientific disciplines. Given this, however, one's *lived experience* cannot currently be examined nor verified by external means. That said, by systematically describing the way subjective experiences appear from the first-person perspective, we can understand the ways in which these experiences are predicated on the body's morphology and its physiological systems. The domain of phenomenology, as introduced in Chapter 2, is thus important to consider to better understand human minds and their behaviours. A significant concept which has greatly influenced phenomenology is Martin Heidegger's *Dasein* or "being-in-the-world."[473] It describes humans as a particular kind of entity, one which is inseparable from the world in which we live as self-

472. Varela, Rosch, and Thompson, 151.
473. Varela, Rosch, and Thompson, 19; Wheeler, 'Martin Heidegger', sec. 2.1.3.

aware beings.[474] Because human cultures are a part of this wider environment, each human individual understands itself as existing within some culture and participating in a specific way of being.[475] In contrast, scientific examinations of humans and their behaviour aim to understand "context-free elements, cues, attributes, features, factors, primitives, etc., and relate them through covering laws, as in natural science and behaviourism, or through rules and programs as in structuralism and cognitivism."[476] Though the sciences and humanities including phenomenology begin from distinct vantage points on questions related to human life and behaviour, both must be considered to sufficiently address these topics of inquiry. While phenomenology and its postulates cannot be reduced to and be fully articulated in scientific terms, a connection can be identified and articulated by concepts like autopoiesis. It suggests that through a variety of physiological self-organizing processes within the body, subjective experiences manifest as internal presentations of elements of the external world. Since biological organisms are capable of responding and adapting to environmental demands, their bodies and behaviours are products of environmental pressures and regularities influencing physiological processes.

Evan Thompson would further the connection between phenomenology and biology in *Mind in Life* published in 2007. Thompson argues that phenomenology needs to be able to interpret and understand its inquiry in relation to scientific domains, and that mind sciences can be informed by the analysis of phenomenological considerations.[477] By appealing to autopoiesis,

474. Dreyfus, *Being-in-the-World*, 14–15.
475. Dreyfus, 24.
476. Dreyfus, 2.
477. Thompson, *Mind in Life*, 14.

he explains how the mind emerges from self-organizing processes which interconnect the brain, body, and environment, resulting in a self-determining individual who remains nevertheless continuously involved with its environment.[478] As the individual is dependent on aspects of the world for its continued survival, and is harmed by some of its features and contents, value and meaning is established for individuals.[479] Sugar, for example, gains its value as food and fuel source because it is registered as a nutrient which satisfies an organism's metabolic needs.[480] The evolved pleasant experiences of consuming sugar indicate a beneficial element to interacting with this chemical, motivating individuals to consume more of it to meet their biological needs.

Included in the environment, for humans and some animals, is a social element which adds complexity to the dynamic between internal and external environments. The actions and inactions of others have the ability to impact the way individuals experience the world and subsequently behave within it.[481] This added complexity is due to the way cultural practices and norms evolve over time. Since culture involves systems of information, the interaction between individuals and groups, along with their ideas and practices, further contributed to the complexity of cognitive capacities, behaviours, and cultural products like technology.

Over time and as a result of the work of Varela and others like Thompson, a growing understanding and appreciation of environmental influences on the body shifted conceptions of the mind and brain, giving rise to *embodied cognition* as an approach to scientific research.[482] From this perspective, individuals and their behaviours are seen as outcomes of physiological

---

478. Thompson, 37.
479. Thompson, 153.
480. Thompson, 74.
481. Thompson, 410.
482. Shapiro and Spaulding, 'Embodied Cognition', sec. 1.2; Thompson, *Mind in Life*, 10.

systems fluctuating in response to shifting environments. Social and cultural factors also play a role in shaping human thinking and actions taken in response to environmental conditions. Rather than conceiving of the mind as a separate entity which governs the body, as proposed by Descartes, contemporary views see the mind as emerging from bodily processes over the course of human evolution and through childhood development, occurring in particular sociocultural contexts and physical environments. Thus, behaviour is not the outcome of a biological computer, but the organism as a complex system of interrelated subsystems.

Therefore, an organism's awareness and ability to experience the world emerged through evolutionary processes to facilitate the individual's continued survival, in which these capacities are generated from interconnected, self-organizing physiological processes. The capacity to detect and respond to environmental changes facilitates the continued functioning of necessary physiological processes. One form of sensorimotor functionality is the association of pain and pleasure signals to specific stimuli, consequently imbuing them with meaning and significance for the sake of the individual's survival. It is this association between stimuli and their effects on the body and its processes which gives language and symbols their meaning, as they refer to objects or phenomena which have the potential to impact these self-organizing processes. This brief discussion of autopoiesis and biological development aims to explain and motivate the proposed solution to building a robot capable of machine empathy. The next section continues the story of Varela, as he was indirectly related to developments in robotics which gave rise to iCub. Given the issues identified with the aforementioned embodied robot iCub, a new approach is required, one which takes *embodiment* further to ground the meanings of emotions.

5.2 Self-Organization in Machines

The principle of self-organization has been applied to machines and software, first beginning with cybernetics in the 1940's and later in a domain called *artificial life* (ALife). As mentioned in the previous chapter in the section on the history of AI, cybernetics played an important role in the history of AI with the McCulloch Pitts neuron, a mathematical model of neuronal activity. In fact, cybernetics inspired Varela while working on his PhD at Harvard during 1968 and 1969.[483] Varela would eventually go on to assist in the development of ALife as a new domain of research, and contributions made by Rodney Brooks would introduce embodiment to AI. Despite the promising directions proposed by cybernetics and ALife, the significance of emotions for embodiment was overlooked, and as such, led to the development of robots without the ability to understand what emotions are. This section returns to the mid 20th century with an examination of cybernetics prior to presenting a brief history of ALife to demonstrate the promising, albeit insufficient, direction robotics has pursued since the advent of the McCulloch Pitts neuron. While this direction led to the creation of iCub, it nonetheless failed to notice the significance of emotion for behaviour.

As mentioned in the beginning of the previous chapter, in 1948 and before the advent of AI, Norbert Wiener published *Cybernetics: or Control and Communication in the Animal and the Machine*. 'Cybernetics' comes from the Greek word *kybernetes* which means "steersman"

483. Luisi, 'Autopoiesis', 50.

and refers to the person steering a boat or ship.[484] The idea is that this role watches to see where the ship is headed and corrects the course accordingly by using visual feedback. The word *governor* is also derived from *kybernetes*.[485] Thus, cybernetics concerns itself with mechanisms for control and regulation through relations and the flow of information.[486] Instead of focusing on what systems are made of or what the system *is*, cybernetics is interested in what systems do.[487] This interest in functionality and behaviours means that cybernetics can be applied to a number of domains, from engineering to biology to sociology.[488] As such, the domain has influenced many disciplines in modern sciences, including computer science, information theory, cognitive science, artificial intelligence and artificial life.[489]

The specific functionality of interest to cybernetics is circularity, in which feedback processes use system outputs or effects as inputs for the system.[490] The systems characterized by these circular processes are self-regulating and able to maintain their stability through feedback loops.[491] Given its focus on functionality, cybernetics is interested in the similarities between machines and living systems, in which autonomous agents are capable of regulating their own behaviours through circular processes.[492] To do so, these *automata* receive information from the external environment and use it to perform actions, and are thus *coupled* to the world in which

484. Wiener, *Cybernetics*, 18.
485. Wiener, 18.
486. Heylighen and Joslyn, 'Cybernetics and Second-Order Cybernetics', 157.
487. Ashby, *An Introduction to Cybernetics*, 1.
488. Heylighen and Joslyn, 'Cybernetics and Second-Order Cybernetics', 155.
489. Heylighen and Joslyn, 156.
490. Heylighen and Joslyn, 160.
491. Johnston, *The Allure of Machinic Life*, 26.
492. Heylighen and Joslyn, 'Cybernetics and Second-Order Cybernetics', 156.

they exist.[493] Elements of the environment play a significant role in governing the behaviours of automata, as the responses generated are in relation to these environmental factors.

An example of an automatic device is a thermostat which uses sensors to measure air temperature.[494] The reading is compared to the set threshold, and heat is generated if the value is below the desired temperature. Once the heat has increased to meet the threshold, the thermostat detects the change and shuts the heater off. If the temperature drops below the threshold, the thermostat repeats the process automatically. The difference between the actual air temperature and set threshold is deemed the *error* and the thermostat's behaviour aims to reduce this value. This functionality is called *negative feedback* and is used by the system to generate or maintain a specific outcome by adjusting behaviours which reduce the error value. This phenomenon also observed in the Watt governor,[495] a spinning device which regulates the amount of steam passed into a steam engine. It is comprised of two metal balls attached to upper and lower links, where these links are secured to a central sleeve which raises and lowers over an inner spindle as a result of the spinning action.[496] Since the governor is attached to a valve, the centrifugal forces acting on the metal balls pulls the sleeve upward, subsequently closing the valve to reduce the amount of steam passed to the engine.[497] As the spinning slows, the central bar drops down causing the valve to open. In contrast to negative feedback, *positive feedback* increases the difference or *error* between two states or values.[498] An example of positive feedback is the noise

493. Wiener, *Cybernetics*, 62.
494. Wiener, 131.
495. Wiener, 132.
496. Hasan, 'Comparative Study of Watt, Porter, Proell and Hartnell Governor Mechanism', 481.
497. Hasan, 482.
498. Black, 'Stabilized Feedback Amplifiers', 5.

generated from a speaker when a nearby microphone picks up and amplifies the sound produced by the speaker.[499]

In humans and animals, the nervous system acts as a mechanism for the regulation of behaviour. As an integrated whole, this physiological system monitors behavioural outputs and adjusts motor actions accordingly based on the difference between desired outcomes and actual outcomes.[500] This includes an array of sensory mechanisms, including an organism's skin and body, both of which provide persistent feedback to self-regulate the organism's own behavioural responses.[501] For example, a cat can learn to pounce by noting the difference between its desired end-state, say leaping on top of a mouse, and its actual end-state, which could include falling short and missing it, allowing the mouse to escape. In this case, visual and tactile cues are processed by the cat's nervous system and it is thus able to adjust its muscle movements on the next jump, reducing the error until it lands successfully on the mouse. Though this form of feedback and adjustment is notably different than engineered systems which can modify their behaviours as they are in operation, it nonetheless demonstrates a way in which bodily systems use negative feedback for corrective measures. The cat must make additional jumps to reduce the error value, in this case the distance from the target, and is not able to correct its behaviour mid-jump. Through trial and error, however, the cat's learning mechanisms enable it to close the distance between where it lands and where it wants to land. In this way, negative feedback is

499. Hofstadter, *I Am a Strange Loop*, 54.
500. Wiener, *Cybernetics*, 130.
501. Wiener, 12.

used by organisms to correct behaviours, accomplishing this in a slightly different way than devices like thermostats and governors.

In *The Human Use of Human Beings*, Wiener includes a discussion on the significance of affective states which motivate behaviours in animals. Referring to Ivan Pavlov's influential study on associative conditioning, he notes that animal behaviour is often motivated by "emotion" which usually concerns food.[502] Although many stimuli unassociated with food like certain sounds and visual cues do not produce a behavioural effect, a connection between the two may be established provided the stimulus is presented at the same time. The resulting association can be learned from a number of concurrent presentations, leading the animal to respond to the stimulus alone without the presence of food. Pavlov's research led Wiener to conclude that these associations are created from "something important to the life of the animal: in this case, food," though it is not required to generate the behaviour once the association has been learned.[503] To reinforce his claim, he also discusses the same phenomenon with respect to pain by appealing to an electric fence to enclose cattle. Although a cow is strong enough to break the fence and escape, it does not because the pain generated from electric shock deters it from touching the fence.[504] The cow learns the association between the fence and its painful effects, influencing its interest in escaping from the enclosure altogether. This discussion is interesting and worth mentioning because it highlights the significance of affect, which Wiener calls 'emotion', for generating and influencing motivation. Affect directs attentional processes to determine which aspects of the environment are salient to intentional agents like humans and animals. Therefore,

502. Wiener, *The Human Use of Human Beings*, 68.
503. Wiener, 68–69.
504. Wiener, 69.

to recreate humanlike abilities, affect and motivation should be included in the design of social robots. The section which follows describes this functionality in further detail.

Cybernetics would eventually lead to the creation of the domain of *artificial life* in the late 1980's.[505] Though the concept of self-organizing systems had been introduced to cybernetics by Heinz von Foerster in the late 1950's,[506] the implementation of these systems in computer software would not arise until the development of modern computers.[507] After the discovery of DNA by Watson and Crick in 1953, a student of von Neumann, Arthur Burks, would develop abstract logical structures of self-reproduction in *cellular automata*.[508] Consisting of a two-dimensional array divided into square cells, cellular automata would be programmed with a set of rules for self-reproduction, in which cells would grow into unoccupied adjacent spaces within the array. Development on the mathematics of cellular automata would continue with the help of von Neumann and others, including Edgar Codd and John Conway.[509] Conway would go on to create the well-known "Game of Life" which, using specific configurations, aimed to increase the number of live cells within the grid space.[510] In 1987, Christopher Langton would hold a conference on the "Synthesis and Simulation of Living Systems" which introduced a number of models for living systems, leading to the developing of ALife.[511] The approach used in creating cellular automata was a novel development for its time. This bottom-up approach directly

505. Johnston, *The Allure of Machinic Life*, 165.
506. Johnston, 54.
507. Johnston, 165–66.
508. Johnston, 169.
509. Johnston, 170.
510. Bays, 'Introduction to Cellular Automata and Conway's Game of Life', 4.
511. Johnston, *The Allure of Machinic Life*, 171–72.

contrasted the top-down, rule-governed functionality of symbolic AI which was still popular during this period.[512]

In 1979, Varela would publish *Principles of Biological Autonomy* which included a discussion of cellular automata as a model for simple autopoietic systems capable of creating and repairing self-enclosing boundaries.[513] These formulations would provide an alternative basis for ALife, and Varela would credit von Foerster as inspiration for his and Maturana's work on autopoiesis. Thus, congruency between animals and mechanical artifacts had been identified in practice, just as Wiener's book title theorized forty years prior.

In 1991, Varela and Paul Bourgine would organize a conference titled the "First European Conference on Artificial Life"[514] with the proceedings subsequently published and titled *Towards a Practice of Autonomous Systems*.[515] Included in these presentations was one by Rodney Brooks, asking whether techniques from ALife could be useful for generating programs to control robots by mimicking evolution.[516] The reason for this suggestion was because robots at that time were unable to learn to adapt to dynamic environments, and new behaviours needed to be specified in code. Given that ALife had developed techniques for evolving new programs for robots in simulated environments, Brooks was interested in applying these techniques to physically embodied robots. Of particular interest was *genetic programming*, in which behaviours could be developed in simulated environments and later implemented in physical

512. Johnston, 173.
513. Varela, *Principles of Biological Autonomy*; Johnston, *The Allure of Machinic Life*, 188.
514. Johnston, *The Allure of Machinic Life*, 189.
515. Johnston, 198.
516. Brooks, 'Artificial Life and Real Robots', 3.

versions of the robot.[517] Genetic programming involves the incremental design and implementation of behaviours, where a minimal amount of architecture is written and tested before adding the next-simplest layer of behaviour.[518] Since robots are required to adapt to their environment via machine learning, control structures could be developed through evolutionary means in simulated environments,[519] removing the need of performing large numbers of "runtime trials" in physically-embodied robots.[520] Machine learning, in this case, was generated through symbolic approaches to AI using the programming language LISP,[521] unlike the machine learning techniques of today which involve neural networks in robots like iCub.

Despite their shared interest in ALife, Langton and Varela held contrasting points of view on the discipline's primary characteristic and research motivation. Langton's interests were centred around framing life in terms of information processing, in which a specific configuration of information reproduces itself to sustain life.[522] Varela, on the other hand, considered information processing as a secondary factor of a system's organization, working as an allopoietic submachine within self-organizing dynamics. Instead, Varela considers the most fundamental capacity of living systems to be autonomy.[523] In the front matter of *Towards a Practice of Autonomous Systems*, he states that autonomous systems are better defined by the ways in which they act within the world and are shaped by environmental factors.[524] The reason

517. Brooks, 4.
518. Harvey et al., 'Evolutionary Robotics', 80.
519. Brooks, 'Artificial Life and Real Robots', 9.
520. Brooks, 3–4.
521. Brooks, 6.
522. Johnston, *The Allure of Machinic Life*, 197.
523. Johnston, 198.
524. Johnston, 198; Bourgine and Varela, 'Toward a Practice of Autonomous Systems', xi.

for this framing is because it corresponds to a variety of different autonomous systems, from cells and animals to human societies. Moreover, this characterization moves beyond the "disembodied abstractions" of information processing to consider the situated nature of these systems.[525] Because the environment and its features play a crucial role in the functioning and organization of living systems, these factors are a significant component of the way living systems are considered. In contrast, Langton's concerns were oriented toward an excessive focus on computational approaches which generate formal abstractions of life. This approach separates living beings from their physical instantiation, ignoring the embodiment of a system and its *situatedness* within an environment.[526]

While initially, ALife focused on computational processes enacted in simulated environments, research today involves both Langton's and Varela's approaches to studying life.[527] To implement physically embodied systems, an understanding of the requisite computational processes is required; however, simulated environments remain distinct from the physical world. As such, ALife research is divided into three categories: hardware, software, and "wetware" or artificial cells synthesized in a laboratory.[528] While software approaches to ALife rely on computer simulations of cells and organisms, hardware approaches involve robots which learn to navigate and respond to their environment, as seen in the work of Rodney Brooks. His contributions to ALife would eventually become adopted by AI researchers as a way for robots

525. Bourgine and Varela, 'Toward a Practice of Autonomous Systems', xi.
526. Johnston, *The Allure of Machinic Life*, 200.
527. Johnston, 200.
528. Bedau, 'Artificial Life', 296.

to adapt to and learn about dynamic environments, leading to the development of embodied robots like iCub.[529]

This shift in directions for AI development also saw support from AI critic Hubert Dreyfus, particularly in his book *What Computers Can't Do* originally published in 1972. Sceptical of the excitement and hype around symbolic-reasoning approaches to AI, Dreyfus's first publication was a report written for the RAND Corporation in 1965 titled 'Alchemy and Artificial Intelligence'.[530] In it, he suggests that human problem-solving is not reducible to the kinds of discrete operations used in symbolic-reasoning. Furthermore, he argues that digital computers are not simulating the manner in which the brain functions but instead, approximate the outcomes of human intelligence through discrete operations such as heuristic search.[531] These ideas would be further developed and explained in *What Computers Can't Do*, where Dreyfus recommends a new approach to AI based on his study of phenomenologists like Heidegger and Merleau-Ponty.[532] Because much of human knowledge is based on *know-how*, or "inarticulate, preconceptual background understanding of what it is like to be a human being," any attempt at formalizing this know-how into symbolic representations appeared to be a "hopeless task."[533] The alternative Dreyfus suggests is one which produces machine intelligence from experience, learning, and the acquisition of skills.[534] In *What Computers Still Can't Do*, published in

529. Johnston, *The Allure of Machinic Life*, 347–48.
530. Dreyfus, 'Alchemy and Artificial Intelligence'.
531. Dreyfus, 63.
532. Dreyfus, *What Computers Still Can't Do*, xi.
533. Dreyfus, xi–xii.
534. Dreyfus, *What Computers Can't Do*, 215.

