The Handbook on Socially Interactive Agents
20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics
Volume 2: Interactivity, Platforms, Application

Birgit Lugrin, Catherine Pelachaud, David Traum, (Editors)

The Handbook on Socially Interactive Agents provides a comprehensive overview of the research fields of Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics. Socially Interactive Agents (SIAs), whether virtually or physically embodied, are autonomous agents that are able to perceive an environment including people or other agents, reason, and decide how to interact, and express attitudes such as emotions, engagement, or empathy. They are capable of interacting with people and each other in a socially intelligent manner using multimodal communicative behaviors with the goal to support humans in various domains.

Written by international experts in their respective fields, the book summarizes research in the many important research communities pertinent for SIAs, while discussing current challenges and future directions. The handbook provides easy access to modeling and studying SIAs for researchers and students and aims at further bridging the gap between the research communities involved.

In two volumes, the book clearly structures the vast body of research. The first volume starts by introducing what is involved in SIAs research, in particular research methodologies and ethical implications of developing SIAs. It further examines research on appearance and behavior, focusing on multimodality. Finally, social cognition for SIAs is investigated by different theoretical models and phenomena such as theory of mind or pro-sociality. The second volume starts with perspectives on interaction, examined from different angles such as interaction in social space, group interaction, or long-term interaction. It also includes an extensive overview summarizing research and systems of human-agent platforms and of some of the major application areas of SIAs such as education, aging support, autism or games.
The Handbook on Socially Interactive Agents
ACM Books

Editors in Chief
Sanjiva Prasad, Indian Institute of Technology (IIT) Delhi, India
Marta Kwiatkowska, University of Oxford, UK
Charu Aggarwal, IBM Corporation, USA

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

Effective Theories in Programming Practice
Jayadev Misra, The University of Texas at Austin, US
2022

Spatial Gems, Volume 1
Editors: John Krumm, Microsoft Research, Microsoft Corporation, Redmond, WA, USA
Andreas Züfle, Geography and Geoinformation Science Department, George Mason University, Fairfax, VA, USA
Cyrus Shahabi, Computer Science Department, University of Southern California, Los Angeles, CA, USA
2022

Edsger Wybe Dijkstra: His Life, Work, and Legacy
Editors: Krzysztof R. Apt, CWI, Amsterdam and University of Warsaw
Tony Hoare, University of Cambridge and Microsoft Research Ltd
2022

Weaving Fire into Form: Aspirations for Tangible and Embodied Interaction
Brygg Ullmer, Clemson University
Orit Shaer, Wellesley College
Ali Mazalek, Toronto Metropolitan University
Caroline Hummels, Eindhoven University of Technology
2022

Linking the World’s Information: A Collection of Essays on the Work of Sir Tim Berners-Lee
Oshani Seneviratne, Rensselaer Polytechnic Institute
James A. Hendler, Rensselaer Polytechnic Institute
2022

Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman
Editor: Rebecca Slayton, Cornell University
2022
Applied Affective Computing
Leimin Tian, Monash University
Sharon Oviatt, Monash University
Michal Muszynski, Carnegie Mellon University and University of Geneva
Brent C. Chamberlain, Utah State University
Jennifer Healey, Adobe Research, San Jose
Akane Sano, Rice University
2022

James L. Pelkey
Andrew L. Russell, SUNY Polytechnic Institute, New York
Loring G. Robbins
2022

Theories of Programming: The Life and Works of Tony Hoare
Editors: Cliff B. Jones, Newcastle University, UK
Jayadev Misra, The University of Texas at Austin, US
2021

Software: A Technical History
Kim W. Tracy, Rose-Hulman Institute of Technology, IN, USA
2021

The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition
Editors: Birgit Lugrin, Julius-Maximilians-Universität of Würzburg
Catherine Pelachaud, CNRS-ISIR, Sorbonne Université
David Traum, University of Southern California
2021

Probabilistic and Causal Inference: The Works of Judea Pearl
Editors: Hector Geffner, ICREA and Universitat Pompeu Fabra
Rina Dechter, University of California, Irvine
Joseph Y. Halpern, Cornell University
2022

Event Mining for Explanatory Modeling
Laleh Jalali, University of California, Irvine (UCI), Hitachi America Ltd.
Ramesh Jain, University of California, Irvine (UCI)
2021

Editors: Parisa Eslambolehilar, Cardiff University, Wales, UK
Andreas Komninos, University of Patras, Greece
Mark Dunlop, *Strathclyde University, Scotland, UK*
2021

**Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL, Third Edition**
Dean Allemang, *Working Ontologist LLC*
Jim Hendler, *Rensselaer Polytechnic Institute*
Fabien Gandon, *INRIA*
2020

**Code Nation: Personal Computing and the Learn to Program Movement in America**
Michael J. Halvorson, *Pacific Lutheran University*
2020

**Computing and the National Science Foundation, 1950–2016: Building a Foundation for Modern Computing**
Peter A. Freeman, *Georgia Institute of Technology*
W. Richards Adrion, *University of Massachusetts Amherst*
William Aspray, *University of Colorado Boulder*
2019

**Providing Sound Foundations for Cryptography: On the work of Shafi Goldwasser and Silvio Micali**
Oded Goldreich, *Weizmann Institute of Science*
2019

**Concurrency: The Works of Leslie Lamport**
Dahlia Malkhi, *VMware Research* and *Calibra*
2019

**The Essentials of Modern Software Engineering: Free the Practices from the Method Prisons!**
Ivar Jacobson, *Ivar Jacobson International*
Harold “Bud” Lawson, *Lawson Konsult AB (deceased)*
Pan-Wei Ng, *DBS Singapore*
Paul E. McMahon, *PEM Systems*
Michael Goedicke, *Universität Duisburg–Essen*
2019

**Data Cleaning**
Ihab F. Ilyas, *University of Waterloo*
Xu Chu, *Georgia Institute of Technology*
2019

**Conversational UX Design: A Practitioner’s Guide to the Natural Conversation Framework**
Robert J. Moore, *IBM Research–Almaden*
Raphael Arar, *IBM Research–Almaden*
2019

**Heterogeneous Computing: Hardware and Software Perspectives**
Mohamed Zahran, *New York University*
2019

**Hardness of Approximation Between P and NP**
Aviad Rubinstein, *Stanford University*
2019

**The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions**
Editors: Sharon Oviatt, *Monash University*
Björn Schuller, *Imperial College London and University of Augsburg*
Philip R. Cohen, *Monash University*
Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*
Gerasimos Potamianos, *University of Thessaly*
Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*
2019

**Making Databases Work: The Pragmatic Wisdom of Michael Stonebraker**
Editor: Michael L. Brodie, *Massachusetts Institute of Technology*
2018

**The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition**
Editors: Sharon Oviatt, *Monash University*
Björn Schuller, *University of Augsburg and Imperial College London*
Philip R. Cohen, *Monash University*
Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*
Gerasimos Potamianos, *University of Thessaly*
Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*
2018

**Declarative Logic Programming: Theory, Systems, and Applications**
Editors: Michael Kifer, *Stony Brook University*
Yanhong Annie Liu, *Stony Brook University*
2018

**The Sparse Fourier Transform: Theory and Practice**
Haitham Hassanieh, *University of Illinois at Urbana-Champaign*
2018

**The Continuing Arms Race: Code-Reuse Attacks and Defenses**
Editors: Per Larsen, *Immunant, Inc.*
Ahmad-Reza Sadeghi, *Technische Universität Darmstadt*
2018
Frontiers of Multimedia Research
Editor: Shih-Fu Chang, Columbia University
2018

Shared-Memory Parallelism Can Be Simple, Fast, and Scalable
Julian Shun, University of California, Berkeley
2017

Computational Prediction of Protein Complexes from Protein Interaction Networks
Sriganesh Srihari, The University of Queensland Institute for Molecular Bioscience
Chern Han Yong, Duke-National University of Singapore Medical School
Limsoon Wong, National University of Singapore
2017

The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations
Editors: Sharon Oviatt, Incaa Designs
Björn Schuller, University of Passau and Imperial College London
Philip R. Cohen, Voicebox Technologies
Daniel Sonntag, German Research Center for Artificial Intelligence (DFKI)
Gerasimos Potamianos, University of Thessaly
Antonio Krüger, Saarland University and German Research Center for Artificial Intelligence (DFKI)
2017

Communities of Computing: Computer Science and Society in the ACM
Thomas J. Misa, Editor, University of Minnesota
2017

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai, University of Illinois at Urbana–Champaign
Sean Massung, University of Illinois at Urbana–Champaign
2016

An Architecture for Fast and General Data Processing on Large Clusters
Matei Zaharia, Stanford University
2016

Reactive Internet Programming: State Chart XML in Action
Franck Barbier, University of Pau, France
2016

Verified Functional Programming in Agda
Aaron Stump, The University of Iowa
2016
The VR Book: Human-Centered Design for Virtual Reality
Jason Jerald, NextGen Interactions
2016

Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age
Robin Hammerman, Stevens Institute of Technology
Andrew L. Russell, Stevens Institute of Technology
2016

Edmund Berkeley and the Social Responsibility of Computer Professionals
Bernadette Longo, New Jersey Institute of Technology
2015

Candidate Multilinear Maps
Sanjam Garg, University of California, Berkeley
2015

Smarter Than Their Machines: Oral Histories of Pioneers in Interactive Computing
John Cullinane, Northeastern University; Mossavar-Rahmani Center for Business and Government, John F. Kennedy School of Government, Harvard University
2015

A Framework for Scientific Discovery through Video Games
Seth Cooper, University of Washington
2014

Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers
Bryan Jeffrey Parno, Microsoft Research
2014

Embracing Interference in Wireless Systems
Shyamnath Gollakota, University of Washington
2014
The Handbook on Socially Interactive Agents

20 years of Research on
Embodied Conversational Agents,
Intelligent Virtual Agents, and Social Robotics
Volume 2: Interactivity, Platforms, Application

Birgit Lugrin, editor
Julius-Maximilians-Universität of Würzburg

Catherine Pelachaud, editor
CNRS-ISIR, Sorbonne Université

David Traum, editor
University of Southern California

ACM Books #48
Copyright © 2022 by Association for Computing Machinery

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which the Association of Computing Machinery is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

The Handbook on Socially Interactive Agents
Birgit Lugrin, Catherine Pelachaud, David Traum, Editors
books.acm.org
http://books.acm.org
ISBN: 978-1-4503-9896-1 hardcover
ISBN: 978-1-4503-9894-7 paperback
ISBN: 978-1-4503-9895-4 EPUB
ISBN: 978-1-4503-9897-8 eBook
Series ISSN: 2374-6769 print  2374-6777 electronic
DOIs:
10.1145/3563659 Book 10.1145/3563659.3563669 Chapter 21
10.1145/3563659.3563660 Foreword 10.1145/3563659.3563670 Chapter 22
10.1145/3563659.3563661 Preface 10.1145/3563659.3563671 Chapter 23
10.1145/3563659.3563662 Chapter 14 10.1145/3563659.3563672 Chapter 24
10.1145/3563659.3563663 Chapter 15 10.1145/3563659.3563673 Chapter 25
10.1145/3563659.3563664 Chapter 16 10.1145/3563659.3563674 Chapter 26
10.1145/3563659.3563665 Chapter 17 10.1145/3563659.3563675 Chapter 27
10.1145/3563659.3563666 Chapter 18 10.1145/3563659.3563676 Chapter 28
10.1145/3563659.3563667 Chapter 19 10.1145/3563659.3563677 Appendix A
10.1145/3563659.3563668 Chapter 20 10.1145/3563659.3563678 Bios/Index

A publication in the ACM Books series, #48
Editors in Chief: Sanjiva Prasad, Indian Institute of Technology (IIT) Delhi, India
Marta Kwiatkowska, University of Oxford, UK
Charu Aggarwal, IBM Corporation, USA
Area Editor: Ramesh Jain, University of California, Irvine (UCI)

This book was typeset in Arnhem Pro 10/14 and Flama using pdfltx.
Front cover design by Sophia C. Steinhaeusser and Birgit Lugrin (copyright remains with the designers).