1992,[535] Dreyfus states that although neural networks offer a more promising approach to AI development, a new problem emerges into view.[536] He states "One needs a learning device that shares enough human concerns and human structure to learn to generalize the way human beings do. And as improbable as it was that one could build a device that could capture our humanity in a physical symbol system, it seems at least as unlikely that one could build a device sufficiently like us to act and learn in our world."[537] While Dreyfus appears to see the benefit to approaching AI from a perspective rooted in embodied cognition, he suggests that something else might be required, something akin to "human concerns."

My suggestion is that this missing element consists of an analogue of emotion for machines to support a form of self-organization. This version would, at the most basic level, involve the manifestation of behaviours derived from experiences of pain and pleasure. Additionally, learned associations between these sensations and the various stimuli agents encounter within the environment would contribute to self-organization. As the robot establishes these associations, it organizes its behaviour around aspects of the environment and their impact on its functioning, seeking out stimuli which generate pleasure and avoiding those which cause pain. From this foundation of learned associations, a form of "inarticulate, preconceptual

535. Dreyfus, *What Computers Still Can't Do*, iv. The contents of this publication are almost identical to its older version titled *What Computers Can't Do*. In the "Introduction to the MIT Press Edition" on page ix, Dreyfus states that this edition "marks not only a change of publisher and a slight change of title; it also marks a change of status. The book now offers not a controversial position in an ongoing debate but a view of a bygone period of history."
536. Dreyfus, xlv.
537. Dreyfus, xlv–xlvi.

background understanding"[538] could emerge, generating automatic or reflex-like responses to specific stimuli.

At this point, it would be reasonable to inquire about the differences between hardware approaches to ALife and the development of robotics in AI, as the two domains appear to be rather similar. Though there is considerable overlap, hardware-based ALife "explicitly and extensively" applies inspiration from all forms of life, not just humans.[539] One approach has involved the use of supercomputers to coevolve a robot's morphology with its functional controllers, coupling its body and functionality to produce behaviours, where a three-dimensional printer subsequently builds a physical version of the robot.[540] Another approach investigates ways for robots to continuously diagnose and repair damage to their bodies.[541] While some AI robots may have drawn inspiration from ALife, AI approaches are generally concerned with creating functionality to bring about a specific behavioural output, like picking up objects or detecting features.[542] Robots designed from an ALife perspective, however, would instead be designed in such a way where behaviours are generated to achieve a specific outcome *for* the machine's needs or functioning. In living beings, if an organism is hungry, it regulates its behaviour for the sake of acquiring food, where the food intake is the input which satisfies an internal motivation to eat. Because living beings are motivated to act in order to fulfill specific goals which satisfy internal needs, robots built to mimic living systems should follow suit. A

538. Dreyfus, xii.
539. Bedau, 'Artificial Life', 297.
540. Pollack et al., 'Three Generations of Automatically Designed Robots', 216.
541. Bongard, Zykov, and Lipson, 'Resilient Machines Through Continuous Self-Modeling', 1118.
542. Young, 'A General Architecture for Robotics Systems', 238.

robot's goals should focus on satisfying their own needs and desires as indicated by their perceptual systems.[543] Likewise, if the robot's perceptual system indicates an environment is too noisy, for example, it modifies its behaviour to reduce the amount of noise entering its auditory-processing channels.

To overcome the limitations presented in Chapter 3.3, social robots should be developed in a manner which gives rise to experiences, particularly emotional experiences. Given the significance of experience for providing an organism with information about the physical world and the ways it impacts the body, social robots should be built with a form of subjective experience. This would enable them to become analogously invested in their own continued functioning. Included in this requirement is an ability to feel emotions and various affective states, as emotions infuse stimuli with meaning and significance which the agent uses to act and make decisions. This aspect of biology is currently missing in current robotic applications like iCub, an omission which results in both a lack of interest in its own survival, and subsequently, a lack of understanding of words related to emotions.

Over these last two sections, some important features of machines have been identified for recreating central aspects of living beings. With these ideas in mind, we can now analyze a new strategy and understand why it constitutes a better approach than iCub for generating social robots. Though the creator's interest resides in machine consciousness, a topic beyond the scope of the present work, it nevertheless provides a good alternative due to its explicit appeal to theories generated from empirical evidence. Specifically, it provides a method for creating

543. Young, 259.

machine emotions in an analogous manner to animals, where these emotions can become associated with stimuli the robot encounters. Through this associative mechanism, a robot is able to learn about the world and the ways in which it impacts its own body and functioning, imbuing the robot with subjective experiences.

5.3 A Candidate Solution

The new direction discussed in this section has been developed by Pentti Haikonen, the Principal Scientist on Cognitive Technology at the Nokia Research Center in Helsinki, Finland, from 1991 to 2009.[544] His particular interest is in consciousness and its implementation in machines, in which 'consciousness' refers to an agent's ability to report on perceptions of stimuli from internal and external environments.[545] Given this interest, Haikonen has deliberately added experience into his cognitive architecture. His solution uses associative learning, memory, and evocation to establish connections between incoming sensory information from signals generated by sensors on the robot body. By incorporating signals which are analogous to pain and pleasure in animals, his architecture enables the robot to determine what stimuli mean through how they impact its body.

544. University of Illinois Springfield Philosophy Department, 'Faculty & Staff'. Haikonen is listed under the heading 'Adjunct Faculty' and has supplied his curriculum vitae which includes his tenure as Principal Scientist at the Nokia Research Center.
545. Haikonen, *Consciousness and Robot Sentience*, 42.

Haikonen also appeals to differences between digital computers and biological brains to argue that all meaning is ungrounded in robots governed by software running on digital computers. Although further explanation of this argument requires a lengthy discussion beyond the scope of this work, it is worth mentioning here as a significant reason for Haikonen's motivation for creating this new architecture. For our purposes, however, we are interested in the Haikonen Cognitive Architecture (HCA) for its inclusion of emotion at a fundamental level of robotic operation.

According to Haikonen's perspective on cognitive machines, digital computers should not constitute the foundation for social robots because computers operate only using syntactical structures, resulting in the *symbol grounding problem*.[546] This problem is due to the fact that the meanings of these symbols are explained by other symbols, rather than "self-explanatory information or real-world phenomena."[547] Since symbols are explained by other symbols, the reference becomes circular and is thus uninformative and ungrounded. The reason is due to the fact that digital computers only operate on syntactic structures and rule-based operations without appealing to semantics. There is no understanding in a digital computer because it cannot assign meaning to the words it receives and displays, an idea also mentioned by Varela, Thompson, and Rosch in *The Embodied Mind*.[548] Since computers are not structured as organizational unities, the execution of any computer algorithm will not be sufficient to produce the experiences

546. Haikonen, 6.
547. Haikonen, 13.
548. Varela, Rosch, and Thompson, *The Embodied Mind*, 41.

necessary for consciousness. As such, a robot with an ability to feel and experience the world must be built in a way which mimics the self-organizing structures observed in living beings.

Before discussing the specifics of this approach, it is vital to briefly discuss what model neurons are and how they function. A neuron is one type of cell located in the brain and together, via the central nervous system, neurons are responsible for sending and receiving signals throughout the body.[549] To accomplish this, a neuron *fires* and transmits trains of electric pulses, provided the amount of incoming energy from one or more neurons exceeds a specific threshold. At the beginning of this process, *dendrites*, which resemble tree branches extending outward from the neuron's *cell body*, receive incoming signals in the form of chemical compounds called *neurotransmitters*.[550] With sufficient stimulation from neurotransmitters, a chemical reaction occurs at the *axon hillock* located at the base of the cell body, transmitting the signal through the tail-like structure of the neuron called the *axon*. This transmission occurs from a series of electrochemical reactions to propel the signal along the axon, until it reaches the *axon terminals* and releases neurotransmitters into the *synapse*. The synapse is the space between neurons where the dendrites in the neighbouring neuron detect their presence, causing it to alter its pattern of activity if sufficiently stimulated.[551] Though neurons maintain a base-rate of firing, the neuron itself acts in an *all-or-nothing* manner, firing at the same strength every time.[552] Neuronal signals

549. Krause et al., *An Introduction to Psychological Science*, 93.
550. Krause et al., 94.
551. Krause et al., 96.
552. Krause et al., 97.

are thus generated when their base-rate of firing changes in response to the presence of neurotransmitters.

There are a variety of different types of neurotransmitters. While some create an excitatory effect in the surrounding neurons, increasing their likelihood of firing, other types of neurotransmitters have an inhibitory effect.[553] These neurotransmitters thus decrease the likelihood of the firing of adjacent neurons, playing an important role in facilitating relaxation and a decrease in arousal. An example of one such type of neurotransmitter is *gamma-amino butyric acid* (GABA), significant for both inducing sleep and promoting deeper, slow-wave periods of sleep.[554]

As one neuron's firing leads to the firing of adjacent neurons, these neurons grow toward each other, leading to an increased likelihood of firing in the future.[555] This provides a significant mechanism for learning, first discovered by Canadian neuroscientist Donald Hebb in 1949.[556] This phenomenon has since been named Hebb's Law or *Hebbian learning* and is often summarized by the phrase "cells that fire together wire together."[557] This mechanism is responsible for the associative learning observed by Ivan Pavlov in the early 1900's in his experiments with dogs and food, as mentioned by Wiener and noted in the previous section of this chapter. Pavlov discovered that dogs would be *conditioned* or taught to salivate in response to hearing a tone, provided the same stimuli had been previously introduced as the dogs were

553. Krause et al., 98.
554. Gottesmann, 'GABA Mechanisms and Sleep', 235.
555. Hebb, *The Organization of Behavior*, 62.
556. Krause et al., *An Introduction to Psychological Science*, 23.
557. Haikonen, *Consciousness and Robot Sentience*, 90.

given food.[558] After a number of concurrent instances of presenting food with the tone, the dogs would no longer require the presence of food to begin salivating; the tone alone was sufficient. Thus, the sound became associated with being fed, initiating salivation. These experiments indicated the power of association, as two stimuli can be cognitively connected such that the presence of one stimulus evokes an association with another stimulus.[559] Forty years later, Hebb would discover the neural mechanism responsible for association.

Modified Hebbian learning is also responsible for machine learning in the cognitive architecture developed by Haikonen.[560] Rather than using numerical values to perform calculations, as observed in digital computers, the Haikonen Cognitive Architecture (HCA) is controlled by an assortment of threshold levels which affect the robot's behaviours and emotional state. Learning is established through the use of dedicated circuits representing synapses which are embedded within a larger electrical circuit called the Haikonen Associative Neuron (HAN). These associative neurons are designed to compare signal patterns generated by sensors on the robot's body. The HAN fires when an incoming signal, the *main signal*,[561] matches or sufficiently resembles a stored signal pattern, known as the *associative signal pattern*[562] or *vector*.[563] Both the main signal and associative signal pattern are comprised of a number of input signals representing *features* of a particular stimulus.[564] Each feature within a pattern is connected to a synaptic weight circuit.[565] For example, if an associative signal pattern

558. Krause et al., *An Introduction to Psychological Science*, 21.
559. Haikonen, *Robot Brains*, 17.
560. Haikonen, *Consciousness and Robot Sentience*, 90.
561. Haikonen, 'XCR-1', 361.
562. Haikonen, *Consciousness and Robot Sentience*, 92.
563. Haikonen, 'XCR-1', 361.
564. Haikonen, *Robot Brains*, 45.
565. Haikonen, 19.

consists of seven features, the HAN contains seven synaptic weight circuits. These features are determined based on a particular sensory modality, made up of parts which together, represent the qualities of the stimulus detected within the external environment. This will be explained in further detail shortly.

Within each of these synaptic circuits, the main signal is *correlated* with the associative input signal to produce a *weight* in the form of a binary value, either 1 or 0.[566] To do so, the main signal is multiplied by the associative input signal and the product is passed to an *accumulator* which stores a *correlation sum*, which will be compared to a preestablished threshold. If the product passed to the accumulator equals 1, the value of the correlation sum is incremented by a value of 1.5.[567] Alternatively, if the product passed equals 0, the correlation sum is decremented by a value of 0.5. The correlation sum is then compared to the threshold value and if it exceeds the value set, a switch is closed and the synaptic weight is set to 1. Initially, prior to any learning, the weight of the synapse is set to 0 and the switch is open. When the threshold is set to a low value, learning occurs quickly since the first instance of multiplication pushes the correlation sum over the threshold. This closes the switch, setting the weight to 1. At this point, the associative signal passes through the synapse and is multiplied by the synaptic weight.

The synaptic circuit also contains a *learn enable signal* in the form of a binary value. When set to 1, the learn enable signal is on and learning can occur when the main signal and associative signal occur at the same time. When learning is not enabled and the signal is set to 0,

566. Haikonen, 20.
567. Haikonen, 20.

learning cannot occur. This aims to control the occasions in which the robot is capable of learning, mimicking attentional processes in animals and humans.[568] If this signal were not included, new learning may conflict with previously learned information.

Just as the brain contains neurotransmitters which inhibit the firing of adjacent neurons, an inhibitory synaptic weight circuit has been developed to mimic this functionality in Haikonen's model. The *inhibiting neuron group* acts similarly to the GABA neurotransmitter by decreasing the likelihood of the surrounding neurons to fire.[569] This neuron group is useful for instances which require the robot to avoid performing an activity, for example, stopping its forward movement to avoid running into an obstacle.[570]

Returning to the functionality of the HAN, the neuron produces four binary outputs: the output signal representing the main signal, a match signal, a mismatch signal, and a novelty signal. The output signal is determined by summing the synaptic weights and comparing this value to the HAN's threshold. If the sum of the weights exceeds the value of the threshold, a match between the main signal and the associative signal pattern has been identified and the match value is set to 1.[571] At this stage, the output signal and match signals are set to 1, while the mismatch and novelty signals are set to 0. A mismatch is identified when the main signal does not match the associative signal pattern and the synaptic weights do not total a value greater than the threshold. In this case, the mismatch value is set to 1 while the output signal, match, and

568. Haikonen, *Consciousness and Robot Sentience*, 94.
569. Haikonen, *Robot Brains*, 46.
570. Haikonen, *Consciousness and Robot Sentience*, 216.
571. Haikonen, 118.

novelty signals are set to 0. Novelty is detected when a main signal exists without a corresponding associative signal pattern, and in this case, the novelty signal is set to 1 while the other signals are set to 0.[572]

As mentioned above, both the main signal and associative signal patterns entering into the HAN are comprised of features which are then compared.[573] The features of the main signal are determined by preprocessing circuits or *filters* and are separated into parts depending on the particular sensory mechanism. The robot's architecture would contain a number of sensory modalities, all represented by different sensors, such as audio or light sensitive receptors that provide sensory information about some aspect(s) of the external environment.[574] After the sensory information has been preprocessed by filters to detect features of stimuli, these signal vectors are passed to feedback neurons comprised of associative neuron groups to determine whether a match, mismatch, or a novel event is present. These associative signal patterns represent the robot's expectations or predictions based on previous learning. The stage of processing which involves the use of HANs is called the *perception/response feedback loop* as it determines the action the robot should take as an outcome of the match, mismatch, or novelty situation. Within this feedback loop, each extracted feature signal for a particular sensory module is compared to its corresponding associative signal.[575] Once this comparison is complete, the associated signal pattern is now considered a *percept* and is subsequently fed back into the same feedback neuron group of HANs. Here, it becomes a new "virtual" percept, one which depicts an

572. Haikonen, *Robot Brains*, 19; Haikonen, *Consciousness and Robot Sentience*, 118.
573. Haikonen, 'XCR-1', 362; Haikonen, *Robot Brains*, 72–73.
574. Haikonen, 'XCR-1', 363.
575. Haikonen, *Robot Brains*, 72–73.

internally evoked or "imagined" entity.[576] This percept is also broadcast to other neuron groups to become associated with other signal vectors, including ones representing distinct sensory modalities.[577]

To get a better sense of how this feedback loop gives rise to robot behaviours, an overview of the core sensory modalities is necessary. Mimicking animals and humans, the robot's cognitive architecture includes inputs for audio, visual signals, haptic information for a sense of touch, kinesthetic information about the robot's body, as well as specific sensors for representing pain and pleasure. Each sensory modality contains its own perception/response feedback loop, and these are cross-connected for learning and governing the robot's general behaviour.[578] The framework which constitutes the robot's "brain" is called the Haikonen Cognitive Architecture (HCA).[579] Interestingly, while the HCA aims to recreate human abilities, Haikonen notes this architecture could be altered to include additional sensory modalities absent in humans, including the sensing of electric, magnetic, or electromagnetic fields.[580] That said, as it exists currently, each modality will now described in further detail.

The HCA supports the use of one or more microphones[581] for detecting sounds in the environment. Incoming audio signals are first processed by a set of filters which perform frequency analysis by separating and isolating individual frequencies of a sound,[582] as well as the

576. Haikonen, 'XCR-1', 362.
577. Haikonen, *Consciousness and Robot Sentience*, 114.
578. Haikonen, *Robot Brains*, 180.
579. Haikonen, *Consciousness and Robot Sentience*, 155.
580. Haikonen, *Robot Brains*, 180.
581. Haikonen, 'XCR-1', 362; Haikonen, *Robot Brains*, 104–5.
582. Haikonen, *Robot Brains*, 100.

direction they are coming from.[583] Additionally, the temporal durations and the rhythms of sounds are also learned by associative neuron groups,[584] in which a circuit dedicated to autoassociative memory supports both prediction and the instant replay of sounds.[585]

For vision, the HCA captures and processes light through the use of hardware such as a digital cameras or photodiode sensors. Digital cameras can be used to create a two-dimensional array of *pixels* or picture elements which are assigned a numerical value proportional to the intensity of illumination.[586] For this pixel map, higher values indicate a higher amount of light present within that area of the image. Colour values are indicated by three additional pixel maps, one for each primary colour of light: blue, green, and red. To detect other visual features of an image, additional pixel maps are required, for example, pixel maps indicating line features, temporal changes, and motion.[587] Alternatively, the HCA could use two photodiodes which produce an electrical current from the absorption of photons.[588] The image of the target is projected on to the photodiodes through small lenses, and due to the parallax effect, the movement and direction of the target can be inferred from the relative amplitude of the electrical signals.[589] These output signals are then used to direct the motor actions of the robot, either approaching the target or avoiding it, depending on the robot's learning and the nature of the target.[590]

---

583. Haikonen, 105.
584. Haikonen, 101.
585. Haikonen, *Consciousness and Robot Sentience*, 129.
586. Haikonen, *Robot Brains*, 85.
587. Haikonen, 86.
588. Murari et al., 'Which Photodiode to Use', 753.
589. Haikonen, 'XCR-1', 363.
590. Haikonen, *Consciousness and Robot Sentience*, 211–12.

The actions performed by a robot utilizing the HCA depend on its inner state which is influenced by feelings and associations related to pain or pleasure.[591] The robot's "shock sensor" consists of a small magnetic earphone and is sensitive to vibrations within its body.[592] This sensor indicates the presence of a stimuli capable of causing damage to the robot, sending signals to the displeasure neuron group.[593] For pleasure, a separate "petting sensor" registers a touch and sends signals to a pleasure neuron group, acting as an intrinsic reward to reinforce behaviours.[594] Stimuli associated with rewards teach the robot to approach and further interact with the source of pleasure. On the other hand, stimuli associated with pain lead to situations of avoidance to reduce the negative impact the stimuli has on the robot itself. These pleasure and pain signals can also be associated with words like 'good' or 'bad', such that when these learned words are paired with specific stimuli, the robot can learn which stimuli to avoid and which to approach.[595] From these elementary signals related to pleasure and pain, the robot is able to develop its own versions of emotions, as suggested by Haikonen's Systems Reaction Theory of Emotions (SRTE).[596] This theory suggests that robot cognition and behaviours are predicated on its reaction to stimuli in the environment. An emotional evaluation of a stimulus initiates cognitive processes for performing actions, in addition to generating reactions and percepts about the stimulus itself. Machine emotions are created from combinations of machine sensations analogous to pain and pleasure in animals, along with other features of cognitive processing,

591. Haikonen, *Robot Brains*, 152.
592. Haikonen, 'XCR-1', 364.
593. Haikonen, *Robot Brains*, 152.
594. Haikonen, 'XCR-1', 364.
595. Haikonen, *Robot Brains*, 154.
596. Haikonen, 154.

such as mismatch and novelty.[597] For example, machine sadness is comprised of sensory mismatch and strong pain signals, while machine surprise follows from a sudden and large degree of mismatch.[598]

The robot's body also contains sensors related to its own movements and positions, along with perception/response feedback loops for comparing expected information to information actually captured from the environment. Sensors for kinesthetic information about the robot's body indicate its direction of movement and the robot's position within a space.[599] The robot is able to alter its movements and behaviours as a result of the match, mismatch, and novelty outcomes from feedback processes. Kinesthetic sensors are also used to produce information about the direction of the robot's gaze, allowing it to adjust its own position to better track objects and their movement.[600] Haptic feedback representing the robot's sense of touch is provided by sensors in its gripper "hand" to let the robot know when it has grasped an object.[601] The percepts generated from haptic processing also support the sensations of different shapes, textures, as well as the hardness or softness of an object. Haptic feedback is combined with kinesthetic and visual feedback to explore the properties of objects, in which sensors on the robot's body provide information about the object it is interacting with.[602]

The HCA and the various sensory modalities which contribute to the robot's abilities and functionality are separated into *modules*. Each module is responsible for processes related to

---

597. Haikonen, 155.
598. Haikonen, 156.
599. Haikonen, 82.
600. Haikonen, 87.
601. Haikonen, 83.
602. Haikonen, 83.

specific elements of the robot's functionality, where modules 1 and 2 specialize in motivations related to survival.[603] Specifically, module 1 contains the feedback loops for pain and pleasure, determining what should be considered "good" and "bad" to the robot based on learning about the effects of stimuli.[604] Module 2 focuses on information related to the robot's physical requirements, including sensors for its energy levels, motor drive levels, its temperature, balance, mechanical tension, among others.[605] The sensors in this module may send pain information to the first module to motivate certain actions. Module 3 includes haptic information and is related to self-image and sensing the environment, while module 4 includes information about the environment by interpreting visual data. Module 5 also captures information about the environment but through auditory processing, and it is in this module that speech and language are also processed.[606] The robot's direction and orientation is processed in module 6, while module 7 involves kinesthetic processing for the generation of motion and motor activities through percepts generated by other modules.[607] Each of these modules is capable of communicating with other modules, as percepts generated from feedback loops within each module are broadcast to other modules.[608] By generating more complex and multilayered associations between percepts from distinct modules, the robot is able to learn about its environment and how aspects of the physical world impact its own being and functionality. The fundamental role of pain and pleasure influence the robot's own internal motivations by informing it about which objects or environmental aspects should be approached and explored

603. Haikonen, 180.
604. Haikonen, *Consciousness and Robot Sentience*, 159.
605. Haikonen, *Robot Brains*, 180.
606. Haikonen, 181.
607. Haikonen, 182.
608. Haikonen, *Consciousness and Robot Sentience*, 159.

further, and which are to be avoided.[609] In this way, the decisions the robot makes are influenced by information related to affect and emotion,[610] just as observed in humans and animals given the research findings made by Antonio Damasio[611] as mentioned in the previous chapter.

By using the associative signal patterns from each module's set of feedback loops, the robot is capable of internally evoking percepts associated with a particular cue.[612] For example, if the robot learns that the word 'cat' is associated with a particular animal, speaking the word "cat" evokes an internal or "virtual" percept of the image it has associated with the word. This internally generated percept is created by its various feedback loops, as learned associations can evoke percepts upon registering an associated cue, such as a spoken word or visual image of an item. This functionality mimics inner imagery and inner speech in humans, as the robot can "imagine" or internally evoke a stimulus from the presence of associated cues. The evocation of the robot's inner speech or inner imagery would present the percept in a similar way to how it appears in humans, in which some features of the percept are less vivid or missing altogether if they cannot be recalled.[613] Returning to the cat example, the details of the cat's colouring may not be as clear in an imagined percept, however, the general body shape of the cat may be fully recalled.

Associations can be stored in memory through the use of Accept-and-Hold (AH) circuits, which separately store the signal vectors of percepts broadcast from perception/response

609. Haikonen, 160–61.
610. Haikonen, *Robot Brains*, 153.
611. Haikonen, 149.
612. Haikonen, *Consciousness and Robot Sentience*, 156.
613. Haikonen, *Robot Brains*, 79.

feedback loops.[614] Each AH circuit stores one signal vector, where multiple signal vectors are then cross-associated with others within an associative neuron group. Here, the associative neuron group links percepts together such that one percept acts as a cue to retrieve information associated with it, in a cross-referential manner. An example to further illustrate this structure involves an item left in specific locations at certain times;[615] after coming home from work, a cellphone is plugged in to charge in the kitchen, and after dinner, the phone is moved to the living room. Recalling where the phone is located can be cued by the time of day as, prior to dinner, the phone is in the kitchen, while afterwards, the phone is in the living room. The cross-associations generated within the associative neuron group generate synaptic weights and act as memory traces for the two scenarios of described above. Match, mismatch, and novelty outcomes also arise from this associative neuron group, indicating instances in which a set of percepts are either associated with each other or are not.[616] For example, a mismatch would occur if an incoming set of signals contained vectors representing 'cellphone', "before dinner," and 'hallway'.

This version of memory formation is similar to the creation of short-term memories in people, in which recent situations can be recalled for a limited period of time before they are forgotten.[617] The HCA is also capable of storing long-term memories for situations characterized by high emotional significance. Alternatively, long-term memories can be formed when information is circulated within perception/response feedback loops and is subsequently recalled

614. Haikonen, 141.
615. Haikonen, 140–41.
616. Haikonen, 142.
617. Haikonen, 140.

by short-term memory to be memorized again.[618] In the case of emotionally-charged memories, the threshold for association in the HAN is lowered such that an instantaneous association can be made.[619] For circulated memories, a separate long-term neuron group receives signal vectors which enter into short-term memory circuits. This performs the same cross-association seen elsewhere in the HCA, albeit with lower HAN thresholds. Thus, a higher number of repeated associations between incoming signal vectors is necessary for the association of stimuli. Recall is performed in the same manner in long-term neuron groups as short-term neuron groups, such that the presentation of a stimulus cues associated percepts to be remembered.

It is through these learned associations in the robot that meaning can be understood and language can be acquired. The associative neuron groups within the perception/response feedback loops enable the transition from sub-symbolic processing to symbolic processing, in which the main signal vector acts as a symbol for the corresponding associative pattern.[620] The associative signal patterns within the feedback loops used for comparison against the incoming main signal consist of the sub-symbolic features which represent a stimulus. For example, sub-symbolic features of visual stimuli may include lines and edges, shading, and light reflectances. Items are identified based on how an incoming signal compares to expectations and predictions. For example, a cellphone is generally rectangular and thin, and when the screen is off, is black and somewhat reflective of light. The visual information representing a cellphone can be associated with the spoken word ‘cellphone’, where hearing the word evokes the associated

618. Haikonen, 142.
619. Haikonen, 143.
620. Haikonen, ‘XCR-1’, 361.

visual signal pattern which represents the object.[621] The association between these two stimuli grounds the meaning of the word ‘cellphone’, as the sub-symbolic information generated from sense data are combined to create an understanding of what the symbol ‘cellphone’ refers to.[622]

Because robot emotions also arise from internally-generated percepts from sensors on the robot’s body, it is able to ground the meaning of words like ‘pleasure’ and ‘sadness’ in a similar manner. The sub-symbolic information generated from its “petting sensor” becomes the referent of the word ‘pleasure’ and similar terms like ‘joy’ or ‘comfort’. When sub-symbolic information is combined with other sub-symbolic information from system reactions, more complex robot emotions or affective states can be created.[623] If the robot associates the word ‘good’ with pleasure signals from the “petting sensor,” and subsequently the word ‘good’ with ‘cellphone’, a desire for interacting with a cellphone can be learned.[624] Alternatively, ‘sadness’ can be grounded in a combination of pain sensations and sensory mismatch, where an expectation for something good like a cellphone is taken away as the robot’s body is being struck. When the robot is in this mood as a result of its experiences, the utterance of ‘sadness’ can be learned over time through the repetition of this scenario. Therefore, the robot understands what words like ‘sadness’ and ‘pleasure’ mean, as the referents of these words are grounded in the robot’s own sensations and subjective experience.

From this grounding of words and meaning, language can be generated through *vertical* and *horizontal grounding*. Vertical grounding involves the association of sensory information

621. Haikonen, *Consciousness and Robot Sentience*, 138.
622. Haikonen, 139.
623. Haikonen, *The Cognitive Approach to Conscious Machines*, 214.
624. Haikonen, *Robot Brains*, 156.

and percepts to words, as seen in the previous examples regarding emotions. Adjectives like 'square' and 'black' can be determined through correlative learning, through which repeated interactions involving different objects sharing the same characteristic indicate the feature being referred to.[625] Horizontal grounding, on the other hand, associates the words within a sentence or multiple sentences to one another.[626] Just as objects, locations, and times of day can become associated to form a memory of where an item was left after dinner, the parts of a sentence can also become associated with one another. For example, the sentences "the cellphone is good" and "the cellphone is black and rectangular" associates 'cellphone' with the adjectives 'good', 'black', and 'rectangular'. When horizontal grounding and vertical grounding are combined, the system is able to report its experiences through language based on the associations it has learned.[627] Moreover, linguistic syntax can be learned by associating elements of sentences together, like word order and tense. Additionally, vocal inflection or the use of punctuation can also be learned through association.[628] Asking the robot "what is black and rectangular?" will generate the answer "cellphone" provided the associations have been learned, and upon querying the robot about whether the cellphone is bad, the answer will be "no."

The HCA also supports robot speech by imitating the sounds and words the robot hears.[629] To accomplish this, sound features processed by the auditory feedback loop are forwarded to a *sequence neuron assembly* which acts as a temporary memory store with

625. Haikonen, *The Cognitive Approach to Conscious Machines*, 223.
626. Haikonen, 229.
627. Haikonen, 231.
628. Haikonen, 236–37.
629. Haikonen, 219–20.

instantaneous learning.[630] The reason for this store is to collect and remember auditory information as it is being heard, only repeating sounds and words after the person speaking has finished.[631] Auditory percepts are then passed to a kinesthetic perception/response feedback loop which imitates the detected temporal signal patterns.[632] Speech sounds can be produced by synthesizing various sounds to generate a wider, more flexible range of sounds and words.[633] Signals are subsequently forwarded to hardware like an audio amplifier and loudspeaker, where the output sound is captured by the robot's own auditory processing channels. Initially, this is the only way in which the robot can hear its own inner speech. Over time however, an internal feedback loop can emerge, allowing the inner speech to be heard as silent speech without the need to talk aloud. Subsequent rehearsal of these words and sounds from reentrant auditory processing allows the robot to permanently learn the words it is exposed to.[634] Moreover, the established connection between perceived sounds and output signals is able to cause the robot to mimic the sounds it hears in the environment, acting similar to the so-called mirror neuron activity discussed in Chapter 2, albeit without the use of any specific or unique "mirror neurons."[635] Through the generation of associations between heard words and spoken sounds, in conjunction with the vertical and horizontal grounding of language, the HCA supports language acquisition. This enables the robot to both understand phrases and produce linguistic responses.

630. Haikonen, *Robot Brains*, 162.
631. Haikonen, 163.
632. Haikonen, 162; Haikonen, *Consciousness and Robot Sentience*, 207–8.
633. Haikonen, *Consciousness and Robot Sentience*, 208.
634. Haikonen, *Robot Brains*, 163.
635. Haikonen, *The Cognitive Approach to Conscious Machines*, 220.

A functional prototype of a robot using HCA and HANs can be viewed in demonstration videos uploaded by Haikonen to his YouTube channel.[636] In these videos, the prototype robot XCR-1 exhibits an ability to learn its own name and respond to being called.[637] The robot can also learn to avoid objects which have been previously associated with pain signals, as the robot associates an object with the sensation of its body being struck.[638] Moreover, the associative processing also enables the robot to recognize itself in a mirror since it learns the connection between visual stimuli detected from the mirror and its own kinesthetic activity.[639] These videos demonstrate the ways in which associative neurons are able to generate appropriate behaviours to various stimuli, mimicking the way animals and humans react and respond to aspects of their environment.

Haikonen's approach indicates a promising direction for overcoming limitations for social robots. Since language and empathy depend on capacities which cannot be replicated in existing robots like iCub. The reason this architecture is suitable is because it follows processes identified in living beings and aims to recreate these processes in a machine. Haikonen's SRTE is a good model for machine emotions because it reflects physiological regularities observed in nature. The resulting combinations of robot experiences provide a model which supports a rudimentary foundation for empathy in machines. Moreover, these feelings are comprised of electrical signals which are generated by the body to establish associations between stimuli, in which these associations provide a sense of meaning for the robot. As such, its "mind" and

636. Youtube, 'Pentti Haikonen'. See https://www.youtube.com/@PenHaiko for a list of videos.
637. *Calling Robot by Name Feat. Sound Direction Detection*.
638. *Robot Sequence Memory*.
639. *Robot Self-Consciousness. XCR-1 Passes the Mirror Test*.

behaviour is *for* the body to keep the system operating, as observed in animals and humans. This enables the agent to ground the meanings of emotions, as the robot has been constructed to generate its understanding of the world from specific sensors on its body related to pain and pleasure. Though XCR-1 is a simplistic model for the purposes of showing a functioning proof-of-concept, this approach can be expanded upon such that the robot is able to learn more words and objects. This would eventually enable the robot to produce sentences reflecting its own experiences operating in its environment.

Further development of this approach would generate a robotic analogue of empathy since a capacity to imitate is already present even in its current iteration. Due to XCR-1 being able to understand basic pain and pleasure, with further training, the robot could associate its own feelings with the detection of emotional states in others. If the robot were to learn to feel distress upon detecting distress in others, it could become motivated to alleviate its own negative feelings through attending to and comforting others, as witnessed in some animals. Since an act of empathy arises from an interaction between two agents, the robot could appeal to various behavioural cues to act appropriately, modifying its behaviour in response to the match, mismatch, or novelty signals produced from internal feedback loops. Since it is capable of experiencing inner states, the robot's act of empathy is not a mere simulation but a recreation of the processes involved in human and animal empathy.

Therefore, for more lifelike social human-robot interactions, a new approach to developing robotics is required. The previous chapter established the limitations of existing AIs like iCub, and in response, this chapter has offered an alternative and appealing solution. The reason for adopting Haikonen's approach for the development of social robots is due to its ability

to generate internal sensations from bodily sensors, and from these sensations, cognition and behaviour responds to environmental stimuli. This is in contrast to current AIs which generate behaviours without the use of internal sensations and experiences. While some robotic applications can make do with this approach, social robots are better off if they are able to feel and experience the world like humans and animals. The architecture invented by Haikonen also better mimics biological regularities as it makes use of feedback loops for machine learning and the regulation of behaviours. By basing behavioural outcomes on the robot's own physical experiences, it is able to operate in a way more analogous to humans and animals, leading to better artificial agents for socializing with people.