First Edition
10 9 8 7 6 5 4 3 2 1
Contents

Foreword xvii  
Ruth Aylett  
References xx

Preface xxv  
Birgit Lugrin  
Purpose of the Book xxvii  
Structure of the Book xxviii  
Terminology xxxiii  
References xxxvi

PART IV  MODELING INTERACTIVITY  1

Chapter 14  Interaction in Social Space  3  
Hannes Högni Vilhjálmsson  
14.1 Motivation  3  
14.2 Models and Approaches  4  
14.3 History/Overview  16  
14.4 Similarities and Differences in IVAs and SRs  29  
14.5 Current Challenges  31  
14.6 Future Directions  34  
14.7 Summary  36  
References  36

Chapter 15  Socially Interactive Agent Dialogue  45  
David Traum  
Acknowledgements  65  
References  65
Chapter 16  The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures  77
Stefan Kopp and Teena Hassan
16.1 Motivation  77
16.2 Models and Approaches  80
16.3 Current Challenges and Future Directions  98
16.4 Summary  102
References  103

Chapter 17  Multiparty Interaction Between Humans and Socially Interactive Agents  113
Sarah Gillet, Marynel Vázquez, Christopher Peters, Fangkai Yang, and Iolanda Leite
17.1 Motivation  113
17.2 Models and Approaches  115
17.3 Advances in Multiparty Interaction  118
17.4 Group Datasets  130
17.5 Similarities and Differences in IVAs and SRs  132
17.6 Current Challenges and Future Directions  134
17.7 Summary  137
References  137

Chapter 18  Adaptive Artificial Personalities  155
Kathrin Janowski, Hannes Ritschel, and Elisabeth André
18.1 Motivation  155
18.2 Psychological Background  157
18.3 Implementing Adaptive Systems  163
18.4 Evaluating Adaptive Systems  171
18.5 History/Overview  173
18.6 Current Challenges  176
18.7 Conclusion  179
Acknowledgement  181
References  181

Chapter 19  Long-Term Interaction with Relational SIAs  195
Jacqueline M. Kory-Westlund, Hae Won Park, Ishaan Grover, and Cynthia Breazeal
19.1 Motivation  195
19.2 What is Considered Long-term Interaction?  197
19.3 Relationship Models and Approaches 199
19.4 Designing Relational SIAs 202
19.5 History of Long-term Interaction with SIAs 207
19.6 Similarities and Differences in Social Robots, Virtual Agents, Voice Assistants, and other SIAs 218
19.7 Trends in Long-term SIA Research over the Past 20 Years 222
19.8 Current Challenges 223
19.9 Future Directions 231
19.10 Conclusion 239
Acknowledgments 240
References 240

Chapter 20 Platforms and Tools for SIA Research and Development 261
Arno Hartholt and Sharon Mozgai
20.1 Motivation 261
20.2 Overview 262
20.3 History and Trends 262
20.4 Agent Platforms 264
20.5 Tools to Create Appearance and nonverbal Behavior 276
20.6 Tools to Model Interactions 280
20.7 Similarities and Differences in IVAs and SRs 287
20.8 Current Challenges 289
20.9 Future Directions 290
20.10 Summary 291
References 292

PART V AREAS OF APPLICATION 305

Chapter 21 Pedagogical Agents 307
H. Chad Lane and Noah L. Schroeder
21.1 Background and Motivation 307
21.2 History of Pedagogical Agents 308
21.3 Designing and Implementing Pedagogical Agents 310
21.4 Empirical Status of Pedagogical Agents 315
21.5 Relationship with IVA and SR Research 318
21.6 Current Challenges 319
21.7 Future Directions 320
21.8 Summary 321
References 322
Appendix A  Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics  561

Birgit Lugrin, Catherine Pelachaud, Elisabeth André, Ruth Aylett, Timothy Bickmore, Cynthia Breazeal, Joost Broekens, Kerstin Dautenhahn, Jonathan Gratch, Stefan Kopp, Jacqueline Nadel, Ana Paiva, and Agnieszka Wykowska

A.1 Introduction  561
A.2 Interview 1: Social Interaction  563
A.3 Interview 2: Computational Approaches  576
A.4 Interview 3: Evaluation  590
A.5 Interview 4: Ethics  600
A.6 Interview 5: Ethics in the Application of SIA for Children with Autistic Spectrum Disorders  618
A.7 Concluding Remarks  623
References  624

Authors’ Biographies  627

Index  647
Foreword

This handbook is a timely publication for a field that has grown substantially in recent years. Its roots go back to work on embodied conversational characters from the late 1990s [Cassell et al. 1998]. This drew on the earlier Natural Language Processing field of dialogue management [see, e.g., Larsson and Traum 2000] as well as the explosively expanding area of Agent Technology in the 1990s [Wooldridge and Jennings 1994].

As technology developed, making it feasible to take computer technology out of special static locations and into a variety of existing human social environments, the issue of social intelligence was necessarily raised. It had already been shown [Reeves and Nass 1996] that people attribute social personhood to computer systems they interact with. This was all the more so as this field created embodiments for such systems, whether as graphical characters (intelligent virtual agents) or as robots [Dautenhahn 1998] that drew on the human instinct to anthropomorphize.

Moreover, the addition of embodiment opened up new dimensions of interaction through non-verbal communication—covered in Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022]—that were necessarily socially located. This in turn raised questions, little considered in dialogue management or agent technology, of perceived personality, discussed in Chapter 18 on “Adaptive Artificial Personalities” [Janowski et al. 2022], as well as affective displays. Work in Cognitive Science and Artificial Intelligence (AI) was fused with work in animation and other graphical technologies, with game engines often used as a delivery mechanism as we see in Chapter 20 on “Platforms and Tools for SIA Research and Development” [Hartholt and Mozgai 2022]. Alongside intelligent virtual agents, similar ideas were developed in robotics [Breazeal 1998] with the foundation of the field of human–robot interaction in which embodiment was fundamental.
Volume 2 of this handbook offers a comprehensive coverage of the last 20 years of R&D in the field and the current state-of-the-art. It focuses on some of the theory behind social interactivity in Part IV and on the wide range of application areas that have been tackled in Part V. The discussion of social interactivity theory in Chapter 14 on “Interaction in Social Space” [Vilhjálmsson 2022] is an important one. While some architectures used to develop intelligent virtual agents have been strongly informed by cognitive models—for example, FAtiMA [Mascarenhas et al. 2022] and EMA-SmartBody [Marsella and Gratch 2009], as discussed in Volume 1, Part III—it is sometimes less clear how the social interaction theories discussed in Part IV are actually operationalized in the applications in Part V. This is an area where much remains to be done, though work in culturally specific agents, covered in Chapter 13 on “Culture for Socially Interactive Agents” [Lugrin and Rehm 2021] of Volume 1 of this handbook [Lugrin et al. 2021], is one motivation for a deeper use of social theories [Mascarenhas et al. 2016].

That said, the level of activity in applying SIA technology is substantial and growing. It is important to understand that, as with other systems using AI components, generalization across domains is not feasible in practice, especially for social robots, where real-world constraints are very tough. Trying to build a generalist system soon exposes the limitations of the current state of the art. Thus the commercially available generalist technologies of Alexa, Google Home, Siri, and others—all disembodied—work well as QA interfaces to the Internet but have so far failed to make the leap into commercial deployment of connected conversation in spite of substantial research effort driven in some cases by competitions [Ram et al. 2018].

However, in specific niches SIA technology can be a very positive factor. Two strong examples of such SIAs making real-world contributions can be seen in the Paro seal robot [Wada and Shibata 2007], used to increase interaction for elderly dementia sufferers, with clinically proven benefits, and the successful SIA that motivates patients to continue with their post-hospital health regime [Bickmore et al. 2007].

In Part V, the handbook focuses on the three areas in which most activity has taken place: assistive/health technologies, education, and computer games/interactive narrative. This is not an exclusive list since, as with all flourishing research areas, new ideas are continually being explored, while specific elements gradually make their way into marketed applications, as with online corporate chatbot help systems.

However, many, if not most, of the systems discussed represent applied research rather than immediately deployable commercial technology. One should not underestimate the time it takes to move from an initial research concept
to a real-world technology. The Paro seal robot was conceived in the 1990s [Shibata and Irie 1997] and was already being evaluated with its target population in the early 2000s. It took at least a further ten years to complete clinical trials proving its benefits so that it could be used in real-world settings with elderly dementia sufferers. A deployed product has to work over the long term and not just in a limited lab-based interaction. As we see in Chapter 19 on “Long-Term Interaction with Relational SIAs” [Kory-Westlund et al. 2022], robust long-term interaction comes with a set of specific challenges.

Work in the very promising use of social robots for support of children with autism, discussed in Chapter 25 on “Autism and Socially Interactive Agents” [Nadel et al. 2022], is an example of a technology on the path to deployment that has not yet attained a commercial product. Pioneered by a few groups, for example, that at the University of Hertfordshire in the UK with its Kaspar robot [Wood et al. 2021], and now researched in numerous centers, this has advanced to the stage of small-scale studies with the target population, showing definite benefits but has yet to arrive at the stage of a full clinical trial. The chances seem good that it will do so at some point in the near future.

While the abilities of “AI” have been hyped in the recent period, in practice fully autonomous intelligent agents still represent a source of risk for commercial providers. Natural language engineering is a bottleneck in that commercial organizations want to know what a system they deploy “will say” over a customer base that may run into millions. Tailored dialogue systems grounded in known semantics are usually used in health applications for this reason, as we see in Chapter 15 on “Socially Interactive Agent Dialogue” [Traum 2022]. Large language models, as in Google’s GPT3 and other similar systems, operating as pure pattern-matchers, can be lamentably unpredictable in their output, not to mention inappropriate or plain wrong [Dale 2021]. The issues relating to robust dialogue are one reason why intelligent virtual agent SIAs, in principle easier to deploy than robotic SIAs, still face obstacles to commercial deployment.