Altogether, this chapter aims to illustrate the rationale for a new approach to building AIs given the shortcomings identified in the previous chapter. Specifically, machine behaviour is generated by the way signals generated by bodily sensors influence its own interests, as indicated by the presence of analogues of pain and pleasure. These sensations not only give rise to analogous affective states, a requirement for empathy, but they also provide the machine with the meanings of the stimuli it encounters. An object associated with analogous pain signals is deemed "bad" given its impact on the machine's functioning. This gives rise to behaviours which aim to mitigate the stimulus's effects to promote its own continued functioning. In this manner, the robot is able to organize its own functionality around its experiences with the environment, demonstrating a similarity to the self-organizing processes central to biological organisms. Without these analogous affective states, environmental stimuli and aspects of the wider world are meaningless, as the robot has no stake in its own survival as an autonomous system. The rudimentary self-organizing behaviours observed in the robot XCR-1 arise from feedback loops, in which Haikonen Associative Neurons are used to process and reprocess percepts for the sake

of governing its behaviour according to the demands of the environment. Although affective states and emotions in humans and animals are often considered reactions or responses to interactions between agents and aspects of the world, their valence provides vital information about the environment. This information pertains to the individual's survival and point of view as a living being as it navigates the world. Decision-making without this underlying emotional framework in animals like humans is not feasible nor desirable, as it overlooks an integral element of *evaluation*. This evaluation occurs from the ways in which stimuli give rise to affect and emotional experiences, in which stimuli are evaluated as 'bad' if they negatively impact the individual. While computers and other machines may not require the implementation of analogous affective states to perform behaviours, any true attempt at recreating human cognition arguably does. Given that computers are incapable of having experiences, our current attempts at creating "artificial intelligence" are not as well-informed as they could be. By building a machine with the capacity to experience its own form of emotions and affective states, it could understand for itself what emotions and feelings mean to humans.

The following chapter concludes this work by integrating the discussions presented throughout these four chapters. The goal of this work is to demonstrate the importance of developing affect for empathy in social robots, as empathy has been identified as a significant characteristic for social robots. This involves communication with a degree of social appropriateness and emotional sensitivity, one learned through interacting with a myriad of different individuals from different backgrounds. To accomplish this, the robot must be able to experience the world in an analogous manner to humans and animals. Any social robot without this understanding of emotion consists of a mere simulation, potentially introducing harms to

human lives and societies. Until these experiences are implemented, it is insufficient to conclude that the robot acts with empathy simply because it is behaving as if it does.

6 Concluding Remarks

We live in a world excited by the prospect of new machines capable of humanlike abilities. With recent advances in artificial intelligence and robotics, it seems that just about all aspects of human life could one day be outsourced or augmented by new technologies. Specifically, social robots designed to interact with people could see wide-spread implementation in a variety of settings, from personal or domestic contexts to public or professional environments. Though each of these situations will require social robots to possess specific abilities, some capacities will be similar between models as well. To be a 'social' robot, some degree of interaction and communication is required, engaging with users in an appropriate or informative manner.

Prior to discussing desirable social robot capacities, Chapter 1 began with a discussion of shifting demographics to demonstrate a source of demand for these agents. Aging populations in societies around the world require novel solutions to meet growing demand for care and services. One approach includes using robots for support in eldercare, a solution which emerged from Japan given their status as a world-leader in both robotics development and aging populations. As time has progressed, however, more regions around the world are witnessing similar demographic trends, indicating an increasing need for elderly care. With fewer working-age adults able to meet this demand worldwide, technological solutions provide a means to assist in caring for aging individuals.

This approach introduces questions surrounding what care involves. While robots and AI can be used in a variety of ways to assist human workers in caring for the elderly, some forms of

work will require social skills and empathy. Patients must be treated with respect as individuals and not as tasks to be dealt with. Even in non-healthcare related settings, surveys indicate that people want social robots to act with empathy, expecting them to act sufficiently humanlike as to be easy and intuitive to interact with. It turns out, as demonstrated in Chapter 3, that this requirement is no easy feat. In fact, if we want our robots to act with empathy, a new approach to robotics is required.

Human interactions are complex, and in order to discuss empathy in general, a closer look at interpersonal communications and social behaviour is required. Not only does language involve the use of syntax and vocabulary, it also relies on references to and interaction with aspects of the physical world. Additionally, human communication relies on behaviours like body language and non-linguistic vocalizations to convey information. Social robots must be able to understand the meanings of these communicative actions to act appropriately, in addition to responding in a sufficiently familiar manner.

For social robots to understand the meanings of words, however, the symbol grounding problem must be resolved. Although the introduction of robotics to the field of AI was thought to address this problem, the issue still remains, as these robots are not structured in an analogous way to living beings. While current versions of AI may not understand or possess an analogous mind, the reason is due to the way they are designed as set of decoupled modules of functionality. If structured differently, in a way which gives rise to a self-organizing unity, an artificial agent could possess an analogue of a mind which grounds the meanings of words. Throughout this work, I have focused on terms related to emotion and affect for illustrating the requirements of machine empathy. It turns out, however, that affect is required for motivation,

and motivation is the key to directing attention and organizing behaviour. So although robots like iCub are not sentient or capable of understanding or expressing empathy, this does not mean that AIs are *necessarily* incapable of possessing an analogous mind of their own. It does, however, require an alternative approach to robotics, as was demonstrated in Chapter 4.3 with the introduction of Haikonen's cognitive architecture. Only when a robot can experience the world in which it exists can it be capable of understanding emotions and expressing an analogue of empathy.

Empathy is an important capacity to consider in relation to social robots given its requirement in socialization and healthcare. The ability to adopt the perspective of another, to some degree at least, is a significant aspect of human life, providing a means for understanding and relating to others. Given the general aim of healthcare to reduce the suffering of individuals, empathy and respect is of utmost importance, especially with respect to nursing for patient recovery.[640] Since we generally all know what it feels like to suffer, most people are interested in doing what they can to assist others in feeling better. In fact, researchers[641] suggest that the motivation to assist others is directed by an individual's own ability to comprehend and feel the sensations witnessed in the behaviours of others. In Chapter 2, after discussing the philosophical background of empathy, an empirical theory of empathy was presented which suggests that in human and some mammals, a desire to assist others is driven by the presence of an emotional contagion. Upon seeing another suffering, some animals and many humans begin to feel a degree of stress or discomfort which motivates them to address this suffering in some way.

640. Moudatsou et al., 'The Role of Empathy in Health and Social Care Professionals', 3.
641. de Waal and Preston, 'Mammalian Empathy', 498.

Although empathy involves an ability to simulate within oneself the perceived experiences of others, acts of empathy cannot be reduced to simulation alone. In some situations, efforts to simulate the experiences of others must be expanded upon by using one's own knowledge to better understand how another is feeling. Cultural differences between individuals may complicate one's ability to empathize if emotional display rules and customs are sufficiently distinct, as one may be unsure of the complete context of another's experiences. As such, by appealing to one's own knowledge of the individual in question or other information like cultural differences, individuals may better understand how another is feeling when simulation alone is insufficient.

In this way, empathy involves an interaction between two or more people, a theory of mind proposed by Shaun Gallagher. Specific situations involving unique individuals in a variety of roles and relationships ultimately determine whether empathy can be expressed, and whether these acts are accurate given the situation at hand. If the context and individuals involved in a particular situation is somewhat familiar, perhaps involving friends or family members, an act of empathy is more likely to be accurate. In cases where individuals are strangers to each other, it is more likely that acts of empathy may not accurately understand the situation, as the behaviours of others may be misinterpreted due to personal differences or other contextual factors. Gallagher's *interaction theory* provides a good explanation of empathy given its sensitivity to a variety of considerations, factors which are able to explain why acts of empathy fail or succeed. Specifically, this theory characterizes an act of empathy as emerging from the second-person perspective, a subjective point of view which is oriented outward and toward another person in

an attempt to understand, communicate, or interact with them.[642] Rather than focusing on one's own experiences from a first-person point of view, a second-person perspective aims to establish an intersubjective understanding of the world for the sake of communication.[643]

Because empathy relies on an understanding of the internal states of others, it inherently deals with interpreting behavioural cues which signal emotional states. This poses a problem for social robots because emotions are produced by the body as an outcome of biology, a consequence of both evolutionary and developmental processes. Since current approaches to social robots involve the use of embodied computers, robots are not provided with a suitable analogue for human emotion. We saw this in our discussion of the developmental robot iCub in Chapter 3. Any information iCub learns about human emotion is provided to it by humans, rather than emerging from its own body. Moreover, iCub and similar robots do not experience the world and its elements, they merely react in response to inputs and data. In contrast, living organisms interact with aspects of their environment based on the ways in which these aspects impact their bodies, resulting in emotions which signal the impact or significance of these aspects. Without experiencing for itself what emotions are, or what they are grounded in, the robot is unable to determine what a human means by 'happiness' or 'sadness'. This lack of understanding indicates that any act of apparent empathy from iCub is a simulation rather than a machine analogue of the real capacity.

642. Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 90–91.
643. Gallagher, 'Neurons, Neonates and Narrative', 174; Gallagher, 'The Practice of Mind. Theory, Simulation or Primary Interaction?', 99.

Despite AI developers building embodied machines capable of learning through interactions with the environment, these robots are insufficiently embodied. These robots do not recreate emotions and other affective states through their bodies, as animals and humans do. Sensations of pain, for example, do not emerge in developmental robots like iCub. As such, the concept of pain cannot be understood by robots like iCub because this understanding relies on the subjective experience of it. iCub may be able to recognize human emotions, yet this functionality is treated like a module to be added or removed depending on the robot's use, rather than a core feature of the robot's body as an agent. Instead of serving as a fundamental mechanism for learning and guiding behaviour, affect and emotion seem to be considered as just another element in a suite of computational abilities. This design decision results in an insufficiency in the way robot bodies are created, as a significant aspect of embodiment for living creatures is absent.

This idea of embodiment is derived from a form of organization observed in living beings, namely in the form of autopoiesis or self-organization. As nested levels of physical systems, organisms are able to remain alive and reproduce as a product of the physiological interactions which comprise its body and behaviours. Subjective experiences and emotions emerged as a result of evolutionary processes because they facilitate these self-organizing processes, such as signalling damage in the form of pain. Without subjective experiences, individuals would not be able to respond to changes in the environment, and as a result, would be more likely to perish. Awareness and responsivity observed in living organisms exists, directly or indirectly, for the sake of maintaining these self-organizing processes. In fact, when Hubert Dreyfus hints at a missing element within AI, it is exactly this inherent interest or "concern"

which robots do not share with living beings.[644] Although a robot may be embodied and capable of learning over time, it does not experience its interactions with aspects of the environment, unlike biological organisms. The learning which occurs is distinct from the learning which transpires in animals because it does not arise from subjective experience, but from the modification of numerical weights within its neural networks. Though these adjustments produce changes in behaviour, the robot's actions are not motivated by subjective experience. Instead, they result from the way its programming has been developed for a particular purpose or specification, rather than being driven by its own inherent self-organizing processes. Even if the machine's purpose were to be a self-governing autonomous agent, its functional design nonetheless remains fundamentally distinct from the functional organization of biological organisms.

For robots to feel emotions in an analogous manner to humans and animals, a new approach to robotics is required. The solution proposed by Haikonen offers a better biological model of neurons as it does not use numerical values, but operates on specifically formatted signals generated by dedicated sensors. This includes signals representing pain and pleasure which serve to disrupt or soothe the robot's internal state. Moreover, because these signals provide a direct source of meaning about stimuli sensed in the environment, Haikonen considers them to be self-explanatory. As a result, they can serve as symbols when associated with other percepts; for example, the association between a spoken word and visual information generates an understanding of a name for the detected object or feature. The learning which occurs in the Haikonen Cognitive Architecture is unlike the learning which occurs in existing robots, as it does

644. Dreyfus, *What Computers Still Can't Do*, xlv–xlvi.

not use syntactical structures, but direct experiences of stimuli in the environment. When stimuli are associated with pain signals, for example, they are deemed "bad" as a consequence of the ways in which they impact the body and appear within subjective experience. As a result, this causes the robot to avoid further engagement with the stimulus.

Therefore, this new approach provides a better candidate for developing robot empathy because it mimics the biological processes which provide meaning to humans and animals. Although robots of the future would not be able to know what it is like for a *human* to experience pain, it could use its own understanding and experiences to infer how a person may be feeling. Moreover, the use of intrinsic rewards and punishments, like pleasure and pain, could be used to further develop robot empathy. A robot that witnesses others in pain could be motivated to alleviate the suffering of others, if its associative neurons were to simulate that experience within its own cognitive architecture. This would provide an analogous mechanism for the purported way in which acts of empathy can emerge in some animals. Additionally, robot empathy could be fostered by interactions which generate feelings of pleasure or happiness. As discussed in Chapter 2, the establishment of intersubjective communication is learned through simple interactions between infants and adults, often involving mimicry and positive emotions. A smiling or babbling infant often elicits a suitable response from a caregiver, generally in the form of a smile, a similar sound or word, or some other form of positive reinforcement. This exchange is typically accompanied by some degree of pleasure in both the infant and the adult, increasing the likelihood of further interaction. As the child continues to grow, the subject matter of these interactions expands to include various aspects in the environment, establishing joint attention and improved intersubjective communication. In robots, a capacity for empathy could also be fostered by following a similar paradigm, one which establishes the learning of intersubjective

communication through the use of positive emotions and playful mimicry. Doing so introduces an analogue of Gallagher's *interaction theory* which could further develop empathy in social robots.

In conclusion, although the current trajectory of existing social robots does not enable them to possess a capacity for empathy, a new solution has been identified. By following a range of principles and theories within biology and cognitive science, a robot with expanded social abilities could be developed. This alternative would be better suited to engaging with people in an intuitive, humanlike manner, assisting individuals and societies in a range of different roles all around the world. While it is especially important to focus on supporting elderly populations, empathic social robots could be used for other purposes as well, from entertainment and companionship to assisting professionals in their workplace and endeavours. Effective communication can be further improved by providing these robots with an understanding of emotions and an ability to directly perceive the internal states of others. Not only would these robots appear to be more intelligent and better equipped for understanding humans, they would also have the ability to respond appropriately to the feelings and experiences of people from a wide range of sociocultural backgrounds.

7 Bibliography

Adami, Christoph. ‘A Brief History of Artificial Intelligence Research’. *Artificial Life* 27, no. 2 (2 May 2021): 131–37. https://doi.org/10.1162/artl_a_00349.

Agosta, Lou. *Empathy in the Context of Philosophy*. London: Palgrave Macmillan UK, 2010. https://doi.org/10.1057/9780230275249.

Alonso, Eduardo. ‘Actions and Agents’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 232–46. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.015.

Angell, Frank. ‘Titchener at Leipzig’. *The Journal of General Psychology* 1, no. 2 (1 April 1928): 195–98. https://doi.org/10.1080/00221309.1928.9920123.

Arkoudas, Konstantine, and Selmer Bringsjord. ‘Philosophical Foundations’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 34–63. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.004.

Ashby, William Ross. *An Introduction to Cybernetics*. London: Chapman & Hall, 1956. http://pcp.vub.ac.be/books/IntroCyb.pdf.

Aymerich-Franch, Laura. ‘Why It Is Time to Stop Ostracizing Social Robots’. *Nature Machine Intelligence* 2, no. 7 (July 2020): 364–364. https://doi.org/10.1038/s42256-020-0202-5.

Baaren, Rick B. van, Jean Decety, Ap Dijksterhuis, Andries van de Leij, and Matthijs L. van Leeuwen. ‘Being Imitated: Consequences of Nonconsciously Showing Empathy’. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 31–42. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0004.

Bagheri, Elahe, Oliver Roesler, Hoang-Long Cao, and Bram Vanderborght. ‘A Reinforcement Learning Based Cognitive Empathy Framework for Social Robots’. *International Journal of Social Robotics* 13, no. 5 (1 August 2021): 1079–93. https://doi.org/10.1007/s12369-020-00683-4.

Baldwin, Richard. *The Globotics Upheaval: Globalization, Robotics, and the Future of Work*. New York, NY: Oxford University Press, 2019.

Barrett, Lisa Feldman. ‘Are Emotions Natural Kinds?’ *Perspectives on Psychological Science* 1, no. 1 (1 March 2006): 28–58. https://doi.org/10.1111/j.1745-6916.2006.00003.x.

———. ‘Psychological Construction: The Darwinian Approach to the Science of Emotion’. *Emotion Review* 5, no. 4 (1 October 2013): 379–89. https://doi.org/10.1177/1754073913489753.

Barrett, Lisa Feldman, and James J. Gross. ‘Emotional Intelligence: A Process Model of Emotion Representation and Regulation’. In *Emotions: Currrent Issues and Future Directions*, 286–310. Emotions and Social Behavior. New York, NY, US: The Guilford Press, 2001.

Barrett, Lisa Feldman, and Kristen A. Lindquist. ‘The Embodiment of Emotion’. In *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, 237–62. Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511805837.011.

Barrett, Lisa Feldman, Kristen A. Lindquist, and Maria Gendron. ‘Language as Context for the Perception of Emotion’. *Trends in Cognitive Sciences* 11, no. 8 (August 2007): 327–32. https://doi.org/10.1016/j.tics.2007.06.003.

Barrett, Lisa Feldman, Batja Mesquita, Kevin N. Ochsner, and James J. Gross. ‘The Experience of Emotion’. *Annual Review of Psychology* 58, no. 1 (2007): 373–403. https://doi.org/10.1146/annurev.psych.58.110405.085709.

Barsalou, Lawrence W., W. Kyle Simmons, Aron K. Barbey, and Christine D. Wilson. ‘Grounding Conceptual Knowledge in Modality-Specific Systems’. *Trends in Cognitive Sciences* 7, no. 2 (1 February 2003): 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3.