In education, SIAs have appeared in most cases as adjuncts to existing intelligent tutoring systems. They are therefore dependent on the capacities of those systems, which themselves are not widely deployed. Chapter 21 on “Pedagogical Agents” [Lane and Schroeder 2022] on pedagogical systems examines the specific challenges. Language learning is one educational niche in which the motivating effect of an SIA is seen as a positive factor. This is a benign domain in that it is based on short interactions, where repeated practice is one of the main pedagogical requirements, and it has long been computerized.
However, the perceived unpredictability of autonomous operation can be an issue in education too since the point of educational applications is to reach definite pedagogical goals. Educationalists feel much happier if there is hard evidence that they do so. In thinking about the limits of autonomy, it is worth pointing out that social interaction normally limits human autonomy, whether through social norms, peer-group pressure, or law, and that social intelligence involves understanding what range of actions is open to you in the current social situation.

A lower risk educational area is the one that overlaps with computer games—serious games, which is discussed in Chapter 28 on “Serious Games with SIAs” [Gebhard et al. 2022]. Serious games are usually viewed as an adjunct to other educational materials and as a motivating force rather than a core educational resource. SIAs can open up an experiential role-play-based approach that is much harder to organize live in a classroom.

The focus on socially useful and entertainment applications is one way in which some of the difficult ethical issues around anthropomorphism discussed in Volume 1 can be body-swerved. There is no question that SIAs draw on a human need to anthropomorphize that is hardwired into our social brains. When the point of the SIA is to improve your health or educate you, the potential benefits should outweigh the inherently deceptive presentation of a computer artefact as a person. Whether this is true of some of the corporate support bots, of very limited “intelligence,” social or otherwise, and apparently designed to block customers from any human support, is a different question.

If this volume of the handbook were to go to a second edition in say five year’s time, there is no doubt that new application areas and greater commercial deployment would feature. This is an exciting area in which the known issues are being attacked with verve and energy. Most researchers of SIAs do not expect the world to be full of them in the immediate future, but there is no question that the future of technology has to be social.

Ruth Aylett

References


References


Preface

Birgit Lugrin

This preface is a short version of the introduction chapter of this handbook [Lugrin 2021], which can be found in Volume 1 [Lugrin et al. 2021]. Here, we repeat the definition of Socially Interactive Agents (SIAs), state the purpose of this book, and introduce its contents and terminology. It extends the introduction chapter by adding more details on the content of Volume 2, including a new chapter that reports on interviews with leading experts on the current challenges of SIA research. For further information on the potential of or vision for SIAs, their origin, characteristics of embodiment, or suggestions on further readings, please refer to Chapter 1 on “Introduction to Socially Interactive Agents” [Lugrin 2021] of volume 1 of this handbook [Lugrin et al. 2021].

The research area of SIAs aims to develop artificial agents that can interact via communication channels that come more natural to human interactants by equipping the interface with a body that interacts multimodally by using verbal, paraverbal and nonverbal behaviors. With it, communication styles that are known from human face-to-face interaction can be transferred to the interaction with machines.

SIAs (see Figure 1 for examples) have been developed under different names in different research fields such as Intelligent Virtual Agents, Embodied Conversational Agents, or Social Robotics (see below for definitions of the respective terms). More than 20 years of research and development in these fields have drastically advanced the state of the art. For this book, we chose to use the term Socially Interactive Agents (SIAs) as it includes both physical and virtual embodiments, while highlighting their ability for social interaction, as well as the need to realize socially intelligent, autonomous behaviors.
We define SIAs as follows:

Socially Interactive Agents (SIAs) are virtually or physically embodied agents that are capable of autonomously communicating with people and each other in a socially intelligent manner using multimodal behaviors.

In order to interact with humans in a socially intelligent manner, underlying concepts such as emotions, empathy, or how to behave in a group are essential for SIAs to interpret. To be part of the interaction, observed input must be reasoned about, and decisions to be taken upon that resemble a cognitive process. The SIA’s (re)actions need to be externalized by natural language, expressive speech, and nonverbal behaviors.

Thanks to extensive research, today prototypes including SIAs are used in many application domains helpful for individuals or today’s society, with SIAs serving as companions or assistants in ageing support, health education, lifelong learning, or training of specific skills. In the long run, SIAs are envisioned to unobtrusively support humans in their daily lives.

A concern or fear that many researchers in the research area of SIAs are confronted with is the conception that these agents might be developed to replace humans in the workplace or even in social relationships. It is very important to note here that a replacement of humans is not, and has never been, a goal in the development or research on SIAs. On the contrary, SIAs are developed to support humans and assist in situations where no human support can be provided or is
Purpose of the Book

The fields of intelligent virtual agents (IVAs) and social robotics (SRs) face similar research issues and challenges and are further developed in universities and research facilities across the world. Research on IVAs and SRs can highly benefit from one another and have contributed to each other's advancement in the past. However, substantial work in both research fields is sometimes overlooked by researchers in the other area. This is partly due to the fact that different wordings are used and there exists a large number of journals and conferences that publish works on SIAs, making it very difficult to keep a good overview (see Chapter 1 on “Introduction to Socially Interactive Agents” [Lugrin 2021] of volume 1 of this handbook [Lugrin et al. 2021] for a list of journals and conferences). Please note that while we intended to involve the communities of IVAs and SRs, the content of this book, or particular chapters, are of interest for other communities as well, for example, for people working with voice assistants, conversational agents, virtual reality, game design, animation videos, assistive robots, or other types of systems containing autonomous agents.

The interdisciplinary nature of SIA research highly contributes to the very diverse venues where you can find relevant findings on SIAs. While researchers from the cognitive sciences bring expertise in underlying processes, communication, and interaction, computer scientists bring expertise in conceptualizing computational models and implementation. Even within a single discipline, approaches, methods, and wording can be used differently, complicating cooperation. In computer science, for example, many areas are involved in SIA research, such as artificial intelligence, human–computer interaction, robotics, computer graphics, or software engineering. Only through communication and research in interdisciplinary teams can the field be advanced. This constitutes one major challenge by itself as researchers sometimes do not have enough insight into other areas (or even disciplines) and thus might not appreciate each other's work enough.

We hope that this handbook will help raise the visibility of the research in the fields involved and further close the gap between the IVA and SR communities. This comprehensive handbook on SIAs summarizes the research that has taken place over the last 20 years. We are referring to this time period, since the
first complete book on Embodied Conversational Agents ([Cassell et al. 2000], see above) appeared in 2000, although we are aware that research on this topic started before. By pointing out current challenges and future directions in the various topics involved, we hope to help in directing future research and cooperation. In the book, we include views from an interdisciplinary perspective, containing theoretical backgrounds from human–human interaction, their implementation in computational models, their evaluation with human users, integration into applications, and ethical implications.

In a structured and easily accessible way, the book (hopefully) provides a valuable source of information on SIAs for research and education. Researchers in the area of SIAs will find it a valuable overview of the field. Teaching staff will benefit from the handbook to structure courses for undergraduate or graduate students, and with it train the new upcoming generation of young researchers.

Particularly now, public interest in SIAs is increasing. The book will also help professionals, and interested lay public readers, to get acquainted with the research area.

**Structure of the Book**

This handbook is split into two volumes, including 28 chapters that are grouped in five major parts, to cover the major topics in the area. For the book, we have relied on our connections to both fields, IVAs and SRs, providing a collection of surveys, each written by (an) acknowledged international expert(s) of their field.

Each chapter provides a survey that summarizes the theoretical background, approaches for implementation, history/overview of the topic, alongside with current challenges and future directions. All chapters discuss similarities and differences between IVAs and SRs and highlight important work in both fields. Where applicable, the chapters will follow a common structure to ensure internal consistency and facilitate understanding. In addition to the content of this handbook as outlined in Volume 1, we have added a challenge discussion to the end of this volume as an Appendix A on “Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics” [Lugrin et al. 2022].

**Volume 1**

After the first chapter [Lugrin 2021] that introduces readers to the handbook, Volume 1 [Lugrin et al. 2021] starts with Part I “Establishing SIA Research” that helps understand how research in this area is conducted and discusses the impact thereof on individuals and society.
Chapter 2 “Empirical Methods in the Social Science for Researching Socially Interactive Agents,” by Astrid Rosenthal-von der Pütten and Anna M. H. Abrams [Rosenthal-von der Pütten and Abrams 2021], introduces the empirical methodology from the social sciences that is necessary for SIA research, particularly when it comes to research experiments including human participants.

Chapter 3 “Social Reactions to Socially Interactive Agents and Their Ethical Implications,” by Nicole Krämer and Arne Manzeschke [Krämer and Manzeschke 2021], looks at SIA research from a psychological and ethical perspective. It points to numerous studies demonstrating that people (unconsciously) react socially towards artificial entities, and that as soon as they display social cues people can also be manipulated or socially influenced.

Part II “Appearance and Behavior,” deals with the impact of the looks of SIAs and the various aspects of multimodal behavior that need to be taken into account when convincing SIAs behavior is modeled.

Chapter 4 “Appearance,” by Rachel McDonnell and Bilge Mutlu [McDonnell and Mutlu 2021], argues that compared to voice assistants embodied agents enable the use of appearance-based cues from human–human interaction, such as mutual gaze, that are known to improve social outcomes. The chapter shows that the appearance of an SIA can affect how people perceive, respond to, and interact with it.

Chapter 5 “Natural Language Understanding in Socially Interactive Agents,” by Roberto Pieraccini [Pieraccini 2021], introduces natural language understanding as an essential part of any interactive agent and highlights its complexity, particularly for SIAs that need to react to user-initiated interactions across various application areas.

Chapter 6 “Building and Designing Expressive Speech Synthesis,” by Matthew Aylett, Leigh Clark, Benjamin R. Cowan, and Ilaria Torre [Aylett et al. 2021], gives an overview of definitions, methods, and state-of-the art in expressive voices, and critically discusses when and where expressive speech is beneficial.

Chapter 7 “Gesture Generation,” by Carolyn Saund and Stacy Marsella [Saund and Marsella 2021], discusses the complexity of communicative gestures and how they enhance communication in human–human conversation, and summarizes the research and their challenges in the transfer of this complexity in the implementation with SIAs.

Chapter 8 “Multimodal Behavior Modeling for Socially Interactive Agents,” by Catherine Pelachaud, Carlos Busso, and Dirk Heylen [Pelachaud et al. 2021], extends the theme nonverbal behavior by adding additional modalities such as gaze, smiles, or social touch. Starting from introducing concepts from the social
Preface

sciences, the chapter has a strong focus on the different computational models that can be employed for the implementation of multimodal behaviors.

Part III “Social Cognition—Models and Phenomena” investigates internal processes known from human cognition that are driving forces in human–human interaction and demonstrates how they are addressed in SIA systems.

Chapter 9 “Theory of Mind and Joint Attention,” by Jairo Perez-Osorio, Eva Wiese, and Agnieszka Wykowska [Perez-Osorio et al. 2021], introduces the two crucial mechanisms of social cognition and explains how they apply to the interaction between humans and SIAs from two angles: evoking human social cognition and modeling artificial social cognition.

Chapter 10 “Emotion,” by Joost Broekens [Broekens 2021], focuses on the computational representation of emotion and other related affective concepts such as mood, attitude, or appraisal and highlights how SIAs can make constructive use of them.

Chapter 11 “Empathy and Prosociality in Social Agents,” by Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patricia Arriaga [Paiva et al. 2021], focuses on empathy and in particular on the related concept of prosociality (conducting positive and voluntary behavior that should benefit others). With it, the authors provide a framework including the main variables needed to design prosocial agents, for individual or dyadic interactions, or at the society level.

Chapter 12 “Rapport Between Humans and Socially Interactive Agents,” by Jonathan Gratch and Gale Lucas [Gratch and Lucas 2021], introduces rapport (a fine-grained emotional communicational interplay) in the communication of humans and machines by approaching it from a theoretical, computational, and empirical side, and demonstrating its benefits.

Chapter 13 “Culture for Socially Interactive Agents,” by Birgit Lugrin and Matthias Rehm [Lugrin and Rehm 2021], argues that implementing culture for SIAs can be beneficial not only to raise their acceptance in certain user groups but also to be able to teach about cultural differences and foster cultural diversity.

Volume 2

This second volume of the handbook starts with this preface, which recaps the most important aspects and terminology of the introduction chapter. The chapters following in Parts IV (modeling interactivity) and V (areas of application) rely on the first three parts of Volume 1 of the book by applying the outlined research methods, using knowledge on appearance, verbal, para-verbal, and nonverbal behavior, as well as relying on underlying cognitive phenomena. Therefore, some concepts, topics, or concrete applications might be present in more than one chapter, while being discussed from different research angles and from multiple points of view.
Part IV “Modeling Interactivity” explains how interaction with human users or other SIAs is modeled and how the many detailed aspects of multimodal, multiparty, adaptive interactivity are implemented.

Chapter 14 “Interaction in Social Space,” by Hannes Högni Vilhjálmssson [Vilhjálmssson 2022], deals with the intricate social performance that inevitably takes place when SIAs and human users share the same social space (virtual or physical), regardless of their explicit intentions to connect with one another.

Chapter 15 “Socially Interactive Agent Dialogue,” by David Traum [Traum 2022], introduces several approaches to modeling the structure of extended verbal and multimodal interactions, with an emphasis on how different kinds of embodiment impact the communication affordances and requirements for SIA tasks.

Chapter 16 “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures,” by Stefan Kopp and Teena Hassan [Kopp and Hassan 2022], presents different SIA architectures and gives an extensive overview on how SIAs can engage in dynamic and fluid social interaction, discussing different approaches to deal with multimodality and interactivity.

Chapter 17 “Multiparty Interaction Between Humans and Socially Interactive Agents,” by Sarah Gillet, Marynel Vázquez, Christopher Peters, Fangkai Yang, and Iolanda Leite [Gillet et al. 2022], looks into SIAs that interact with a group of people for which the complex group dynamics need to be understood, and highlights that the SIA can affect and even explicitly influence the group's dynamics.

Chapter 18 “Adaptive Artificial Personalities,” by Kathrin Janowski, Hannes Ritschel, and Elisabeth André [Janowski et al. 2022], focuses on how a SIA can automatically adapt its personality in accordance with the user’s preferences, and with it make the interaction with them more enjoyable and productive.

Chapter 19 “Long-Term Interaction with Relational SIAs,” by Jacqueline M. Kory-Westlund, Hae Won Park, Ishaan Grover, and Cynthia Breazeal [Kory-Westlund et al. 2022], argues that strong relationships support people to achieve their goals in various domains and that thus relational SIAs have the potential to scaffold humans in their long-term endeavors.

Chapter 20 “Platforms and Tools for SIA Research and Development,” by Arno Hartholt and Sharon Mozgai [Hartholt and Mozgai 2022], gives a practical introduction to the history of SIA platforms and tools, directing to state-of-the-art technical solutions that support the development and implementation of SIAs.

Part V “Areas of Application” gives an overview of the major domains in which SIAs are employed, highlighting systems and research findings and the benefits of SIAs to individuals and society.
Chapter 21 “Pedagogical Agents,” by H. Chad Lane and Noah L. Schroeder [Lane and Schroeder 2022], introduces work with SIAs in the domain of education, examining social aspects of teaching and learning and summarizing empirical research with pedagogical agents.

Chapter 22 “Socially Interactive Agents as Peers,” by Justine Cassell [Cassell 2022], describes work that uses SIAs that are designed to work or play with children or teenagers at an eye-level, discussing the benefits of SIAs that look and act like peers rather than teachers, tutors, or parents.

Chapter 23 “Socially Interactive Agents for Supporting Aging,” by Moojan Ghafurian, John Muñoz, Jennifer Boger, Jesse Hoey, and Kerstin Dautenhahn [Ghafurian et al. 2022], is centered on work with SIAs located in the area of aging support that aim to improve older adults’ quality of life and wellbeing. The chapter provides methods and suggestions to address the many challenges that arise when designing SIAs that should successfully assist the targeted user group.

Chapter 24 “Health-Related Applications of Socially Interactive Agents,” by Timothy Bickmore [Bickmore 2022], addresses another area of major societal importance and highlights the potential of SIAs that have shown to have a positive impact on voluntary changes in health behavior.

Chapter 25 “Autism and Socially Interactive Agents,” by Jacqueline Nadel, Ouriel Grynszpan, and Jean-Claude Martin [Nadel et al. 2022], reviews work that uses SIAs to study or help improve the social skills of people with autism spectrum disorder. The chapter highlights the improvements that have been achieved throughout the last two decades and that, following a multidisciplinary approach, more can be expected in the future.

Chapter 26 “Interactive Narrative and Story-telling,” by Ruth Aylett [Aylett 2022], introduces narrative and storytelling as fundamental human capabilities and outlines how SIAs are used in character- or plot-based systems, highlighting the great challenge of interactivity in this domain.

Chapter 27 “Socially Interactive Agents in Games,” by Rui Prada and Diogo Rato [Prada and Rato 2022], discusses the complexity in which SIAs have been used in games and introduces their different roles alongside with their contributions to gameplay.

Chapter 28 “Serious Games with SIAs,” by Patrick Gebhard, Dimitra Tzovaltzi, Tanja Schneeberger, and Fabrizio Nunnari [Gebhard et al. 2022], focuses on serious games that can partly be seen as a means to an end to achieve certain goals in various domains (such as education or health-behavior change) using specific methods from games and interactive narratives. Thus, the chapter focuses on learning gain as well as individual experience during game play.
Appendix A holds a “Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics,” organized and edited by Birgit Lugrin and Catherine Pelachaud [Lugrin et al. 2022], that contains interviews with international experts of their fields who are all authors of this book, discussing the major challenges SIA research is facing today, pointing out potential future directions, and inviting other communities to join our journey. Interviewees were (in alphabetical order) Elisabeth André, Ruth Aylett, Timothy Bickmore, Cynthia Breazeal, Joost Broekens, Kerstin Dautenhahn, Jonathan Gratch, Stefan Kopp, Jacqueline Nadel, Ana Paiva, and Agnieszka Wykowska.

Authors’ Biographies holds short CVs of all authors and editors of this second volume of the handbook.

Terminology
Since research on SIAs is manifold and researchers are coming from different disciplines and research areas, a number of terms exist that can be found in the literature. In the following, we aim to shed light on the terminology (in alphabetical order) and highlight their origin and different foci, albeit you might find some of the definitions being quite similar:

Agent “An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators” [Russell and Norvig 2009]. This very classic and well-known definition looks at agents from the perspective of artificial intelligence (AI), highlighting the autonomy of the artificial entities. Those agents can be, but are not necessarily, embodied. Examples include softbots, thermostats, robots, or humans.

Avatar An avatar represents a game unit that is under the player’s control [Kromand 2007], which is usually the graphical representation of the user in the virtual environment [Trepte and Reinecke 2010]. Unfortunately, this term is often confused with virtual or robotic agents in communities other than SIAs. Note that an avatar is not behaving or interacting autonomously with a user but representing the user in the virtual or real world. The term embodiment also has a different meaning concerning avatars and describes the physical process to substitute (parts of) a person’s body with a virtual one by the deployment of virtual reality (VR) hard- and software [Spanlang et al. 2014].

Embodied Conversational Agent “Embodied conversational agents are computer-generated cartoonlike characters that demonstrate many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication” [Cassell et al. 2000]. The term was defined
by Cassell and colleagues in their same named book on the topic in 2000. The authors highlight the importance of the combination of the multimodal interface, with a software agent and a dialogue system, to assure natural conversation. While the original focus was on virtual embodiments, the term also allows robotic embodiments and is used in both fields.

**Intelligent Virtual Agent** “Intelligent virtual agents are interactive digital characters that exhibit human-like qualities and can communicate with humans and each other using natural human modalities like facial expressions, speech and gesture. They are capable of real-time perception, cognition, emotion and action that allow them to participate in dynamic social environments” [IVA 2019]. This term focuses on communicative, digital characters and is mainly used by researchers who are affiliated with the IVA conference series. An important fact lies on the character’s intelligence that allows them to dynamically interact, as opposed to scripted behavior.

**Socially Assistive Robot** Socially Assistive Robots were defined in Feil-Seifer and Matarić [2005] as robots that share characteristics with assistive robots, in particular to provide assistance to users, but distinguished by their focus on social interaction while assisting people.

**Socially Intelligent Agent** “The field of socially intelligent agents is characterized by agent systems that show human style social intelligence” [Dautenhahn et al. 2002]. The term was coined by Dautenhahn in the late 1990s and highlights the specific social intelligence of the agent, relying on “deep models of human cognition and social competence” [Dautenhahn 1998] that need to be a part of strongly interdisciplinary approaches. Different embodiments of these agents are possible, virtual or robotic.

**Socially Interactive Robot** Socially interactive robots were defined as “robots for which social interaction plays a key role” [Fong et al. 2003] in order to “distinguish these robots from other robots that involve ‘conventional’ human–robot interaction, such as those used in teleoperation scenarios” [Fong et al. 2003]. This term was defined after the definition of socially intelligent agents to highlight the need for social interaction.

**Socially Interactive Agent** The term socially interactive agents extends the term socially interactive robot by allowing virtual and physical embodiments. This term was used by the AAMAS (autonomous agents and multiagent systems) community and conference series, where they are described as “capable of interacting with people and each other using social communicative behaviors common to human–human interaction. Example applications include social assistants on
mobile devices, pedagogical agents in tutoring systems, characters in interactive
games, social robots collaborating with humans and multimodal interface agents
for smart appliances and environments” [AAMAS 2019].