Barsalou, Lawrence W., and Katja Wiemer-Hastings. ‘Situating Abstract Concepts’. In *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, edited by Diane Pecher and Rolf A. Zwaan, 129–63. Cambridge: Cambridge University Press, 2005. https://doi.org/10.1017/CBO9780511499968.007.

Bartneck, Christoph, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-Robot Interaction: An Introduction*. Cambridge University Press, 2020.

Bays, Carter. ‘Introduction to Cellular Automata and Conway’s Game of Life’. In *Game of Life Cellular Automata*, edited by Andrew Adamatzky, 1–7. London: Springer, 2010. https://doi.org/10.1007/978-1-84996-217-9_1.

Bedau, Mark A. ‘Artificial Life’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 296–315. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.019.

Beer, Jenay M., Karina R. Liles, Xian Wu, and Sujan Pakala. ‘Affective Human–Robot Interaction’. In *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Myounghoon Jeon, 359–81. San Diego: Academic Press, 2017. https://doi.org/10.1016/B978-0-12-801851-4.00015-X.

Beer, Randall D. ‘Dynamical Systems and Embedded Cognition’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 128–48. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.009.

Belpaeme, Tony, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. ‘Social Robots for Education: A Review’. *Science Robotics* 3, no. 21 (15 August 2018). https://doi.org/10.1126/scirobotics.aat5954.

Benita, Moti, Talia Shechter, Shahar Nudler-Muzikant, and Reout Arbel. ‘Emotion Regulation during Personal Goal Pursuit: Integration versus Suppression of Emotions’. *Journal of Personality* 89, no. 3 (2021): 565–79. https://doi.org/10.1111/jopy.12599.

Berkeley, Edmund Callis. *Giant Brains, or Machines That Think*. New York: John Wiley & Sons, 1949. https://www.gutenberg.org/files/68991/68991-h/68991-h.htm.

Bhaumik, Arkapravo. *From AI to Robotics: Mobile, Social, and Sentient Robots*. 1st edition. Boca Raton: CRC Press, 2018.

Bickhard, Mark H. ‘Robot Sociality: Genuine or Simulation?’ In *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*, edited by Raul Hakli and Johanna Seibt, 41–66. Studies in the Philosophy of Sociality. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-53133-5_3.

Black, H. S. ‘Stabilized Feedback Amplifiers’. *Bell System Technical Journal* 13, no. 1 (1934): 1–18. https://doi.org/10.1002/j.1538-7305.1934.tb00652.x.

Bloom, Paul, and Tim P German. ‘Two Reasons to Abandon the False Belief Task as a Test of Theory of Mind’. *Cognition* 77, no. 1 (16 October 2000): B25–31. https://doi.org/10.1016/S0010-0277(00)00096-2.

Boden, Margaret A. ‘GOFAI’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 89–107. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.007.

Bongard, Josh, Victor Zykov, and Hod Lipson. ‘Resilient Machines Through Continuous Self-Modeling’. *Science* 314, no. 5802 (17 November 2006): 1118–21. https://doi.org/10.1126/science.1133687.

Bourgine, Paul, and Francisco J. Varela. ‘Toward a Practice of Autonomous Systems’. In *Toward a Practice of Autonomous Systems Proceedings of the First European Conference on Artificial Life*, xi–xvii. Complex Adaptive Systems. Cambridge, Mass: The MIT Press, 1991. https://mitpress.mit.edu/9780262720199/toward-a-practice-of-autonomous-systems/.

Boyd, Danah, and Kate Crawford. ‘Critical Questions for Big Data’. *Information, Communication & Society* 15, no. 5 (1 June 2012): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Brandtzaeg, Petter Bae, and Asbjørn Følstad. ‘Chatbots: Changing User Needs and Motivations’. *Interactions* 25, no. 5 (22 August 2018): 38–43. https://doi.org/10.1145/3236669.

Broekens, Joost, Marcel Heerink, and Henk Rosendal. ‘Assistive Social Robots in Elderly Care: A Review’. *Gerontechnology* 8, no. 2 (2009): 94–103.

Brooks, Rodney. ‘Artificial Life and Real Robots’. In *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, edited by Francisco J. Varela and Paul Bourgine, 3–10. Complex Adaptive Systems. Cambridge, Mass: The MIT Press, 1991.

Brooks, Rodney A. ‘Elephants Don’t Play Chess’. *Robotics and Autonomous Systems*, Designing Autonomous Agents, 6, no. 1 (1 June 1990): 3–15. https://doi.org/10.1016/S0921-8890(05)80025-9.

Brooks, Rodney A., Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. ‘The Cog Project: Building a Humanoid Robot’. In *Computation for*

Metaphors, Analogy, and Agents, edited by Chrystopher L. Nehaniv, 52–87. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1999. https://doi.org/10.1007/3-540-48834-0_5.

Buchanan, Bruce G. 'A (Very) Brief History of Artificial Intelligence'. *AI Magazine* 26, no. 4 (15 December 2005): 53–53. https://doi.org/10.1609/aimag.v26i4.1848.

Burns, Timothy A. 'Empathy, Simulation, and Neuroscience: A Phenomenological Case against Simulation-Theory'. *Phenomenology and Mind*, no. 12 (9 August 2017): 208–16. https://doi.org/10.13128/Phe_Mi-21119.

Cabibihan, John-John, Hifza Javed, Marcelo Ang, and Sharifah Mariam Aljunied. 'Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism'. *International Journal of Social Robotics* 5, no. 4 (1 November 2013): 593–618. https://doi.org/10.1007/s12369-013-0202-2.

Čaić, Martina, Dominik Mahr, and Gaby Oderkerken-Schröder. 'Value of Social Robots in Services: Social Cognition Perspective'. *Journal of Services Marketing* 33, no. 4 (1 January 2019): 463–78. https://doi.org/10.1108/JSM-02-2018-0080.

Calling Robot by Name Feat. Sound Direction Detection, 2020. https://www.youtube.com/watch?v=b_5cohcRRfM.

Campa, Riccardo. 'The Rise of Social Robots : A Review of the Recent Literature'. *Journal of Evolution and Technology* 26, no. 1 (2016). https://ruj.uj.edu.pl/xmlui/handle/item/42187.

Cangelosi, Angelo, Tony Belpaeme, Giulio Sandini, Giorgio Metta, Luciano Fadiga, Gerhard Sagerer, Katherina Rohlfing, Britta Wrede, Stefano Nolfi, and Domenico Parisi. 'The ITALK Project: Integration and Transfer of Action and Language Knowledge in Robots'. In *Proceedings of Third ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)*, 12:15, 2008.

Cangelosi, Angelo, and Matthew Schlesinger. *Developmental Robotics: From Babies to Robots*. Intelligent Robotics & Autonomous Agents Series. US: The MIT Press, 2015.

———. 'From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology'. *Child Development Perspectives* 12, no. 3 (2018): 183–88. https://doi.org/10.1111/cdep.12282.

Carpenter, Malinda, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 'Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age'. *Monographs of the Society for Research in Child Development* 63, no. 4 (1998): i–174. https://doi.org/10.2307/1166214.

Carr, Evan W., Anne Kever, and Piotr Winkielman. 'Embodiment of Emotion and Its Situated Nature'. In *The Oxford Handbook of 4E Cognition*, edited by Albert Newen, Leon De Bruin, and Shaun Gallagher, 529–52. Oxford University Press, 2018. https://doi.org/10.1093/oxfordhb/9780198735410.013.30.

Carter, C. Sue, James Harri, and Stephen W. Porges. 'Neural and Evolutionary Perspectives on Empathy'. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 169–82. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0014.

Chatterji, Somnath, Paul Kowal, Colin Mathers, Nirmala Naidoo, Emese Verdes, James P. Smith, and Richard Suzman. 'The Health Of Aging Populations In China And India'. *Health Affairs* 27, no. 4 (July 2008): 1052–63. https://doi.org/10.1377/hlthaff.27.4.1052.

Chawla, Mukesh, Gordon Betcherman, and Arup Banerji. *From Red to Gray: The 'Third Transition' of Aging Populations in Eastern Europe and the Former Soviet Union*. World Bank Publications, 2007.

Chen, Lincoln, Timothy Evans, Sudhir Anand, Jo Ivey Boufford, Hilary Brown, Mushtaque Chowdhury, Marcos Cueto, et al. 'Human Resources for Health: Overcoming the Crisis'. *The Lancet* 364, no. 9449 (27 November 2004): 1984–90. https://doi.org/10.1016/S0140-6736(04)17482-5.

Chong, Trevor T. J., and Jason B. Mattingley. 'Automatic and Controlled Processing within the Mirror Neuron System'. In *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition*, edited by Jaime A. Pineda, 213–33. Contemporary Neuroscience. Totowa, NJ: Humana Press, 2009. https://doi.org/10.1007/978-1-59745-479-7_10.

Churamani, Nikhil, Francisco Cruz, Sascha Griffiths, and Pablo Barros. 'iCub: Learning Emotion Expressions Using Human Reward'. *arXiv:2003.13483 [Cs]*, 30 March 2020. http://arxiv.org/abs/2003.13483.

Cifuentes, Carlos A., Maria J. Pinto, Nathalia Céspedes, and Marcela Múnera. 'Social Robots in Therapy and Care'. *Current Robotics Reports* 1, no. 3 (1 September 2020): 59–74. https://doi.org/10.1007/s43154-020-00009-2.

Clark, Andy. *Being There: Putting Brain, Body, and World Together Again*. Bradford Books, 1996.

Clodic, Aurélie, Elisabeth Pacherie, Rachid Alami, and Raja Chatila. 'Key Elements for Human-Robot Joint Action'. In *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*, edited by Raul Hakli and Johanna Seibt, 159–77. Studies in the Philosophy of Sociality. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-53133-5_8.

Colombo, Francesca, and Jérôme Mercier. 'Help Wanted! Balancing Fair Protection and Financial Sustainability in Long-Term Care'. *Eurohealth* 17, no. 2–3 (2011): 3–6.

Cook, Richard, Geoffrey Bird, Caroline Catmur, Clare Press, and Cecilia Heyes. 'Mirror Neurons: From Origin to Function'. *Behavioral and Brain Sciences* 37, no. 2 (April 2014): 177–92. https://doi.org/10.1017/S0140525X13000903.

Coplan, Amy, and Peter Goldie, eds. 'Introduction'. In *Empathy: Philosophical and Psychological Perspectives*, ix–xlviii. Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199539956.002.0007.

Crevier, Daniel. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York, NY: BasicBooks, 1993.

Cuff, Benjamin M.P., Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 'Empathy: A Review of the Concept'. *Emotion Review* 8, no. 2 (1 April 2016): 144–53. https://doi.org/10.1177/1754073914558466.

Czaja, Sara J., and Marco Ceruso. 'The Promise of Artificial Intelligence in Supporting an Aging Population'. *Journal of Cognitive Engineering and Decision Making* 16, no. 4 (1 December 2022): 182–93. https://doi.org/10.1177/15553434221129914.

Damasio, Antonio R. 'Emotion and the Human Brain'. *Annals of the New York Academy of Sciences* 935, no. 1 (2001): 101–6. https://doi.org/10.1111/j.1749-6632.2001.tb03475.x.

———. ‘The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex’. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 29 October 1996. https://doi.org/10.1098/rstb.1996.0125.

Darwall, Stephen. ‘Empathy, Sympathy, Care’. *Philosophical Studies* 89, no. 2 (1 March 1998): 261–82. https://doi.org/10.1023/A:1004289113917.

Dautenhahn, Kerstin. ‘Roles and Functions of Robots in Human Society: Implications from Research in Autism Therapy’. *Robotica* 21, no. 4 (August 2003): 443–52. https://doi.org/10.1017/S0263574703004922.

Davies, D. W. ‘Mechanization of Thought Processes’. *Nature* 183, no. 4656 (1 January 1959): 225–26. https://doi.org/10.1038/183225a0.

Dawkins, Richard. *The Selfish Gene*. 2nd ed. Oxford University Press, 1989.

Debes, Remy. ‘From Einfühlung to Empathy’. In *Sympathy: A History*, edited by Eric Schliesser, 286–322. Oxford University Press, 2015.

———. ‘Which Empathy? Limitations in the Mirrored “Understanding” of Emotion’. *Synthese* 175, no. 2 (1 July 2010): 219–39. https://doi.org/10.1007/s11229-009-9499-7.

Decety, Jean, and Philip L. Jackson. ‘The Functional Architecture of Human Empathy’. *Behavioral and Cognitive Neuroscience Reviews* 3, no. 2 (1 June 2004): 71–100. https://doi.org/10.1177/1534582304267187.

Decety, Jean, and Andrew N. Meltzoff. ‘Empathy, Imitation, and the Social Brain’. In *Empathy: Philosophical and Psychological Perspectives*, edited by Amy Coplan and Peter Goldie, 58–81. Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199539956.003.0006.

Decety, Jean, and Meghan Meyer. ‘From Emotion Resonance to Empathic Understanding: A Social Developmental Neuroscience Account’. *Development and Psychopathology* 20, no. 4 (2008): 1053–80. https://doi.org/10.1017/S0954579408000503.

Dennett, Daniel C. ‘Why You Can’t Make a Computer That Feels Pain’. *Synthese* 38, no. 3 (1978): 415–56.

Depounti, Iliana, Paula Saukko, and Simone Natale. ‘Ideal Technologies, Ideal Women: AI and Gender Imaginaries in Redditors’ Discussions on the Replika Bot Girlfriend’. *Media,*

Culture & Society 45, no. 4 (1 May 2023): 720–36. https://doi.org/10.1177/01634437221119021.

Di Nuovo, Alessandro, Vivian M. De La Cruz, Angelo Cangelosi, and Santo Di Nuovo. 'The iCub Learns Numbers: An Embodied Cognition Study'. In *2014 International Joint Conference on Neural Networks (IJCNN)*, 692–99, 2014. https://doi.org/10.1109/IJCNN.2014.6889795.

Diehl, Joshua J., Lauren M. Schmitt, Michael Villano, and Charles R. Crowell. 'The Clinical Use of Robots for Individuals with Autism Spectrum Disorders: A Critical Review'. *Research in Autism Spectrum Disorders* 6, no. 1 (1 January 2012): 249–62. https://doi.org/10.1016/j.rasd.2011.05.006.

DIGHUM. 'Johanna Seibt'. Accessed 7 June 2024. https://caiml.org/dighum/perspectives-on-digital-humanism/authors/johanna-seibt/.

Dike, Happiness Ugochi, Yimin Zhou, Kranthi Kumar Deveerasetty, and Qingtian Wu. 'Unsupervised Learning Based On Artificial Neural Network: A Review'. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–27, 2018. https://doi.org/10.1109/CBS.2018.8612259.

Dreyfus, Hubert L. 'Alchemy and Artificial Intelligence'. RAND Corporation, 1 January 1965. https://www.rand.org/pubs/papers/P3244.html.

———. *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. 6th ed. Cambridge, Mass: MIT Press, 1995.

———. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972.

———. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, Mass: MIT Press, 1992.

Dubin, Adrienne E., and Ardem Patapoutian. 'Nociceptors: The Sensors of the Pain Pathway'. *The Journal of Clinical Investigation* 120, no. 11 (1 November 2010): 3760–72. https://doi.org/10.1172/JCI42843.

Edwards, Bosede I., and Adrian D. Cheok. 'Why Not Robot Teachers: Artificial Intelligence for Addressing Teacher Shortage'. *Applied Artificial Intelligence* 32, no. 4 (21 April 2018): 345–60. https://doi.org/10.1080/08839514.2018.1464286.

Eisenberg, Nancy, and Natalie D. Eggum. ‘Empathic Responding: Sympathy and Personal Distress’. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 71–83. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0007.

Ekman, Paul. ‘An Argument for Basic Emotions’. *Cognition & Emotion* 6, no. 3–4 (1 May 1992): 169–200. https://doi.org/10.1080/02699939208411068.

———. ‘Are There Basic Emotions?’ *Psychological Review* 99, no. 3 (1992): 550–53. https://doi.org/10.1037/0033-295X.99.3.550.

———. ‘Body Position, Facial Expression, and Verbal Behavior during Interviews’. *The Journal of Abnormal and Social Psychology* 68, no. 3 (1964): 295–301. https://doi.org/10.1037/h0040225.

———. ‘Darwin’s Contributions to Our Understanding of Emotional Expressions’. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, no. 1535 (12 December 2009): 3449. https://doi.org/10.1098/rstb.2009.0189.

———. ‘Universal Facial Expressions of Emotion’. *California Mental Health Research Digest* 8, no. 4 (1970): 151–58.

Ekman, Paul, and Wallace V. Friesen. ‘Head and Body Cues in the Judgment of Emotion: A Reformulation’. *Perceptual and Motor Skills* 24, no. 3, PT. 1 (1967): 711–24. https://doi.org/10.2466/pms.1967.24.3.711.

Ekmekci, Perihan Elif, and Berna Arda. ‘History of Artificial Intelligence’. In *Artificial Intelligence and Bioethics*, edited by Perihan Elif Ekmekci and Berna Arda, 1–15. SpringerBriefs in Ethics. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-52448-7_1.

Esposito, Anna, and Lakhmi C. Jain. ‘Modeling Social Signals and Contexts in Robotic Socially Believable Behaving Systems’. In *Toward Robotic Socially Believable Behaving Systems - Volume II : Modeling Social Signals*, edited by Anna Esposito and Lakhmi C. Jain, 5–11. Intelligent Systems Reference Library. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-31053-4_2.

Feinberg, Todd E., and Jon Mallatt. ‘Phenomenal Consciousness and Emergence: Eliminating the Explanatory Gap’. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.01041.

———. *The Ancient Origins of Consciousness: How the Brain Created Experience*. The MIT Press, 2016.

———. ‘The Nature of Primary Consciousness. A New Synthesis’. *Consciousness and Cognition* 43 (1 July 2016): 113–27. https://doi.org/10.1016/j.concog.2016.05.009.

Feng, Zhanlian, Chang Liu, Xinping Guan, and Vincent Mor. ‘China’s Rapidly Aging Population Creates Policy Challenges In Shaping A Viable Long-Term Care System’. *Health Affairs* 31, no. 12 (December 2012): 2764–73. https://doi.org/10.1377/hlthaff.2012.0535.

Fischer, Tobias, Jordi-Ysard Puigbò, Daniel Camilleri, Phuong D. H. Nguyen, Clément Moulin-Frier, Stéphane Lallée, Giorgio Metta, Tony J. Prescott, Yiannis Demiris, and Paul F. M. J. Verschure. ‘iCub-HRI: A Software Framework for Complex Human–Robot Interaction Scenarios on the iCub Humanoid Robot’. *Frontiers in Robotics and AI* 5 (2018). https://www.frontiersin.org/articles/10.3389/frobt.2018.00022.

Flasiński, Mariusz. ‘History of Artificial Intelligence’. In *Introduction to Artificial Intelligence*, edited by Mariusz Flasiński, 3–13. Switzerland: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-40022-8_1.

Forgas-Coll, Santiago, Ruben Huertas-Garcia, Antonio Andriella, and Guillem Alenyà. ‘How Do Consumers’ Gender and Rational Thinking Affect the Acceptance of Entertainment Social Robots?’ *International Journal of Social Robotics*, 19 December 2021. https://doi.org/10.1007/s12369-021-00845-y.

Frankish, Keith, and William M. Ramsey, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.

Franklin, Stan. ‘History, Motivations, and Core Themes’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 15–33. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.003.

Fukushima, Kunihiko. ‘Cognitron: A Self-Organizing Multilayered Neural Network’. *Biological Cybernetics* 20, no. 3 (1 September 1975): 121–36. https://doi.org/10.1007/BF00342633.

Fulford, Allison J., and Michael S. Harbuz. ‘An Introduction to the HPA Axis’. In *Techniques in the Behavioral and Neural Sciences*, edited by T. Steckler, N. H. Kalin, and J. M. H. M. Reul, 15:43–65. Handbook of Stress and the Brain. Elsevier, 2005. https://doi.org/10.1016/S0921-0709(05)80006-9.

Gallagher, S. ‘The Practice of Mind. Theory, Simulation or Primary Interaction?’ *Journal of Consciousness Studies* 8, no. 5–6 (1 May 2001): 83–108.

Gallagher, Shaun. ‘Direct Perception in the Intersubjective Context’. *Consciousness and Cognition*, Social Cognition, Emotion, and Self-Consciousness, 17, no. 2 (1 June 2008): 535–43. https://doi.org/10.1016/j.concog.2008.03.003.

———. ‘Inference or Interaction: Social Cognition without Precursors’. *Philosophical Explorations* 11, no. 3 (1 September 2008): 163–74. https://doi.org/10.1080/13869790802239227.

———. ‘Neurons, Neonates and Narrative’. In *Moving Ourselves, Moving Others: Motion and Emotion in Intersubjectivity, Consciousness and Language*, edited by Ad Foolen, Ulrike M. Lüdtke, Timothy P. Racine, and Jordan Zlatev, 165–96. Consciousness & Emotion Book Series 6. John Benjamins Publishing Company, 2012. https://doi.org/10.1075/ceb.6.07gal.

Ghafurian, Moojan, Colin Ellard, and Kerstin Dautenhahn. ‘Social Companion Robots to Reduce Isolation: A Perception Change Due to COVID-19’. In *Human-Computer Interaction – INTERACT 2021*, edited by Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, 43–63. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-85616-8_4.

Gibson, James Jerome. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin, 1979.

Goldman, Alvin I. ‘Interpretation Psychologized’. *Mind & Language* 4, no. 3 (1989): 161–85. https://doi.org/10.1111/j.1468-0017.1989.tb00249.x.

———. ‘Two Routes to Empathy: Insights from Cognitive Neuroscience’. In *Empathy: Philosophical and Psychological Perspectives*, edited by Amy Coplan and Peter Goldie, 31–44. Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199539956.003.0004.

Gompei, Takayuki, and Hiroyuki Umemuro. ‘Factors and Development of Cognitive and Affective Trust on Social Robots’. In *Social Robotics*, edited by Shuzhi Sam Ge, John-John Cabibihan, Miguel A. Salichs, Elizabeth Broadbent, Hongsheng He, Alan R. Wagner, and Álvaro Castro-González, 45–54. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-05204-1_5.

González-González, Carina Soledad, Verónica Violant-Holz, and Rosa Maria Gil-Iranzo. ‘Social Robots in Hospitals: A Systematic Review’. *Applied Sciences* 11, no. 13 (January 2021): 5976. https://doi.org/10.3390/app11135976.

Gottesmann, Claude. ‘GABA Mechanisms and Sleep’. *Neuroscience* 111, no. 2 (10 May 2002): 231–39. https://doi.org/10.1016/S0306-4522(02)00034-9.

Graaf, Maartje M. A. de, and Somaya Ben Allouch. ‘Exploring Influencing Variables for the Acceptance of Social Robots’. *Robotics and Autonomous Systems* 61, no. 12 (1 December 2013): 1476–86. https://doi.org/10.1016/j.robot.2013.07.007.

Guemghar, Imane, Paula Pires de Oliveira Padilha, Amal Abdel-Baki, Didier Jutras-Aswad, Jesseca Paquette, and Marie-Pascale Pomey. ‘Social Robot Interventions in Mental Health Care and Their Outcomes, Barriers, and Facilitators: Scoping Review’. *JMIR Mental Health* 9, no. 4 (19 April 2022): e36094. https://doi.org/10.2196/36094.

Gugerty, Leo. ‘Newell and Simon’s Logic Theorist: Historical Background and Impact on Cognitive Modeling’. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, no. 9 (1 October 2006): 880–84. https://doi.org/10.1177/154193120605000904.

Haenlein, Michael, and Andreas Kaplan. ‘A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence’. *California Management Review* 61, no. 4 (1 August 2019): 5–14. https://doi.org/10.1177/0008125619864925.

Haikonen, Pentti O. *Consciousness and Robot Sentience*. 2nd ed. Series on Machine Consciousness. World Scientific, 2019. https://doi.org/10.1142/11404.

———. *Robot Brains: Circuits and Systems for Conscious Machines*. John Wiley & Sons, 2007.

Haikonen, Pentti O. *The Cognitive Approach to Conscious Machines*. UK: Imprint Academic, 2003.

Haikonen, Pentti O. A. 'XCR-1: An Experimental Cognitive Robot Based on an Associative Neural Architecture'. *Cognitive Computation* 3, no. 2 (1 June 2011): 360–66. https://doi.org/10.1007/s12559-011-9100-9.

Hakli, Raul, and Johanna Seibt. 'Sociality and Normativity for Robots: An Introduction'. In *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*, edited by Raul Hakli and Johanna Seibt, 1–10. Studies in the Philosophy of Sociality. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-53133-5_1.

Harnad, Stevan. 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem'. *Minds and Machines* 1, no. 1 (1991): 43–54.

———. 'The Symbol Grounding Problem'. *Physica D: Nonlinear Phenomena* 42, no. 1 (1 June 1990): 335–46. https://doi.org/10.1016/0167-2789(90)90087-6.

Harvey, Inman, Ezequiel Di Paolo, Rachel Wood, Matt Quinn, and Elio Tuci. 'Evolutionary Robotics: A New Scientific Tool for Studying Cognition'. *Artificial Life* 11, no. 1–2 (1 January 2005): 79–98. https://doi.org/10.1162/1064546053278991.

Hasan, Ali. 'Comparative Study of Watt, Porter, Proell and Hartnell Governor Mechanism'. In *Advances in Mechanical Engineering*, edited by Gaurav Manik, Susheel Kalia, Sushanta Kumar Sahoo, Tarun K. Sharma, and Om Prakash Verma, 481–97. Singapore: Springer, 2021. https://doi.org/10.1007/978-981-16-0942-8_46.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 'Unsupervised Learning'. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, edited by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 485–585. Springer Series in Statistics. New York, NY: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7_14.

Hatfield, Elaine, Richard L. Rapson, and Yen-Chi L. Le. 'Emotional Contagion and Empathy'. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 19–30. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0003.

Hebb, D. O. *The Organization of Behavior; A Neuropsychological Theory*. The Organization of Behavior; a Neuropsychological Theory. Oxford, England: Wiley, 1949.

Hegel, Frank, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. 'Understanding Social Robots'. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, 169–74, 2009. https://doi.org/10.1109/ACHI.2009.51.

Hein, Grit, and Tania Singer. 'I Feel How You Feel but Not Always: The Empathic Brain and Its Modulation'. *Current Opinion in Neurobiology*, Cognitive neuroscience, 18, no. 2 (1 April 2008): 153–58. https://doi.org/10.1016/j.conb.2008.07.012.

Helming, Katharina A., Brent Strickland, and Pierre Jacob. 'Making Sense of Early False-Belief Understanding'. *Trends in Cognitive Sciences* 18, no. 4 (1 April 2014): 167–70. https://doi.org/10.1016/j.tics.2014.01.005.

Henderson, Lauren, Bala Maniam, and Hadley Leavell. 'The Silver Tsunami: Evaluating the Impact of Population Aging in the US'. *Journal of Business and Behavioral Sciences* 29, no. 2 (2017): 153–69.

Henrich, Joseph, Robert Boyd, and Peter J. Richerson. 'Five Misunderstandings About Cultural Evolution'. *Human Nature* 19, no. 2 (1 June 2008): 119–37. https://doi.org/10.1007/s12110-008-9037-1.

Heyes, Cecilia. 'Empathy Is Not in Our Genes'. *Neuroscience & Biobehavioral Reviews* 95 (1 December 2018): 499–507. https://doi.org/10.1016/j.neubiorev.2018.11.001.

Heyes, Cecilia, and Caroline Catmur. 'What Happened to Mirror Neurons?' *Perspectives on Psychological Science* 17, no. 1 (1 January 2022): 153–68. https://doi.org/10.1177/1745691621990638.

Heylighen, Francis, and Cliff Joslyn. 'Cybernetics and Second-Order Cybernetics'. In *Encyclopedia of Physical Science and Technology (Third Edition)*, edited by Robert A. Meyers, 155–69. New York: Academic Press, 2003. https://doi.org/10.1016/B0-12-227410-5/00161-7.

Hickok, Gregory. 'Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans'. *Journal of Cognitive Neuroscience* 21, no. 7 (1 July 2009): 1229–43. https://doi.org/10.1162/jocn.2009.21189.

———. *The Myth of Mirror Neurons: The Real Neuroscience of Communication and Cognition*. W. W. Norton & Company, 2014.

Hodson, Hal. 'The First Family Robot'. *New Scientist* 223, no. 2978 (19 July 2014): 21. https://doi.org/10.1016/S0262-4079(14)61389-0.

Hoffmann, Matej, and Rolf Pfeifer. 'Robots as Powerful Allies for the Study of Embodied Cognition from the Bottom Up'. In *The Oxford Handbook of 4E Cognition*, edited by Albert Newen, Leon De Bruin, and Shaun Gallagher, 841–62. Oxford University Press, 2018. https://doi.org/10.1093/oxfordhb/9780198735410.013.45.

Hofstadter, Douglas R. *I Am a Strange Loop*. New York: Basic Books, 2008.

Horowitz, Jason. 'Who Will Take Care of Italy's Older People? Robots, Maybe.' *The New York Times*, 25 March 2023, sec. World. https://www.nytimes.com/2023/03/25/world/europe/who-will-take-care-of-italys-older-people-robots-maybe.html.

Hu, Zebang. 'A Design of Service Robots in Epidemic Disease Isolation Environment'. *Academic Journal of Computing & Information Science* 4, no. 3 (15 May 2021). https://doi.org/10.25236/AJCIS.2021.040302.

Hudson, John. *The Robot Revolution: Understanding the Social and Economic Impact*. *The Robot Revolution*. Edward Elgar Publishing, 2019. https://www.elgaronline.com/view/9781788974479/9781788974479.xml.

Huijnen, Claire AGJ, Monique AS Lexis, and Luc P de Witte. 'Robots as New Tools in Therapy and Education for Children with Autism'. *International Journal of Neurorehabilitation* 04, no. 04 (2017). https://doi.org/10.4172/2376-0281.1000278.

Hume, David. *David Hume: A Treatise of Human Nature: Volume 1: Texts*. Edited by Edited by David Fate Norton and Mary J. Norton. Clarendon Hume Edition Series. Oxford, New York: Oxford University Press, 2007.

Husbands, Phil. 'Robotics'. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 269–95. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.018.

Iacoboni, Marco. 'Imitation, Empathy, and Mirror Neurons'. *Annual Review of Psychology* 60, no. 1 (2009): 653–70. https://doi.org/10.1146/annurev.psych.60.110707.163604.

———. ‘Within Each Other: Neural Mechanisms for Empathy in the Primate Brain’. In *Empathy: Philosophical and Psychological Perspectives*, edited by Amy Coplan and Peter Goldie, 45–57. Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199539956.003.0005.

Ickes, William. ‘Empathic Accuracy: Its Links to Clinical, Cognitive, Developmental, Social, and Physiological Psychology’. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 57–70. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0006.

Ishiguro, Nobu. ‘Care Robots in Japanese Elderly Care: Cultural Values in Focus’. In *The Routledge Handbook of Social Care Work Around the World*. Routledge, 2018.

Istituto Italiano di Tecnologia. ‘iCub History’. Istituto Italiano di Tecnologia iCub. Accessed 14 September 2023. https://icub.iit.it/web/icub/about-us/icub-history.

Ivanov, Stanislav, and Craig Webster. ‘Robots in Tourism: A Research Agenda for Tourism Economics’. *Tourism Economics* 26, no. 7 (1 November 2020): 1065–85. https://doi.org/10.1177/1354816619879583.

Jacob, K. S., P. Sharan, I. Mirza, M. Garrido-Cumbrera, S. Seedat, J. J. Mari, V. Sreenivas, and Shekhar Saxena. ‘Mental Health Systems in Countries: Where Are We Now?’ *The Lancet* 370, no. 9592 (22 September 2007): 1061–77. https://doi.org/10.1016/S0140-6736(07)61241-0.

Jacobs, Kerrin Artemis. ‘Digital Loneliness—Changes of Social Recognition through AI Companions’. *Frontiers in Digital Health* 6 (5 March 2024). https://doi.org/10.3389/fdgth.2024.1281037.

James, Jesin, Catherine Inez Watson, and Bruce MacDonald. ‘Artificial Empathy in Social Robots: An Analysis of Emotions in Speech’. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 632–37, 2018. https://doi.org/10.1109/ROMAN.2018.8525652.

Jamison, Robert N., Winston C. V. Parris, and Wayne S. Maxson. ‘Psychological Factors Influencing Recovery from Outpatient Surgery’. *Behaviour Research and Therapy* 25, no. 1 (1 January 1987): 31–37. https://doi.org/10.1016/0005-7967(87)90112-4.

Jeong, Sooyeon, Cynthia Breazeal, Deirdre Logan, and Peter Weinstock. 'Huggable: Impact of Embodiment on Promoting Verbal and Physical Engagement for Young Pediatric Inpatients'. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 121–26, 2017. https://doi.org/10.1109/ROMAN.2017.8172290.

Johal, Wafa. 'Research Trends in Social Robots for Learning'. *Current Robotics Reports* 1, no. 3 (1 September 2020): 75–83. https://doi.org/10.1007/s43154-020-00008-3.

Johnston, John. *The Allure of Machinic Life: Cybernetics, Artificial Life, and the New AI*. The MIT Press, 2008. https://doi.org/10.7551/mitpress/9780262101264.001.0001.

Jokinen, Kristiina, and Graham Wilcock. 'Expectations and First Experience with a Social Robot'. *HAI*, 2017. https://doi.org/10.1145/3125739.3132610.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education'. *Learning and Individual Differences* 103 (1 April 2023): 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kavaliers, Martin. 'Evolutionary and Comparative Aspects of Nociception'. *Brain Research Bulletin* 21, no. 6 (1 December 1988): 923–31. https://doi.org/10.1016/0361-9230(88)90030-5.

Kaynak, Okyay. 'The Golden Age of Artificial Intelligence'. *Discover Artificial Intelligence* 1, no. 1 (22 September 2021): 1. https://doi.org/10.1007/s44163-021-00009-x.

Keltner, Dacher, and Jonathan Haidt. 'Social Functions of Emotions at Four Levels of Analysis'. *Cognition and Emotion* 13, no. 5 (1 September 1999): 505–21. https://doi.org/10.1080/026999399379168.

Kohonen, Teuvo. 'Correlation Matrix Memories'. *IEEE Transactions on Computers* C–21, no. 4 (April 1972): 353–59. https://doi.org/10.1109/TC.1972.5008975.

Korb, Judith, and Jürgen Heinze. 'Major Hurdles for the Evolution of Sociality'. *Annual Review of Entomology* 61 (2016): 297–316. https://doi.org/10.1146/annurev-ento-010715-023711.

Krause, Mark Andrew, Dan Dolderman, Stephen Smith, and Daniel Paul Corts. *An Introduction to Psychological Science: Modeling Scientific Literacy*. Pearson Education Canada, 2014.

Kwon, Minae, Malte F. Jung, and Ross A. Knepper. 'Human Expectations of Social Robots'. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 463–64, 2016. https://doi.org/10.1109/HRI.2016.7451807.

Law, James, Mark Lee, Martin Hiilse, and Patricia Shaw. 'Infants and iCubs: Applying Developmental Psychology to Robot Shaping'. *Procedia Computer Science*, Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11), 7 (1 January 2011): 272–74. https://doi.org/10.1016/j.procs.2011.09.034.

LeDoux, Joseph. 'The Amygdala'. *Current Biology* 17, no. 20 (2007): R868–74. https://doi.org/10.1016/j.cub.2007.08.005.

LeDoux, Joseph E. 'The Amygdala: Contributions to Fear and Stress'. *Seminars in Neuroscience* 6, no. 4 (1 August 1994): 231–37. https://doi.org/10.1006/smns.1994.1030.

Lee, Mark. *How to Grow a Robot*. The MIT Press, 2020. https://mitpress.mit.edu/books/how-grow-robot.

Lee, Raymond S. T. *Artificial Intelligence in Daily Life*. Singapore: Springer, 2020. https://doi.org/10.1007/978-981-15-7695-9.

Lefkowitz, Melanie. 'Professor's Perceptron Paved the Way for AI – 60 Years Too Soon'. *Cornell Chronicle*, 25 September 2019. https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon.