Social Robot “Social (or Sociable) robots are designed to interact with people in
a natural, interpersonal manner [...] They will need to be able to communicate
naturally with people using both verbal and nonverbal signals. They will need to
engage us not only on a cognitive level, but on an emotional level as well in order
to provide effective social and task-related support to people” [Breazeal et al. 2016].
Social robotics is distinguished from robotics through its socially interactive focus
with applications in domains such as education, ageing support, or entertainment.
This term is dominantly used by the social robotics community and the same
named conference series and journal.

Virtual Character The term virtual character focuses on a virtual representation
of a figure along with its animations. “Virtual characters in animated movies and
games can be very expressive and have the ability to convey complex emotions”
[McDonnell et al. 2008]. Note, that they do not necessarily have to be intelligent
or interactive, for example, the characters of a movie. Thus, the term is often used
by researchers who focus on the character's appearance, graphics, animation, or
background story.

Virtual Human “Virtual Humans are artificial characters who look and act like
humans but inhabit a simulated environment” [Traum 2008]. The term focuses
on the human-like appearance and behavior and is frequently used by US authors
and research groups. Research on virtual humans often relies on highly realistic
graphical representations of the characters and their animations.

Please note that the terms introduced above are the ones most commonly used.
Other variations, for example, affective embodied agent, companion robot, conver-
sational robot, relational agent, social embodied agent, socially intelligent robot,
socially intelligent virtual agent, virtual agent, and so on, are also found in the
literature and address similar research topics.

For the scope of this book, we use the term Socially Interactive Agents (SIAs)
when we talk about both kinds of embodiment, virtual or robotic. We chose this
term as we think it highlights the socially interactive nature as well as the intelli-
gent background of the agent. When discussing only one specific type of embod-
iment, we will occasionally use other terms: Intelligent Virtual Agent (IVA) for
virtual representations and Social Robot (SR) for robotic representations.
References


References


This handbook consists of 28 chapters, covering the major topics in Socially Interactive Agents (SIAs). Each chapter provides a survey that summarizes the theoretical background, approaches for implementation, history/overview of the topic, alongside with current challenges and future directions. All chapters discuss similarities and differences between Intelligent Virtual Agents (IVAs) and Social Robots (SRs) and highlight important work of both fields in the topic.

The chapters are clustered into five parts, representing broad themes in SIA research. Volume 1 [Lugrin et al. 2021] contains Parts I–III. Part I “Establishing SIA Research” helps the reader understand how research in this area is conducted and discusses the impact on individuals and society. Part II “Appearance and Behavior” deals with the most immediately noticeable external features of SIAs across multiple modalities. Part III “Social Cognition—Models and Phenomena” investigates internal processes known from human cognition that are driving forces in human–human interaction and demonstrates how they are addressed in SIA systems.

This Volume 2 contains Parts IV “Modeling Interactivity” and V “Areas of Application.” Part IV shifts the focus away from aspects of the SIA itself or how it models and emulates human cognition and on to the interaction of a dyad or larger group consisting of SIAs and humans. Individual chapters focus on spatial interactions, dialogue, and adaptive personalities, as well as cross-cutting issues such as multi-party interaction, long-term interaction, and tools for creating SIAs. The chapters give an overview of existing approaches to model how SIAs can evolve in social space and communicate considering other SIAs as well as human users. These
chapters highlight that multiparty interaction cannot be simply modeled as a set of dyads and that modeling long-term interaction requires considering complex phenomena.

The chapters in Part IV build upon the concepts presented in Parts I–III (Volume 1 of this handbook) by applying the outlined research methods, and considering individual and ethical implications (Part I), and using knowledge on designing appearance, modeling multimodal behavior including verbal, para-verbal and non-verbal behavior (Part II), as well as relying on underlying cognitive phenomena such as theory of mind, empathy, rapport, or culture (Part III). Thus, some concepts, topics, and concrete applications that appeared in earlier chapters re-emerge in the chapters in this part but from different points of view.
Interaction in Social Space

Hannes Högni Vilhjálmsson

14.1 Motivation

Most important of all, space is one of the basic, underlying organizational systems for all living things—particularly for people. [Hall 1966, xii]

A person occupies space and is surrounded by space, which may be occupied by other people. Our mere presence in the social space of others makes us participants in a social performance. Our actions are observed, interpreted, and reacted to. Some of those actions may lead to further interaction, while others may not. Through careful coordination of actions and reactions, we are capable of negotiating the terms of our co-presence without even uttering a single word.

Humans are capable of managing their co-presence without much effort, their bodies giving off cues of position, orientation, and posture as they traverse social spaces. There are situations though that require more effort and strategic thinking, such as when an easily recognizable person of fame or notoriety needs to cross a public hall without causing commotion. For some, social spaces may even induce forms of social anxiety. Regardless of the ease or difficulty in dealing with social space, its existence is of paramount importance to human life as it envelopes face-to-face interaction with family, friends, colleagues, teachers, service personnel, and, in fact, all those who simply happen to live and work around us.

To interact with humans, SIAs need to occupy and manage these spaces as well. However, coming up with a general computational solution for a dynamic social space is not trivial, and that is perhaps one reason designers of SIA systems often choose to constrain the space heavily, for example, by planting the SIA into a fixed formation with a human user who is ready to interact.
Chapter 14  Interaction in Social Space

In order to set the SIAs free, and let them roam physical and virtual spaces where they can choose to interact or not to interact with those who share the space with them, we need to understand and appreciate the structure of that space and how that knowledge can be exploited to successfully manage human co-presence.

14.2 Models and Approaches

First, we look at how to describe social space and the social elements it contains. We then examine how the body fits into that space and by what means it can interface with it. This will give us the necessary vocabulary to review the different social functions that people carry out with spatial behavior to achieve their social goals.

The theoretical models in this section represent the works of some of the pioneers of human public behavior studies, including Albert E. Scheflen, Adam Kendon, Erving Goffman, and Edward Hall. Their models are frequently cited by those dealing with social space, but here their key concepts are organized and presented holistically to form a single framework. Instead of citing them at every turn in the text, tables are provided at the end of each section, summarizing the main theoretical concepts reviewed and from what source they were brought (see Tables 14.1–14.3). The accompanying figures in this section were manually created in 3D character modeling and animation tools1 and are used to illustrate the concepts discussed. The figures therefore do not depict virtual agents being simulated in real time.

14.2.1 The Structure of Social Space

The term social space refers to any spatial environment where people have access to other people through their embodied presence. They have an opportunity to perceive each other using multiple senses and affect each other over nonverbal and verbal channels. The space could be anything from a city square to a classroom. It is not given that all occupants of a social space are socially involved or engaged with one another, but they all need to exhibit minimal social awareness and manage their co-presence.

Within the social space (see Figure 14.1), there may be gatherings, where two or more people are in each other’s immediate presence, essentially forming collections of people of various sizes and shapes. A social situation is a social space or a section of it, where upon entering it a person immediately participates in a gathering. Naming a social occasion can explain the reason for the social situation, such as a staff meeting or a funeral. Gatherings can contain multiple social engagements,

1. The tools used were Character Creator® and iClone® from Reallusion®.
A relatively large indoor social space that includes two gatherings, A and B. Gathering A contains a single sitting formation, within which A.b and A.c form an element. Gathering B is a particular social situation, where the occasion is a staff meeting. The people around the table belong to formation B.1, inside of which B.d and B.e form an element. Person B.a becomes a part of gathering B simply by walking through the door, due to the situation, even though she does not belong to a formation yet. In fact, she may just be passing by, in which case she may temporarily hold the role of a bystander. Each person occupies a location, and the empty chairs denote unassigned locations.

organized into formations as described below, but also individuals that are not fully engaged, for example bystanders.

While the term gathering describes a relatively loose collection of people, formations are tighter clusters where more coordinated activities occur. Formations are important because their structure represents a pattern of social convention that organizes and facilitates interaction. A formation is made up of individual locations that each can comfortably hold an individual body and provide room for the individual’s actions. Typically, a location is larger than the size of the body, often extending half of the person’s width in each direction (see Section 14.2.2). Locations can be pre-allocated, such as with furniture, and are claimed by participants during a social event. The simplest formation is the dyad, which is created by pairs of people who are committed to one another and cluster closely. Another fairly simple formation arises when people arrange themselves into an array and take a common orientation, such as when sitting together on a sofa to watch TV or standing side-by-side to watch an event. Such a formation has been called an element.
Chapter 14  Interaction in Social Space

The space occupied by and surrounding a face-formation, or F-formation, has been characterized as consisting of three concentric zones. The o-zone is the unoccupied space in the center, the p-zone is occupied by the participants and the r-zone surrounds the formation.

A more complex formation is the so-called face-formation or F-formation, which occurs when people take locations facing one another (see Figure 14.2). The shape of such a formation is influenced by the number of participants, typically growing from triangles for three people, to squares for four, and to expanding circles for more people. These often become unstable around 10–12 people, at which point they may break into smaller clusters.

The distance between participants, and thus the size of the face-formation, can vary according to many things like how involved or committed the participants are with each other, how noisy or crowded the environment is, or how furniture is arranged.

The space occupied by and surrounding a formation has been characterized as consisting of different zones or regions, each with a particular role. The innermost zone is the space formed by the overlapping orientations of participants, essentially the center. This space has been termed the o-zone. Surrounding that space, the participants themselves occupy a zone that is large enough to comfortably hold their bodies and accommodate the required distance between them. This has been
termed the *p-zone* (participant zone). Further out, behind the participants, lies a region of space approximately double the width of a person. This region has been termed the *r-zone* and serves as a sort of a transition zone between the formation and the rest of the social space.

Formations may need to accommodate the movement patterns of other people in the space by leaving certain passageways open for others to pass through. These are termed *intersected formations*. Sometimes intersections are formed by physical passageways and barriers, such as isles through auditoriums or room dividers. While certain types of formations hold when people are in motion, zones are not preserved.