Leite, Iolanda, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 'Empathic Robots for Long-Term Interaction'. *International Journal of Social Robotics* 6, no. 3 (1 August 2014): 329–41. https://doi.org/10.1007/s12369-014-0227-1.

Leite, Iolanda, Carlos Martinho, and Ana Paiva. 'Social Robots for Long-Term Interaction: A Survey'. *International Journal of Social Robotics* 5, no. 2 (2013): 291–308.

Leonelli, Sabina. 'Scientific Research and Big Data'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.

Levenson, Robert W., Sandy J. Lwi, Casey L. Brown, Brett Q. Ford, Marcela C. Otero, and Alice Verstaen. 'Emotion'. In *Handbook of Psychophysiology*, edited by Gary G. Berntson,

John T. Cacioppo, and Louis G. Tassinary, 4th ed., 444–64. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press, 2016. https://doi.org/10.1017/9781107415782.020.

Lighthill, James. 'Artificial Intelligence: A General Survey'. *Artificial Intelligence: A Paper Symposium*, 1973. https://www-formal.stanford.edu/jmc/reviews/lighthill/lighthill.html.

Lindquist, Kristen A. 'Emotions Emerge from More Basic Psychological Ingredients: A Modern Psychological Constructionist Model'. *Emotion Review* 5, no. 4 (1 October 2013): 356–68. https://doi.org/10.1177/1754073913489750.

Lindquist, Kristen A., Tor D. Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. 'The Brain Basis of Emotion: A Meta-Analytic Review'. *Behavioral and Brain Sciences* 35, no. 3 (June 2012): 121–43. https://doi.org/10.1017/S0140525X11000446.

Luisi, Pier Luigi. 'Autopoiesis: A Review and a Reappraisal'. *Naturwissenschaften* 90, no. 2 (1 February 2003): 49–59. https://doi.org/10.1007/s00114-002-0389-9.

Lungarella, Max, Fumiya Iida, Josh C. Bongard, and Rolf Pfeifer. 'AI in the 21st Century – With Historical Reflections'. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, edited by Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, 1–8. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-77296-5_1.

LuxAI S.A. 'Qtrobot Curriculum for Autism'. Accessed 7 June 2024. https://luxai.com/qtrobot-curriculum-for-autism/.

Malle, Bertram, and Stuti Thapa. 'What Kind of Mind Do I Want in My Robot?: Developing a Measure of Desired Mental Capacities in Social Robots', 195–96, 2017. https://doi.org/10.1145/3029798.3038378.

Malsburg, Chr. von der. 'Self-Organization of Orientation Sensitive Cells in the Striate Cortex'. *Kybernetik* 14, no. 2 (1 June 1973): 85–100. https://doi.org/10.1007/BF00288907.

Mar, Tanis, Vadim Tikhanoff, Giorgio Metta, and Lorenzo Natale. 'Self-Supervised Learning of Grasp Dependent Tool Affordances on the iCub Humanoid Robot', Vol. 2015, 2015. https://doi.org/10.1109/ICRA.2015.7139640.

Marocco, Davide, Angelo Cangelosi, Kerstin Fischer, and Tony Belpaeme. 'Grounding Action Words in the Sensorimotor Interaction with the World: Experiments with a Simulated

iCub Humanoid Robot'. *Frontiers in Neurorobotics* 4 (2010). https://www.frontiersin.org/articles/10.3389/fnbot.2010.00007.

Maron, Dina Fine. 'What to Do If You're Attacked by a Bear—or Any of These Other Wild Animals'. *Animals*, 17 August 2023. https://www.nationalgeographic.com/animals/article/survive-wildlife-encounters-bear-bison-shark-alligator.

Maturana, Humberto R., and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy of Science. Dordrecht, Holland: D. Reidel Pub. Co., 1980.

McCorduck, P., M. Minsky, O. Selfridge, and H. A. Simon. 'History of Artificial Intelligence'. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, 951–54. IJCAI'77. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1977.

McCorduck, Pamela. *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. [2nd ed.]. Natick, Massachusetts: AK Peters, 2004.

Mehri, Nader, Mahmood Messkoub, and Suzanne Kunkel. 'Trends, Determinants and the Implications of Population Aging in Iran'. *Ageing International* 45, no. 4 (1 December 2020): 327–43. https://doi.org/10.1007/s12126-020-09364-z.

Merriam-Webster.com Dictionary, s.v. "robot," accessed 23 February 2023. https://www.merriam-webster.com/dictionary/robot.

Merriam-Webster.com Dictionary, s.v. "sentient," accessed 24 April 2025. https://www.merriam-webster.com/dictionary/sentient.

Meskó, Bertalan, Gergely Hetényi, and Zsuzsanna Győrffy. 'Will Artificial Intelligence Solve the Human Resource Crisis in Healthcare?' *BMC Health Services Research* 18, no. 545 (13 July 2018). https://doi.org/10.1186/s12913-018-3359-4.

Mesoudi, Alex. 'Cultural Evolution: A Review of Theory, Findings and Controversies'. *Evolutionary Biology* 43, no. 4 (1 December 2016): 481–97. https://doi.org/10.1007/s11692-015-9320-0.

Metta, Giorgio, Lorenzo Natale, Francesco Nori, Giulio Sandini, David Vernon, Luciano Fadiga, Claes von Hofsten, et al. 'The iCub Humanoid Robot: An Open-Systems Platform for Research in Cognitive Development'. *Neural Networks*, Social Cognition: From Babies

to Robots, 23, no. 8 (1 October 2010): 1125–34. https://doi.org/10.1016/j.neunet.2010.08.010.

Metta, Giorgio, Giulio Sandini, D. Vernon, Lorenzo Natale, and Francesco Nori. 'iCub: The Open Humanoid Robot Designed for Learning and Developing Complex Cognitive Tasks'. *Proceedings of the 2019 IEEE RSJ International Conference on Intelligent Robots and Systems*, 2019.

Minsky, Marvin L. *Computation: Finite and Infinite Machines*. Englewood Cliffs, N.J.: Prentice-Hall, 1967.

MIT. 'People Overview Cynthia Breazeal'. MIT Media Lab. Accessed 22 December 2023. https://www.media.mit.edu/people/cynthiab/overview/.

Montag, Christiane, Jürgen Gallinat, and Andreas Heinz. 'Theodor Lipps and the Concept of Empathy: 1851–1914'. *American Journal of Psychiatry* 165, no. 10 (October 2008): 1261–1261. https://doi.org/10.1176/appi.ajp.2008.07081283.

Moor, James. 'The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years'. *AI Magazine* 27, no. 4 (2006): 87–91. https://doi.org/10.1609/aimag.v27i4.1911.

Moravec, Hans P. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, Mass: Harvard University Press, 1988.

———. *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press, 1999.

Morse, Anthony, Joachim Greeff, Tony Belpaeme, and Angelo Cangelosi. 'Epigenetic Robotics Architecture (ERA)'. *Autonomous Mental Development, IEEE Transactions On* 2 (1 January 2011): 325–39. https://doi.org/10.1109/TAMD.2010.2087020.

Moudatsou, Maria, Areti Stavropoulou, Anastas Philalithis, and Sofia Koukouli. 'The Role of Empathy in Health and Social Care Professionals'. *Healthcare* 8, no. 1 (30 January 2020): 26. https://doi.org/10.3390/healthcare8010026.

Mueller, Bridget, Alex Figueroa, and Jessica Robinson-Papp. 'Structural and Functional Connections Between the Autonomic Nervous System, Hypothalamic–Pituitary–Adrenal Axis, and the Immune System: A Context and Time Dependent Stress Response Network'. *Neurological Sciences* 43, no. 2 (1 February 2022): 951–60. https://doi.org/10.1007/s10072-021-05810-1.

Murari, Kartikeya, Ralph Etienne-Cummings, Nitish Thakor, and Gert Cauwenberghs. 'Which Photodiode to Use: A Comparison of CMOS-Compatible Structures'. *IEEE Sensors Journal* 9, no. 7 (July 2009): 752–60. https://doi.org/10.1109/JSEN.2009.2021805.

Natale, Lorenzo, Ali Paikan, Marco Randazzo, and Daniele E. Domenichelli. 'The iCub Software Architecture: Evolution and Lessons Learned'. *Frontiers in Robotics and AI* 3 (2016). https://www.frontiersin.org/articles/10.3389/frobt.2016.00024.

Nickerson, Raymond S., Susan F. Butler, and Michael Carlin. 'Empathy and Knowledge Projection'. In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 43–56. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0005.

Nilsson, Nils J. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9780511819346.

O'Connor, Jack. 'Undercover Algorithm: A Secret Chapter in the Early History of Artificial Intelligence and Satellite Imagery'. *International Journal of Intelligence and CounterIntelligence* 0, no. 0 (21 June 2022): 1–15. https://doi.org/10.1080/08850607.2022.2073542.

Olaronke, Iroju, Ojerinde Oluwaseun, and Ikono Rhoda. 'State Of The Art: A Study of Human-Robot Interaction in Healthcare'. *International Journal of Information Engineering and Electronic Business* 9, no. 3 (8 May 2017): 43–55. https://doi.org/10.5815/ijieeb.2017.03.06.

Olazaran, Mikel. 'A Sociological Study of the Official History of the Perceptrons Controversy'. *Social Studies of Science* 26, no. 3 (1 August 1996): 611–59. https://doi.org/10.1177/030631296026003005.

OpenAI. 'Sora: Creating Video from Text'. Accessed 27 February 2024. https://openai.com/sora.

O'Regan, Gerard. *A Brief History of Computing*. 2nd ed. London: Springer London, 2012. https://doi.org/10.1007/978-1-4471-2359-0.

Pantic, Maja, and Alessandro Vinciarelli. 'Social Signal Processing'. In *The Oxford Handbook of Affective Computing*, 2015. https://doi.org/10.1093/oxfordhb/9780199942237.013.027.

Park, Sung, and Mincheol Whang. ‘Empathy in Human–Robot Interaction: Designing for Social Robots’. *International Journal of Environmental Research and Public Health* 19, no. 3 (8 February 2022): 1889. https://doi.org/10.3390/ijerph19031889.

Parker, Kim. ‘Family Support in Graying Societies’. *Pew Research Center’s Social & Demographic Trends Project* (blog), 21 May 2015. https://www.pewresearch.org/social-trends/2015/05/21/family-support-in-graying-societies/.

Parmiggiani, Alberto, Marco Maggiali, Lorenzo Natale, Francesco Nori, Alexander Schmitz, Nikos Tsagarakis, José Santos-Victor, Francesco Becchi, Giulio Sandini, and Giorgio Metta. ‘The Design of the iCub Humanoid Robot’. *International Journal of Humanoid Robotics* 9 (1 December 2012): 1–24. https://doi.org/10.1142/S0219843612500272.

Pedersen, Isabel, Samantha Reid, and Kristen Aspevig. ‘Developing Social Robots for Aging Populations: A Literature Review of Recent Academic Sources’. *Sociology Compass* 12, no. 6 (2018): e12585.

Pennisi, Paola, Alessandro Tonacci, Gennaro Tartarisco, Lucia Billeci, Liliana Ruta, Sebastiano Gangemi, and Giovanni Pioggia. ‘Autism and Social Robotics: A Systematic Review’. *Autism Research: Official Journal of the International Society for Autism Research* 9, no. 2 (February 2016): 165–83. https://doi.org/10.1002/aur.1527.

Pepito, Joseph Andrew, Hirokazu Ito, Feni Betriana, Tetsuya Tanioka, and Rozzano C. Locsin. ‘Intelligent Humanoid Robots Expressing Artificial Humanlike Empathy in Nursing Situations’. *Nursing Philosophy* 21, no. 4 (2020): e12318. https://doi.org/10.1111/nup.12318.

Pester, Patrick. ‘Humans Are Practically Defenseless. Why Don’t Wild Animals Attack Us More?’ *Livescience.Com* (blog), 12 July 2021. https://www.livescience.com/why-predators-dont-attack-humans.html.

Pfeifer, Jennifer, and Mirella Dapretto. ‘“Mirror, Mirror, in My Mind”: Empathy, Interpersonal Competence, and the Mirror Neuron System’, 183–98, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0015.

Phelps, Elizabeth A. ‘Emotion and Cognition: Insights from Studies of the Human Amygdala’. *Annual Review of Psychology* 57, no. 1 (2006): 27–53. https://doi.org/10.1146/annurev.psych.56.091103.070234.

Pitts, Walter. ‘Comments on Session on Learning Machines’. In *Proceedings of the March 1-3, 1955*, 108–11. AFIPS ’55 (Western). Los Angeles, CA: Association for Computing Machinery, 1955. https://doi.org/10.1145/1455292.1455313.

Plutchik, Robert. ‘Evolutionary Bases of Empathy’. In *Empathy and Its Development*, edited by Nancy Eisenberg and Janet Strayer, 38–46. Cambridge: Cambridge University Press, 1987.

Pohl, Martin. ‘Robotic Systems in Healthcare with Particular Reference to Innovation in the “Fourth Industrial Revolution”’ 8 (2016): 17–33.

Polak, Ronit Feingold, and Shelly Levy Tzedek. ‘Social Robot for Rehabilitation: Expert Clinicians and Post-Stroke Patients’ Evaluation Following a Long-Term Intervention’. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 151–60. HRI ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3319502.3374797.

Pollack, Jordan B., Hod Lipson, Gregory Hornby, and Pablo Funes. ‘Three Generations of Automatically Designed Robots’. *Artificial Life* 7, no. 3 (1 July 2001): 215–23. https://doi.org/10.1162/106454601753238627.

Porter, Richard E., and Larry A. Samovar. ‘Cultural Influences on Emotional Expression: Implications for Intercultural Communication’. In *Handbook of Communication and Emotion*, edited by Peter A. Andersen and Laura K. Guerrero, 451–72. San Diego: Academic Press, 1996. https://doi.org/10.1016/B978-012057770-5/50019-9.

Portner, Paul. *What Is Meaning? Fundamentals of Formal Semantics*. Fundamentals of Linguistics. Malden, MA: Blackwell Pub, 2005.

Prochazkova, Eliska, and Mariska E. Kret. ‘Connecting Minds and Sharing Emotions through Mimicry: A Neurocognitive Model of Emotional Contagion’. *Neuroscience & Biobehavioral Reviews* 80 (1 September 2017): 99–114. https://doi.org/10.1016/j.neubiorev.2017.05.013.

Quick, Oliver Santiago. ‘Empathizing and Sympathizing With Robots: Implications for Moral Standing’. *Frontiers in Robotics and AI* 8 (2022). https://www.frontiersin.org/articles/10.3389/frobt.2021.791527.

Raczaszek-Leonardi, Joanna, Iris Nomikou, and Katharina J. Rohlfing. ‘Young Children’s Dialogical Actions: The Beginnings of Purposeful Intersubjectivity’. *IEEE Transactions on Autonomous Mental Development* 5, no. 3 (September 2013): 210–21. https://doi.org/10.1109/TAMD.2013.2273258.

Revsbech, Jeppe Kiel. ‘When Humans and Robots Meet’. *School of Culture and Society News* (blog), 28 February 2022. https://cas.au.dk/en/currently/news/nyhedsarkiv/when-humans-and-robots-meet.

Ribas, Jordi. ‘Building the New Bing: Image Creator’. *Microsoft Bing Blogs* (blog), 23 March 2023. https://blogs.bing.com/search-quality-insights/march-2023/Building-the-New-Bing-Image-Creator/.

Robiner, William N. ‘The Mental Health Professions: Workforce Supply and Demand, Issues, and Challenges’. *Clinical Psychology Review* 26, no. 5 (1 September 2006): 600–625. https://doi.org/10.1016/j.cpr.2006.05.002.

Robinson, Nicole Lee, Timothy Vaughan Cottier, and David John Kavanagh. ‘Psychosocial Health Interventions by Social Robots: Systematic Review of Randomized Controlled Trials’. *Journal of Medical Internet Research* 21, no. 5 (10 May 2019): e13203. https://doi.org/10.2196/13203.

Robot Self-Consciousness. XCR-1 Passes the Mirror Test. Accessed 29 August 2024. https://youtube.com/watch?v=WE9QsQqsAdo.

Robot Sequence Memory, 2020. https://www.youtube.com/watch?v=8DNA6_wKVJQ.

Rogers, Stephanie. ‘BRILLO the Bartending Robot Can Fulfill Your Social Needs While Slinging Cocktails’. *Dornob* (blog), 29 August 2022. https://dornob.com/brillo-the-bartending-robot-can-fulfill-your-social-needs-while-slinging-cocktails/.

Rojas, R. ‘Konrad Zuse’s Legacy: The Architecture of the Z1 and Z3’. *IEEE Annals of the History of Computing* 19, no. 2 (April 1997): 5–16. https://doi.org/10.1109/85.586067.

Rojas, Raúl, and Ulf Hashagen. ‘“Nothing New Since von Neumann”: A Historian Looks at Computer Architecture, 1945–1995’. In *The First Computers: History and Architectures*, 195–217. MIT Press, 2002. https://ieeexplore.ieee.org/document/6302806.

Rosenfeld, Edward. ‘Teuvo Kohonen’. In *Talking Nets: An Oral History of Neural Networks*, edited by James A. Andersen, and Edward Rosenfeld, 144–64. The MIT Press, 2000. https://direct.mit.edu/books/book/4886/chapter/622911/Teuvo-Kohonen.

Rowland, Donald T. ‘Global Population Aging: History and Prospects’. In *International Handbook of Population Aging*, edited by Peter Uhlenberg, 37–65. Dordrecht: Springer Netherlands, 2009. https://doi.org/10.1007/978-1-4020-8356-3_3.

Russell, James A. ‘Core Affect and the Psychological Construction of Emotion’. *Psychological Review* 110, no. 1 (2003): 145–72. https://doi.org/10.1037/0033-295X.110.1.145.

———. ‘Emotion, Core Affect, and Psychological Construction’. *Cognition and Emotion* 23, no. 7 (1 November 2009): 1259–83. https://doi.org/10.1080/02699930902809375.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3 edition. Upper Saddle River: Pearson, 2009.

Šabanović, Selma, Casey C. Bennett, Wan-Ling Chang, and Lesa Huber. ‘PARO Robot Affects Diverse Interaction Modalities in Group Sensory Therapy for Older Adults with Dementia’. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, 1–6, 2013. https://doi.org/10.1109/ICORR.2013.6650427.

Sandini, G., G. Metta, and D. Vernon. ‘RobotCub: An Open Framework for Research in Embodied Cognition’. In *4th IEEE/RAS International Conference on Humanoid Robots, 2004.*, 1:13-32 Vol. 1, 2004. https://doi.org/10.1109/ICHR.2004.1442111.

Sandini, Giulio, Giorgio Metta, and David Vernon. ‘The iCub Cognitive Humanoid Robot: An Open-System Research Platform for Enactive Cognition’. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, edited by Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, 358–69. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-77296-5_32.

Scarantino, Andrea, and Paul Griffiths. ‘Don’t Give Up on Basic Emotions’. *Emotion Review* 3, no. 4 (1 October 2011): 444–54. https://doi.org/10.1177/1754073911410745.

Scarantino, Andrea, and Ronald de Sousa. ‘Emotion’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/sum2021/entries/emotion/.

Scassellati, Brian, Henny Admoni, and Maja Matarić. ‘Robots for Use in Autism Research’. *Annual Review of Biomedical Engineering* 14, no. 1 (2012): 275–94. https://doi.org/10.1146/annurev-bioeng-071811-150036.

Schertz, Matthew Victor. ‘Empathy as Intersubjectivity: Resolving Hume and Smith’s Divide’. *Studies in Philosophy and Education* 26, no. 2 (1 March 2007): 165–78. https://doi.org/10.1007/s11217-006-9022-2.

Schidelko, Lydia P., Michael Huemer, Lara M. Schröder, Anna S. Lueb, Josef Perner, and Hannes Rakoczy. ‘Why Do Children Who Solve False Belief Tasks Begin to Find True Belief Control Tasks Difficult? A Test of Pragmatic Performance Factors in Theory of Mind Tasks’. *Frontiers in Psychology* 12 (14 January 2022). https://doi.org/10.3389/fpsyg.2021.797246.

Schimmack, Ulrich, and Alexander Grob. ‘Dimensional Models of Core Affect: A Quantitative Comparison by Means of Structural Equation Modeling’. *European Journal of Personality* 14, no. 4 (2000): 325–45. https://doi.org/10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I.

Scoglio, Arielle AJ, Erin D. Reilly, Jay A. Gorman, and Charles E. Drebing. ‘Use of Social Robots in Mental Health and Well-Being Research: Systematic Review’. *Journal of Medical Internet Research* 21, no. 7 (24 July 2019): e13322. https://doi.org/10.2196/13322.

Seibt, Johanna. ‘Towards an Ontology of Simulated Social Interaction: Varieties of the “As If” for Robots and Humans’. In *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*, edited by Raul Hakli and Johanna Seibt, 11–39. Studies in the Philosophy of Sociality. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-53133-5_2.

Sejnowski, Terrence J. *The Deep Learning Revolution*. The MIT Press, 2018. https://doi.org/10.7551/mitpress/11474.001.0001.

———. *The Deep Learning Revolution*. Cambridge, Massachusetts: MIT Press, 2018. https://mitpress.mit.edu/9780262038034/the-deep-learning-revolution/.

Shamay-Tsoory, Simone G. ‘Empathic Processing: Its Cognitive and Affective Dimensions and Neuroanatomical Basis’. In *The Social Neuroscience of Empathy*, edited by Jean Decety

and William Ickes, 215–32. The MIT Press, 2009. https://doi.org/10.7551/mitpress/9780262012973.003.0017.

Shannon, Claude E. 'Programming a Computer for Playing Chess'. *Philosophical Magazine*, 7, 41, no. 314 (1 March 1950): 256–75. https://doi.org/10.1080/14786445008521796.

Shapiro, Lawrence, and Shannon Spaulding. 'Embodied Cognition'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/.

Sharkey, Amanda, and Noel Sharkey. 'Granny and the Robots: Ethical Issues in Robot Care for the Elderly'. *Ethics and Information Technology* 14, no. 1 (1 March 2012): 27–40. https://doi.org/10.1007/s10676-010-9234-6.

Shaw, Patricia, James Law, and Mark Lee. 'Representations of Body Schemas for Infant Robot Development'. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 123–28, 2015. https://doi.org/10.1109/DEVLRN.2015.7346128.

Shibata, Takanori. 'Therapeutic Seal Robot as Biofeedback Medical Device: Qualitative and Quantitative Evaluations of Robot Therapy in Dementia Care'. *Proceedings of the IEEE* 100, no. 8 (August 2012): 2527–38. https://doi.org/10.1109/JPROC.2012.2200559.

Siegler, Robert S., Judy S. DeLoache, and Nancy Eisenberg. *How Children Develop*. Fourth Canadian Edition. Macmillan, 2014.

Singer, Stephen. 'With Endless Patience and Never Tiring, Robots Are Being Used in Connecticut to Connect with Children with Autism'. *Hartford Courant*, 28 August 2022. https://www.courant.com/news/connecticut/hc-news-robotics-autistic-children-20220828-tvmcwygg45eltmmfqzy24nibqu-story.html.

Snell, K. D. M. 'The Rise of Living Alone and Loneliness in History'. *Social History* 42, no. 1 (2 January 2017): 2–28. https://doi.org/10.1080/03071022.2017.1256093.

Stano, Pasquale, Chrystopher Nehaniv, Takashi Ikegami, Luisa Damiano, and Olaf Witkowski. 'Autopoiesis: Foundations of Life, Cognition, and Emergence of Self/Other'. *Biosystems* 232 (1 October 2023): 105008. https://doi.org/10.1016/j.biosystems.2023.105008.

Steels, Luc. ‘Fifty Years of AI: From Symbols to Embodiment - and Back’. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, edited by Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, 18–28. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-77296-5_3.

Stein, Edith. *On the Problem of Empathy*. Dordrecht: Springer Netherlands, 1964. https://doi.org/10.1007/978-94-017-5546-7.

Stone, Robyn, and Mary F. Harahan. ‘Improving The Long-Term Care Workforce Serving Older Adults’. *Health Affairs* 29, no. 1 (January 2010): 109–15. https://doi.org/10.1377/hlthaff.2009.0554.

Stueber, Karsten. ‘Empathy’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/fall2019/entries/empathy/.

Stueber, Karsten R. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. The MIT Press, 2006. https://www.jstor.org/stable/j.ctt5hhbjr.

Sun, Ron. ‘Connectionism and Neural Networks’. In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 108–27. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139046855.008.

Surkalim, Daniel L., Mengyun Luo, Robert Eres, Klaus Gebel, Joseph van Buskirk, Adrian Bauman, and Ding Ding. ‘The Prevalence of Loneliness across 113 Countries: Systematic Review and Meta-Analysis’. *BMJ* 376 (9 February 2022): e067068. https://doi.org/10.1136/bmj-2021-067068.

Suzuki, Toru. *Low Fertility and Population Aging in Japan and Eastern Asia*. SpringerBriefs in Population Studies. Tokyo: Springer Japan, 2013. https://doi.org/10.1007/978-4-431-54780-8.

Tanevska, Ana, Francesco Rea, Giulio Sandini, Lola Cañamero, and Alessandra Sciutti. ‘A Cognitive Architecture for Socially Adaptable Robots’. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 195–200, 2019. https://doi.org/10.1109/DEVLRN.2019.8850688.

Tanevska, Ana, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. ‘Towards an Affective Cognitive Architecture for Human-Robot Interaction for the iCub Robot’, 2017. https://inria.hal.science/hal-01615491.

TenHouten, Warren D. ‘Basic Emotion Theory, Social Constructionism, and the Universal Ethogram’. *Social Science Information* 60, no. 4 (1 December 2021): 610–30. https://doi.org/10.1177/05390184211046481.

‘The Orocos Project’. Accessed 19 September 2023. https://orocos.org/.

Thomaz, Andrea, Guy Hoffman, and Maya Cakmak. ‘Computational Human-Robot Interaction’. *Foundations and Trends in Robotics* 4, no. 2–3 (20 December 2016): 105–223. https://doi.org/10.1561/2300000049.

Thompson, Evan. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind a Book by Evan Thompson*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 2007.

———. ‘Sensorimotor Subjectivity and the Enactive Approach to Experience’. *Phenomenology and the Cognitive Sciences* 4, no. 4 (1 December 2005): 407–27. https://doi.org/10.1007/s11097-005-9003-x.

Thunberg, Sofia, and Tom Ziemke. ‘Are People Ready for Social Robots in Public Spaces?’ In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 482–84. HRI ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3371382.3378294.

Tikhanoff, V., A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori. ‘An Open-Source Simulator for Cognitive Robotics Research: The Prototype of the iCub Humanoid Robot Simulator’. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, 57–61. PerMIS ’08. New York, NY, USA: Association for Computing Machinery, 2008. https://doi.org/10.1145/1774674.1774684.

Tikhanoff, Vadim, Angelo Cangelosi, and Giorgio Metta. ‘Integration of Speech and Action in Humanoid Robots: iCub Simulation Experiments’. *IEEE Transactions on Autonomous Mental Development* 3, no. 1 (March 2011): 17–29. https://doi.org/10.1109/TAMD.2010.2100390.

Tomasello, Michael, and Ann Cale Kruger. ‘Joint Attention on Actions: Acquiring Verbs in Ostensive and Non-Ostensive Contexts*’. *Journal of Child Language* 19, no. 2 (June 1992): 311–33. https://doi.org/10.1017/S0305000900011430.

Tremblay, Marie-Pier B., Isabelle Deschamps, Béatrice Tousignant, and Philip L. Jackson. ‘Functional Connectivity Patterns of Trait Empathy Are Associated with Age’. *Brain and Cognition* 159 (1 June 2022): 105859. https://doi.org/10.1016/j.bandc.2022.105859.

Tsigos, Constantine, and George P Chrousos. ‘Hypothalamic–Pituitary–Adrenal Axis, Neuroendocrine Factors and Stress’. *Journal of Psychosomatic Research* 53, no. 4 (1 October 2002): 865–71. https://doi.org/10.1016/S0022-3999(02)00429-4.

Turing, Alan M. ‘Computing Machinery and Intelligence.’ *Mind* 49 (1950): 433–60.

Turkle, Sherry. *The Second Self: Computers and the Human Spirit*, 2005. https://doi.org/10.7551/mitpress/6115.001.0001.

Uhlenberg, Peter, ed. *International Handbook of Population Aging*. International Handbooks of Population 1. Springer, 2009. 10.1007/978-1-4020-8356-3.

University of Illinois Springfield Philosophy Department. ‘Faculty & Staff’. Accessed 29 August 2024. https://www.uis.edu/philosophy/faculty-staff.

Vallverdú, Jordi, and David Casacuberta. ‘Ethical and Technical Aspects of Emotions to Create Empathy in Medical Machines’. In *Machine Medical Ethics*, edited by Simon Peter van Rysewyk and Matthijs Pontier, 341–62. Intelligent Systems, Control and Automation: Science and Engineering. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-08108-3_20.

Varela, F. G., H. R. Maturana, and R. Uribe. ‘Autopoiesis: The Organization of Living Systems, Its Characterization and a Model’. *Biosystems* 5, no. 4 (1 May 1974): 187–96. https://doi.org/10.1016/0303-2647(74)90031-8.

Varela, Francisco J. *Principles of Biological Autonomy*. North Holland Series in General Systems Research 2. New York : North Holland, 1979: New York : North Holland, 1979.

Varela, Francisco J., Eleanor Rosch, and Evan Thompson. *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, 1991. https://doi.org/10.7551/mitpress/6730.001.0001.

Vercelli, Alessandro, Innocenzo Rainero, Ludovico Ciferri, Marina Boido, and Fabrizio Pirri. 'Robots in Elderly Care'. *DigitCult - Scientific Journal on Digital Cultures* 2, no. 2 (6 March 2018): 37–50. https://doi.org/10.4399/97888255088954.

Vernon, David, and Dermot Furlong. 'Philosophical Foundations of AI'. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, edited by Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, 53–62. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-77296-5_6.

Vernon, David, Giorgio Metta, and Giulio Sandini. 'The iCub Cognitive Architecture: Interactive Development in a Humanoid Robot'. In *2007 IEEE 6th International Conference on Development and Learning*, 122–27, 2007. https://doi.org/10.1109/DEVLRN.2007.4354038.

Waal, Frans B. M. de. 'Empathy in Primates and Other Mammals'. In *Empathy: From Bench to Bedside*, edited by Jean Decety, 87–106. The MIT Press, 2011. https://doi.org/10.7551/mitpress/9780262016612.003.0006.

Waal, Frans B. M. de, and Stephanie D. Preston. 'Mammalian Empathy: Behavioural Manifestations and Neural Basis'. *Nature Reviews Neuroscience* 18, no. 8 (August 2017): 498–509. https://doi.org/10.1038/nrn.2017.72.

Waltz, Emily. 'Therapy Robot Teaches Social Skills to Children With Autism'. *IEEE Spectrum*, 9 August 2018. https://spectrum.ieee.org/robot-therapy-for-autism.

Wang, Philip S., Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C. Angermeyer, Guilherme Borges, Evelyn J. Bromet, Ronny Bruffaerts, et al. 'Use of Mental Health Services for Anxiety, Mood, and Substance Disorders in 17 Countries in the WHO World Mental Health Surveys'. *The Lancet* 370, no. 9590 (8 September 2007): 841–50. https://doi.org/10.1016/S0140-6736(07)61414-7.

Warneken, Felix, and Michael Tomasello. 'Altruistic Helping in Human Infants and Young Chimpanzees'. *Science* 311, no. 5765 (3 March 2006): 1301–3. https://doi.org/10.1126/science.1121448.

———. 'Extrinsic Rewards Undermine Altruistic Tendencies in 20-Month-Olds'. *Developmental Psychology* 44, no. 6 (2008): 1785–88. https://doi.org/0.1037/a0013860.

Watson, David S. ‘On the Philosophy of Unsupervised Learning’. *Philosophy & Technology* 36, no. 2 (21 April 2023): 28. https://doi.org/10.1007/s13347-023-00635-6.

Waytz, Adam, Kurt Gray, Nicholas Epley, and Daniel M. Wegner. ‘Causes and Consequences of Mind Perception’. *Trends in Cognitive Sciences* 14, no. 8 (1 August 2010): 383–88. https://doi.org/10.1016/j.tics.2010.05.006.

Waytz, Adam, Carey Morewedge, Nicholas Epley, George Monteleone, Jia-Hong Gao, and John Cacioppo. ‘Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism’. *Journal of Personality and Social Psychology* 99 (1 September 2010): 410–35. https://doi.org/10.1037/a0020240.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. 1st edition. San Francisco: W H Freeman & Co, 1976.

———. ‘ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine’. Edited by A.G. Oettinger. *Communications of the ACM* 9, no. 1 (January 1966): 36–45. https://doi.org/10.1145/365153.365168.

Wheeler, Michael. ‘Martin Heidegger’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/fall2020/entries/heidegger/.

Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. Second. Cambridge, MA: The MIT Press, 1948. https://doi.org/10.7551/mitpress/11810.001.0001.

———. *The Human Use of Human Beings: Cybernetics and Society*. 2d ed. rev. Garden City, NY: Doubleday, 1954.

Williams, M.R. ‘The Origins, Uses, and Fate of the EDVAC’. *IEEE Annals of the History of Computing* 15, no. 1 (1993): 22–38. https://doi.org/10.1109/85.194089.

Winkielman, Piotr, Joshua D. Davis, and Seana Coulson. ‘Moving Thoughts: Emotion Concepts from the Perspective of Context Dependent Embodied Simulation’. *Language, Cognition and Neuroscience* 38, no. 10 (26 November 2023): 1531–53. https://doi.org/10.1080/23273798.2023.2236731.

Wirtz, Jochen, Werner Kunz, and Stefanie Paluch. ‘The Service Revolution, Intelligent Automation and Service Robots’. *European Business Review* 29, no. 5 (2021): 909.

Wispé, Lauren. ‘History of the Concept of Empathy’. In *Empathy and Its Development*, edited by Nancy Eisenberg and Janet Strayer, 17–37. Cambridge: Cambridge University Press, 1987.

Wooldridge, Michael. *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*. New York: Flatiron Books, 2020. https://us.macmillan.com/books/9781250770738/abriefhistoryofartificialintelligence.

Wright, James. ‘Inside Japan’s Long Experiment in Automating Elder Care’. *MIT Technology Review* (blog), 9 January 2023. https://www.technologyreview.com/2023/01/09/1065135/japan-automating-eldercare-robots/.

Young, Rupert. ‘A General Architecture for Robotics Systems: A Perception-Based Approach to Artificial Life’. *Artificial Life* 23, no. 2 (1 May 2017): 236–86. https://doi.org/10.1162/ARTL_a_00229.

YouTube. ‘Pentti Haikonen’. Accessed 6 March 2024. https://www.youtube.com/@PenHaiko.

Yu, Ruby, Elsie Hui, Jenny Lee, Dawn Poon, Ashley Ng, Kitty Sit, Kenny Ip, et al. ‘Use of a Therapeutic, Socially Assistive Pet Robot (PARO) in Improving Mood and Stimulating Social Interaction and Communication for People With Dementia: Study Protocol for a Randomized Controlled Trial’. *JMIR Research Protocols* 4, no. 2 (1 May 2015): e45. https://doi.org/10.2196/resprot.4189.

Zahavi, Dan. ‘Empathy and Other-Directed Intentionality’. *Topoi* 33, no. 1 (2014): 129–42. https://doi.org/10.1007/s11245-013-9197-4.

Zahavi, Dan, and Philippe Rochat. ‘Empathy≠sharing: Perspectives from Phenomenology and Developmental Psychology’. *Consciousness and Cognition* 36 (1 November 2015): 543–53. https://doi.org/10.1016/j.concog.2015.05.008.

Ziemke, Tom. ‘The Body of Knowledge: On the Role of the Living Body in Grounding Embodied Cognition’. *Biosystems* 148, no. Complete (2016): 4–11. https://doi.org/10.1016/j.biosystems.2016.08.005.