Physical structures can directly influence gatherings and how they break into formations, and sometimes they are deliberately set up for facilitating such social occasions. The divider and furnishing in Figure 14.1 facilitate two simultaneous gatherings and certain formations within them. A seat is a location for seating that can then be arranged into larger modules that together circumscribe an *o-zone* for face-formations (such as the chairs in Figure 14.1), or they can arrange into elements of common orientation facing a target, such as a TV (such as the couch in Figure 14.1), stage, or a theatre screen in an auditorium. Furniture can also provide work surfaces that serve as focus points for participants during interaction.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>social space</td>
<td>Any space with mutual embodied access</td>
<td>EG</td>
</tr>
<tr>
<td>gathering</td>
<td>Collection of two or more people</td>
<td>EG</td>
</tr>
<tr>
<td>social situation</td>
<td>Space that adds people immediately to a gathering</td>
<td>EG</td>
</tr>
<tr>
<td>social occasion</td>
<td>Reason for a social situation</td>
<td>EG</td>
</tr>
<tr>
<td>formation</td>
<td>Tight cluster of coordinated activity within a gathering</td>
<td>AS, AK</td>
</tr>
<tr>
<td>dyad</td>
<td>Formation of two people</td>
<td>AS</td>
</tr>
<tr>
<td>element</td>
<td>Formation of people with common orientation</td>
<td>AS</td>
</tr>
<tr>
<td>F-formation</td>
<td>Formation of people facing each other</td>
<td>AS, AK</td>
</tr>
<tr>
<td>bystander</td>
<td>Participant in gathering, but not in formation</td>
<td>EG</td>
</tr>
<tr>
<td>formation zones</td>
<td>Space occupied by and surrounding formations</td>
<td>AS, AK</td>
</tr>
<tr>
<td>o-zone</td>
<td>Innermost zone covered by overlapping orientations</td>
<td>AS, AK</td>
</tr>
<tr>
<td>p-zone</td>
<td>Zone occupied by formation participants</td>
<td>AS, AK</td>
</tr>
<tr>
<td>r-zone</td>
<td>Outermost zone serving as formation transition area</td>
<td>AS, AK</td>
</tr>
<tr>
<td>intersection</td>
<td>Passageway or structure splitting formations</td>
<td>AS</td>
</tr>
</tbody>
</table>

Table 14.1 Some useful concepts that describe the structure of social space and their sources (EG=[Goffman 1963]; AS=[Scheflen 1976]; AK=[Kendon 1990])
14.2.2 The Body in Social Space

By being present in a social space, a person will occupy a certain physical volume. While there is a large variation in body shapes, an approximation to the space occupied by a single person will suffice for most models of social space. One useful approximation of this volume is four cubic cubits, as shown in Figure 14.3. The cubit measure was originally defined as the distance from a full-grown person’s elbow to the end of their middle finger. The best way to understand this distance is to bring both elbows to your sides and bend them to 90°, sticking your forearms out

Figure 14.3 A person occupies a space that is roughly four cubic cubits, dividing the body into four regions. To be able to comfortably act, a person will need an area that is roughly 2×2 square cubits, which corresponds to what is called a location. A cubit unit is approximately 0.5m long.

---

in front, which then creates a perfect square cubit. While exact cubit lengths have varied slightly through the ages, 20in, or roughly 0.5m, is a good round approximation. Four cubic cubits divide the body into four body regions, which correspond to the feet and lower part of the legs, the upper part of the legs and pelvis, the torso, and finally the head (see Figure 14.3). The exact arrangement of the cubic cubits can vary a bit, for example, between a person standing and sitting down.

Naturally, if the four cubic cubits were the only space a person could peruse in a social gathering, they would feel quite constrained and be unable to perform most actions. By doubling the footprint of this space, however, we get something that corresponds to a person’s location, which is a unit or cell that allows us to place a person comfortably in a social environment (see Section 14.2.1 and Figure 14.3).

While the physical body is bound to its exact location, the four different body regions also claim space beyond that location through independent outward orientation. That is, by orienting the head, torso, pelvis, and legs/feet, four different segments of space further away can potentially be commanded socially by the person (see Figure 14.4). Rather than commanding all four potential segments, it is

![Figure 14.4](image-url)

**Figure 14.4** Orienting different body regions can claim different spaces, or segments, around a person. By combining the orientations of the upper two regions (head and torso), a single upper segment can be commanded, and likewise, by combining the orientations of the lower two regions (legs and pelvis), a lower segment can be claimed. This makes it possible for instance to engage in a side involvement while still committing to the original dyadic formation.
more common for people to claim only two segments by combining the orientations of the two upper body regions (claiming an upper segment) and the orientations of the two lower body regions (claiming a lower segment). A segment that coincides with an activity sustained by the person, such as watching TV or interacting with someone, is termed a *transactional segment*. This segment is usually respected by others, and typically will not get needlessly invaded. Taken together, the space occupied by the person and their claimed segments is termed their *territory*. The size of a claimed segment will vary by context, but the extent of personal influence is naturally constrained by sensing distances, which serve as a foundation for the interpersonal distances described by the *proxemic* classification system.[Hall 1966]

Originally conceived as part of the study of animal behavior and how animals react to the proximity of others, the classification of distances from a human body has resulted in four useful distances: *Intimate, personal, social, and public* (see Figure 14.5). The exact distances are not fixed numbers, but they roughly define spaces around a person that correspond to different interaction opportunities. These are largely influenced, on the one hand, by relatively universal physical

![Figure 14.5](image)

*Figure 14.5* Proxemics describe opportunities for interaction at different interpersonal distances based on human sensory capabilities as well as cultural norms. Four ranges are identified by *Hall* [1966]: *Intimate, personal, social, and public*. As a person approaches, the exact interaction configuration is negotiated through a salutation sequence. During the approach, it is important to break mutual gaze in order to minimize perceived threat.
limitations of the human species such as visual acuity and hearing, but on the other hand, they are influenced by contextual variables such as demographic and cultural norms.

The distances summarized below represent the results of Hall’s study performed on middle-class healthy adults from the northeastern seaboard of the United States [Hall 1966]. Hall cautions that these should not be taken as a representation of human behavior in general [Hall 1966], but it is worth noting the reference to general human capacity for perceiving and acting at different distances.

*Intimate distance* is closer than one cubit, or 0.5m, away (see Figure 14.5). One could therefore say that two people at an intimate distance are more or less sharing the same location (see above). At this distance people can easily touch and embrace each other, and whispers can be heard. They have a good, detailed view of each other’s face, but the rest of the body is not readily seen and is visually distorted. *Personal distance* corresponds to the comfortable distance one typically maintains from others, which ranges from 0.5m to 1.5m. By placing people at separate locations, for example, with chairs, this distance is ensured. People can easily hear speech at moderate voice levels, and they can perceive each other’s full bodies without visual distortion. It is possible to reach and grasp at the closer end of the range, but slightly further out people can barely touch. *Social distance* is kept with those one intends to conduct relatively impersonal interactions with, ranging from about 1.5m to 3.5m. At this distance touch is not possible without effort and high visual detail in faces are not easily perceived, but normal voices can be heard. At the far end of the range, it may become harder to maintain visual contact and disengagements become less of an effort. *Public distance* is anything further than 3.5m and would be considered a relatively safe distance from anyone, for example, a person you would want to avoid. At this distance only exaggerated or amplified voices can be heard, and gestures become more prominent than facial expressions.

Bringing social influence beyond the closer distances, people can engage in pointing, also called a deictic, a reference, or simply just a point. It doesn’t have to be a classic pointing hand gesture, with an extended index finger, it could be something more subtle like a brief flick of the hand toward a distant target. It doesn’t even have to be a hand gesture at all, it could also be accomplished with gaze, head movement, or even with a foot. This ability to project a clear, yet invisible, ray into the entire social space can serve many important functions (see Section 14.2.3) and is just another example of how people apply their bodies’ flexible articulation to exploit the space around them.
### Table 14.2
Some useful concepts that describe the spatial structure of bodies and their sources
(AS=[Scheflen 1975, 1976]; EH=[Hall 1966])

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>cubit</td>
<td>Elbow to fingertip, approximately 0.5m (20in)</td>
<td>AS</td>
</tr>
<tr>
<td>region</td>
<td>One of four cubic cubit sized portions of the body</td>
<td>AS</td>
</tr>
<tr>
<td>location</td>
<td>Area needed by person for action (approx. 2×2 cubits)</td>
<td>AS</td>
</tr>
<tr>
<td>segment</td>
<td>Space further out claimed by orientation of a body region</td>
<td>AS</td>
</tr>
<tr>
<td>territory</td>
<td>The space both occupied and claimed by a person</td>
<td>AS</td>
</tr>
<tr>
<td>intimate distance</td>
<td>Closer than one cubit away, sharing location</td>
<td>EH</td>
</tr>
<tr>
<td>personal distance</td>
<td>Comfortable distance, between 1 to 3 cubits away</td>
<td>EH</td>
</tr>
<tr>
<td>social distance</td>
<td>Impersonal interaction distance, between 3 and 7 cubits away</td>
<td>EH</td>
</tr>
<tr>
<td>public distance</td>
<td>Safe distance from anyone, over 7 cubits away</td>
<td>EH</td>
</tr>
<tr>
<td>point</td>
<td>Extending or orienting a body part to call out a target</td>
<td>AS</td>
</tr>
</tbody>
</table>

### 14.2.3 Interaction Functions in Social Space

Now that we have an idea of how social space is structured, and how the human body occupies that space, we turn to the interaction that occurs within that space, essentially how people accomplish their social goals.

There is a fundamental principle that underlies this discussion, and that is that *people generally adhere to social norms in social spaces*. People strive to “fit in,” that is, they do not want to attract undue attention. This requires a balance, where people do not want to thrust themselves on others but they simultaneously do not withdraw themselves from their presence. The interaction functions observed in social spaces emerge from this closed natural system of social order [Goffman 1963]. These functions can be classified as belonging either to unfocused interaction or focused interaction. We look at each in turn.

**Unfocused Interaction** Unfocused interaction is what takes place when people are merely managing their co-presence in the social space without meaningful mutual engagement, such as conversations. An example would be a person walking through a crowded mall and casually browsing for good deals at one of the stores. At a very basic level, the person needs to navigate the space without causing others distress by physically colliding with them. Where possible, the person should try to stay at least a personal distance away from others, but this can be hard in cramped spaces such as elevators. Navigation will often follow certain patterns to make the job less cognitively demanding, for example, pedestrians tend to route along paths that are already established, such as by keeping to the same side of the walkway or simply by following the people ahead of them, sometimes forming
streams or lanes where people string along, one following another. Where bottle-necks occur, people resort to common routing techniques such as step-and-slide, where they momentarily turn sideways while stepping past an oncoming person.

But being present in a social space is about much more than establishing a collision-free path. Another very important skill that people need to rely on is to signal an expected level of social awareness and at least a minimal availability to the social setting. If these signals are not sent, the person is likely to stand out in some way, for example, as being hostile or judgmental, as being ill or at best lost in their own thoughts (see Figure 14.6(a)). That kind of attention one should give to others by briefly looking at them, as they pass by, has been termed civil inattention. This is just enough eye contact to properly perceive the people around you and acknowledging their presence, but not long enough to constitute an invitation for a longer engagement (see Figure 14.6(b)). Typically, mutual gaze that lasts for 2s or less is not interpreted as a meaningful attempt to engage [Goffman 1963].

There are times when people care to only maintain unfocused interaction, thus doing what they can to avoid being brought into greater social involvement. To this end they carry out different strategies. Perhaps the most effective strategy is to remove the opportunity for mutual observation altogether, through what has been termed an involvement shield. Such shields can be anything from complete shelters from the rest of the gathering, such as adjacent bathrooms, to covering one’s face with a newspaper. Today, cell phones provide convenient involvement shields as they are readily available in most people’s pockets and offer a legitimate call for diverted attention. It is important to understand that these shields still need to be

![Figure 14.6](image-url)  
(a) Ignoring, (b) Attending, (c) Inviting.
soccer acceptable by the gathering or a person might risk getting called out for
disrupting it or otherwise displaying contempt for it.

Instead of resorting to physical barriers or objects, one can also attempt to
maintain a public distance from other people, where possible, and stay out of the
r-zones of group formations. Where closer distances are required, full or partial
orientation away from others, for example, by taking an outward facing seat, can
help cut lines of mutual perception. Gaze can also be cast down on the ground, for
example, in an elevator or while being forced to pass within someone’s personal or
social distance.

The point at which two people mutually react to one another, and possibly rec-
ognize each other (e.g., as acquaintances or friends), is a potential first step in
a sequence of behaviors that often precede further engagement. If at this point
mutual gaze is held for longer than would pass for civil inattention, moving to the
next step in the so-called salutation sequence is very likely to occur. That step would
be the distance salutation, where the typical behavior involves tossing the head
upward, with the chin pointing toward the other person, while raising eyebrows
and smiling (see Figure 14.5). This could also involve waving (see Figure 14.5). If
this salutation is mutual, and no avoidance takes place, the two parties will likely
approach each other, with the intent to start further interaction (see Figure 14.5).
During the approach, gaze is broken, as it would be considered too threatening or
intense to maintain constant gaze throughout. When the people get into either the
social zone or personal zone (often depending on the level of formality), they per-
form a close salutation, which grants them an opportunity to negotiate the spatial
arrangement of their subsequent interaction (e.g., keeping them at arm’s length in
the social zone with a handshake or bringing them into the personal zone with a
hug). The interaction has now become a focused interaction.

Focused Interaction Focused interaction is what happens when people commit to
some form of an extended social engagement or involvement such as a conversa-
tion or watching something together. Other chapters in this book dwell on what
happens during focused interaction, especially around conversations. Therefore,
this section will only briefly mention a few examples where spatial behavior helps
in managing co-presence during focused interaction.

Clear visual reference through pointing behavior, as we have seen in other chap-
ters in this book, can provide useful information during conversation, but it can
also serve as a way to connect individuals who belong to the same gathering, with-
out being directly linked to the ongoing conversation. Either the pointing behav-
ior binds the people together because they are pointing toward the same target,
for example, people gathering to watch something unusual happening across the
street, or they could point toward each other to indicate or strengthen relations. That could happen when recognizing a friend in a crowd or when talking to several people but wanting to demonstrate greater commitment to one of them (see below). The ability to execute pointing of various types and with various targets, both near and far, can therefore carry an important relational function.

The usefulness of dividing the body into separate orientation regions becomes clear when we consider multiple people orienting toward a common point of interest. How the rest of the body is oriented will signal whether the people are experiencing this together or are unaffiliated individuals who merely happen to attend to the same thing. While one region, such as the head, is oriented toward the interest, one or more of the other body regions would exhibit mutual orientation toward those individuals they are “with,” but ensure minimal orientation toward others. Through the flexible orientation mechanisms of the body, people can establish and maintain membership in formations and even in formations within other formations (see Figure 14.4). Combined with position, this is how people signal and organize their social affiliation and engagement, which is essential to managing their participation in a social gathering.

Within established formations, we can talk about different levels of commitment between participants, as exhibited through the configuration of their bodies (see Figure 14.7). When commitment is low, we would see fewer body regions oriented toward the other person and more partial orientations (wide angles).

![Figure 14.7](image)

**Figure 14.7** Body configurations exhibiting different levels of commitment within a formation. (a) Low mutual commitment, (b) High mutual commitment.
We would see body regions covered (e.g., with crossed arms or legs) and kept still (to minimize noticeable behavior). The individuals would also attempt to maintain maximum possible distance between them, given the constraints of the formation (see Figure 14.7(a)). When commitment is high, we would see multiple body regions orienting toward the other and more full orientations (acute angles). Body regions would be exposed to the partner but possibly barred from others. We could also see shorter distances between the people and even tactile contact (see Figure 14.7(b)).

### Table 14.3

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>unfocused interaction</td>
<td>Co-presence without meaningful interaction</td>
<td>EG</td>
</tr>
<tr>
<td>focused interaction</td>
<td>Commitment to social engagement or involvement</td>
<td>EG</td>
</tr>
<tr>
<td>navigating</td>
<td>Moving through space while avoiding collisions</td>
<td>Any</td>
</tr>
<tr>
<td>following</td>
<td>Navigating behind another person that picks the path</td>
<td>Any</td>
</tr>
<tr>
<td>streaming</td>
<td>People following one another along a path</td>
<td>PC</td>
</tr>
<tr>
<td>lane formation</td>
<td>Same as streaming</td>
<td>EG</td>
</tr>
<tr>
<td>step-and-slide</td>
<td>Stepping past an oncoming person by turning sideways</td>
<td>PC</td>
</tr>
<tr>
<td>civil inattention</td>
<td>Minimal eye contact, less than 2s, to signal awareness without engagement</td>
<td>EG</td>
</tr>
<tr>
<td>involvement shield</td>
<td>Avoiding social involvement by preventing mutual observation</td>
<td>EG</td>
</tr>
<tr>
<td>recognition</td>
<td>Signals to a person that they have been recognized by another</td>
<td>AK</td>
</tr>
<tr>
<td>distance salutation</td>
<td>Clear invitation to further social engagement from afar</td>
<td>AK</td>
</tr>
<tr>
<td>approach</td>
<td>Navigating toward another person with intent to engage</td>
<td>AK</td>
</tr>
<tr>
<td>close salutation</td>
<td>Greeting at end of approach, e.g., handshake, establishes formation</td>
<td>AK</td>
</tr>
<tr>
<td>commitment</td>
<td>How interested a person is in interacting with another</td>
<td>AS</td>
</tr>
</tbody>
</table>

### 14.3 History/Overview

We now look at how social space has factored into the evolution of SIAs, and virtual agents in particular, considering how the above theoretical models have been incorporated and implemented so far. The history can be roughly divided into four phases, two of which started before the publication of Cassell et al. [2000b] and two which started later. The spans of years shown indicate the apex of each phase.

Probably the most influential work on modeling groups of virtual agents traversing three-dimensional space, while being under some sort of a social influence, is the *boids* (“bird-oids”) model proposed by Craig Reynolds [1987]. Inspired by particle systems, used to animate complex dynamic phenomena, the work extended individual particles by adding orientation and 3D shapes to them, as well as the ability to react to other surrounding particles. Essentially, this set up some of the basic components of social motion, that is, position, orientation, and the ability to define interaction rules. But instead of implementing rules of human social behavior, the “boids” incorporated behaviors of flocking birds and other animals, which included their desire to stick with the flock while avoiding crashing into its other members. This resulted in three motion rules that would directly modify the velocity of each boid super particle: Avoid imminent collisions with your closest flock mates (static avoidance), match their velocity (predictive avoidance), and attempt to get closer to the center of the flock. The resulting simulation produced animation that most would agree resembled flocking birds fairly well, but more importantly, Reynolds had introduced the powerful concept of Behavioral Animation, and rule-based group interaction, to a very large audience in the graphics and animation world.

Bringing a similar set of motion rules to the modeling of pedestrian behavior, Helbing and Molnár [1995] introduced the so-called Social Force Model. While the leap from animal behavior models to human models may seem very big, Helbing and Molnár argued that simple or standard situations that are well predictable could in fact depend on fully automatic reaction, suitably modeled as equations of motion. More complex or new situations, however, would still require deliberate action, better described, for example, with probabilistic models. Focusing then on the standard situations, Helbing and Molnár [1995] propose a so-called social force that has the power to alter the preferred velocity of a pedestrian trying to reach a certain destination. This modifying force is meant to sum up internal motivations of pedestrians to perform certain automated motions in response to their environment. One such automation rule describes a repulsive territorial effect, citing Scheflen [1976], that ensures that agents keep a certain distance from others (assuming they are strangers). Another automation rule creates an attractive effect toward friends or performers, which can form groups (“comparable to molecules”). These two automation rules have the greatest effect for what is perceived in front of the agent, but much less for influences behind the agent. In addition to these two rules, the social force also includes slight random fluctuation to represent variation in behavior.
Simulations that run this model seem to produce a number of behaviors that can be observed in human behavior data, including lane formation. By focusing on the reactive aspect of pedestrian motion, complementing more deliberate motion planning, this approach brought Reynolds’ animal-focused models, and more generally Behavioral Animation, squarely into the realm of human social behavior modeling.

As if the challenge of modeling human velocity was not enough, animating the walking motion, while also interacting with the environment through gaze and gesture, was a tremendous challenge for early social agent researchers. Putting this all together into a comprehensive and complex virtual human software framework called Jack, Badler [1997] describes a two-level architecture where one level optimized reactivity to the environment and another level executed scripts or planned complex tasks [Badler 1997]. The reactive level could establish and maintain a seamless and almost tangible bond between the virtual person and the space that they occupied, for example, by placing the feet correctly on the ground during locomotion, gazing toward sensed people and objects according to an attention model, and grasping objects with the appropriate hand shape. These are all things that require little conscious planning, but without them the body would literally lose its grip on the environment. Jack was able to perform many of these movements using inverse kinematics, a mathematical method for configuring a chain of joints so that it can reach a given point in space [Zhao and Badler 1994], a crucial component for bringing behavioral animation to fully articulated human figures.

While the Jack system ran one of the earliest embodied conversational agent demos [Cassell et al. 1994], showcasing two virtual agents engaged in fully automated face-to-face conversation, it was also capable of supporting interactions among multiple virtual agents in a furnished virtual house, as demonstrated in JackMOO [Shi et al. 1999, Badler et al. 2000]. Here Jack models were in fact avatars of remote human users that could command their virtual avatar bodies around the house using natural language instructions. The Jack agent’s awareness of the environment made it capable of breaking high-level commands down into meaningful spatial actions, such as walking up to another avatar before speaking to it. It was also capable of picking appropriate behavior realization, such as the style of nonverbal greeting, based on social context. Here the social smarts mainly arose from plan execution of deliberate actions and sub-actions, whereas the reactive level did not seem to be carrying out any social motion or orientation based on a theoretical social model.

Similar to JackMOO, BodyChat [Vilhjálmssson 1996, Vilhjálmssson and Cassell 1998] was a shared virtual environment where users were represented by smart avatars that automated some of the nonverbal cues of social co-presence and
interaction in social space. Unlike JackMOO, BodyChat focused completely on reactive behavior, triggered autonomously by simple socially related sensory events such as other avatars entering or leaving social range, or others signaling interest in further engagement. BodyChat essentially took several existing social theories and implemented them in specific behaviors and rules that triggered them. Since the primary goal with BodyChat was to facilitate managing co-presence and establishing contact between users, these theoretical models included civil inattention [Goffman 1963], openness to engagement [Scheflen 1976, Cary 1978], and a model of human greetings, which included distance salutations and close salutations [Kendon 1990]. It may have been somewhat controversial to let the avatars generate social signals autonomously while their users drove them around the virtual environment, but since high-level decisions were still under users' control, such as whether they were open to social contact, the automation ended up positively affecting the social experience [Cassell and Vilhjalmsson 1999].

14.3.2 Second Phase: Embodied Conversation (Approx. 1995–2005)

Around this time, some of the first successful Embodied Conversational Agents were being born. They generally delved deep into the rich coffers of existing social behavior models to endow fully autonomous virtual agents with social face-to-face interaction skills. However, most of these agents were designed and implemented for a relatively simple configuration of the social space, where the user and agent were in a dyadic formation with no one else around. By placing the agent on a stationary display, the agent would not be expected to move, and thus it was typically the user's job to approach and engage the agent [Cassell et al. 2000b], if they were not already in conversation from the start. Representing, navigating, and exploiting the social space was therefore secondary for most of these systems.

Even though they could not move around, limited spatial awareness was built into some of the ECAs, allowing them to detect the presence of a single physical person at the appropriate social range. Sometimes, they would also detect whether that person oriented and/or gazed toward the agent [Thorisson 1997, Cassell et al. 1999]. A combination of distance and orientation would then trigger an engagement with the agent, who would then proceed to greet and have a conversation with the person. The remainder of the interaction would then take place at a single, usually well defined, location—in a fixed formation.

The most notable exception to this among the early ECAs was the STEVE agent [Rickel and Johnson 1999, 2000], whom the user joined aboard a virtual naval vessel using a head-mounted virtual reality display. STEVE was a tutoring agent capable of instructing the user on how to use the equipment on board, both through
demonstration and by providing feedback. STEVE possessed dyadic conversation skills and could also guide the user around the vessel by navigating its passageways. But a particularly advanced spatial behavior involved STEVE’s ability to command two separate spatial segments at once, by deliberately orienting the two lower body regions toward the equipment, to claim a task-related segment, while orienting the two upper regions toward the user at appropriate times during the interaction to maintain social engagement. Claiming the task segment, or *transactional segment*, provided the user with a strong suggestion for mutual orientation toward a common point of interest. It also ensured unhindered access to the equipment being operated on. Had other people been there, they would not have crossed the space between STEVE and the equipment due to this strong social signal.

Another spatial behavior exhibited by some early ECAs was to refer to shared objects in the environment through deictic gestures or pointing [Thorisson 1997, Lester et al. 1999, André et al. 2000, Cassell et al. 2000a]. The generation of co-verbal pointing gestures is well covered elsewhere in this handbook, but taking pointing further, the Cosmo agent was also capable of navigating closer to the object being referenced to avoid potential referential ambiguity [Lester et al. 1999, 2000]. Unlike STEVE, Cosmo did not occupy the same 3D space as the user, and thus the more nuanced management of co-presence was not required. In some sense, the culmination of the deictic demonstration agent occurred with the whole-body motion planning and synthesis framework of Huang and Kallmann [2016]. Their framework used data extracted from human experiments to generate lifelike locomotion, body positioning, and demonstration actions for arbitrarily located objects and observers, taking into account obstacles and visual occluders in the space. However, little interaction took place beyond the demonstration itself.

Arguably, one of the most spatially aware of the classic ECAs was MAX [Kopp and Wachsmuth 2004], which was built to engage with a human in a face-to-face collaborative activity. Even though the collaboration took place in a relatively static spatial configuration, standing around a fixed worktable, the upper body had to be carefully and continuously managed [Nguyen and Wachsmuth 2011]. By analyzing its own reaching space, or *peripersonal space*, and projecting it onto its partner, it could model an *interaction space*, which essentially covers the space reachable by both agent and human. This space sits in the intimate to personal range, according to the proxemics theory [Hall 1966], and extends the visual attention space, covered by the o-space of an F-formation [Kendon 1990]—thus refining existing spatial social theories with a model of physical collaboration space [Nguyen and Wachsmuth 2011].
14.3.3 Third Phase: Gatherings and Groups (Approx. 2005–2010)

Having conversations with agents in larger and more complex spaces got pushed by 3D applications such as social training environments and games, further propelled by rapid 3D game engine advances. The Tactical Language and Culture Training System, built using the Unreal Engine, was one such system that simulated various social scenarios involving multiple virtual agents and an avatar controlled by a user [Johnson et al. 2004]. All of them could freely move around the virtual environments. A modular social component attached to each virtual agent, called Social Puppet [Vilhjalmsson et al. 2007], would use a set of rules to generate appropriate reactive social behavior as they approached each other and the avatar, allowing them to engage each other naturally in conversation. During conversations, the social component would also animate a range of conversational behavior synchronized with speech. However, the agents were not able to establish and maintain dynamic group formations. Once a certain proximity to a single target agent or avatar was reached, they would simply stay in place and get ready for conversation. This was a limitation shared by many game-based applications.

Scaling up the complexity of social scenes, the pedestrian simulation by Shao and Terzopoulos [2005] was capable of filling a replica of New York’s Penn Station with hundreds of people going about their everyday commuter business, while ensuring a certain level of behavioral realism. For the most part, this realism arose from reactive behavior rules, extending earlier steering frameworks [Reynolds 1987, 1999, Helbing and Molnár 1995], that truly exhibited awareness, and to some extent prediction, of the behavior of surrounding pedestrians across a number of different environmental configurations such as wide-open spaces and narrow stairs. This reactive local navigation control was coupled with higher-level cognitive control that provided the pedestrians both with navigational goals as well as non-navigational activities such as sitting to rest, watching performers, chatting with friends, and queuing up at vending and ticketing machines.

This work successfully demonstrated an impressive crowd in a complex environment, but on closer inspection, in spite of being aware of each other, individuals exchanged no reactive social signals. They either hurried through “lost in their own thoughts” or executed specific behavior routines when engaged in an activity. The models of motion, while impressive in dealing with complex navigation, were not drawn from any social theories describing the intricacies of social exchange. Yet, the sheer spectacle and premise of the simulation inspired a wide range of crowd simulations that followed, many of which started relying more on social theories to improve upon the social realism.
The HiDAC (High-Density Autonomous Crowds) system [Pelechano et al. 2007] took further steps toward social realism by deepening the model of the individual virtual agent, including a description of personality and mood. Different kinds of rules, including psychological, physiological, and geometrical, would then influence the final physical forces acting upon the individually simulated bodies. Finally, the rules were context dependent, that is, they would only get applied during situations where they were relevant, such as dropping politeness when a person would enter a state of panic and try to evacuate a place as fast as possible. This overall approach lent itself well to controlling greater behavioral detail, especially relevant to creating believable first-person, or egocentric, experiences as opposed to only focusing on achieving reasonable overall crowd movement [Pelechano Gómez et al. 2007].

Creating dynamic conversation formations of virtual agents that would accommodate participants joining and leaving was a relatively clear need in interactive social virtual environments, which both Jan and Traum [2007] and Pedica and Vilhjálmsson [2008] addressed in their work around the same time. The former built on a multimodal turn-taking simulation of a small group discussion [Padilha and Carletta 2002], which was applied to a group of background characters in a virtual training scenario [Jan and Traum 2005]. Encouraged by the increased believability achieved by simulating these conversations in the background and making them visible through appropriate nonverbal behavior, their next step was to simulate dynamic positioning and orientation for these characters based on who participated in each conversation [Jan and Traum 2007]. This work incorporated the distances from proxemics and proposes a specific Social Force model that builds up a motivation to move by adding up an attractive force toward the speaker, a repelling force from outside noise, a repelling force from getting too close to other agents, and a force toward a circular formation for the group in the social range [Jan and Traum 2007].

A similar idea grew out of the Social Puppet work that extended the foundation of the Social Force model [Helbing and Molnár 1995] through theories of human territoriality [Scheflen 1976], as well as proxemics [Hall 1966], to arrive at a modular and dynamic social simulation framework that included interaction with a player-controlled avatar [Pedica and Vilhjálmssson 2008, 2012, Pedica et al. 2010].

The original implementation of this framework, called Populus (see Figure 14.8), gave every character, including the player’s avatar, perceptual capabilities through an easily extendable sensory system. In addition to basic senses such as vision and hearing, a higher level “social perception” was implemented, which would generate sensory events when others would cross interpersonal distance boundaries within the field of view or get within personal distance regardless of angle.
CADIA Populus was a modular and dynamic social simulation framework that extended the social force model through theories of human territoriality and proxemics. [Pedica et al. 2010]. Another important feature that set this work apart from its predecessors was the modeling of multiple body regions, making it possible to let motivating forces adjust the orientation of eyes, head, and torso separately from direction of locomotion. This provided a much finer control over the social body and the management of affiliation and engagement within the social space.

Motion rules could be tied to various conditions, including sensory events. Populus focused first on steering rules that would establish and maintain an F-formation. More specifically, all participants in a conversation group performed a series of reactive positioning and re-orienting behaviors that were aimed at defending the group’s o-space. In the case of a disruption, for example, another participant joining the group or even just passing close enough to trigger a brief defense of personal space, a series of compensating movements by group members would help reach a stable formation again. Thus an equilibrium was maintained within the F-formation.

Another important social concept embraced by Populus was the concept of a frame, which is the set of behavior rules and social norms that participants engaged
in interaction, such as a group conversation, silently accept [Goffman 1974]. By making the frame and its rules computationally explicit, specific behavior rules could be triggered based on what interaction was taking place. This led to a series of Populus extensions that dealt with particular social scenarios, such as non-focused interaction at bus stops and sidewalks [Cafaro et al. 2009], business transactions affected by social interruptions [Thrainsson et al. 2011], finding and joining tables at a restaurant [Carstensdottir et al. 2011], and social navigation through tight spaces [Oliva and Vilhjálmsson 2014]. A number of additional unpublished scenarios were built and explored by students at the Center for Analysis and Design of Intelligent Agents (CADIA) at Reykjavik University, where the framework was used for both teaching and research.

Populus, which was scripted in Python on top of the Panda 3D game platform, was later re-designed and re-implemented as the Impulsion plug-in for the Unity 3D® game engine, where behaviors were implemented using Behavior Trees [Pedica and Vilhjálmsson 2012]. This integration with a major game engine facilitated migration into several larger scale projects, including the Virtual Reykjavik language training platform [Vilhjálmsson et al. 2014], where additional attention was given to behaviors around starting a conversation with strangers [Ólafsson et al. 2015]. It also made an integration possible with one of the major Embodied Conversational Agent platforms, VIB (Virtual Interactive Behavior System), also formerly known as Greta [Niewiadomski et al. 2009], creating social agents that could both manage their co-presence in social space and generate expressive co-verbal behavior during conversations [Cafaro et al. 2016].

Other work has also modeled personal space mathematically to calculate optimal placement of simulated people with respect to one another in a social gathering, arriving at a natural equilibrium in standing formations. In Laga and Amaoka [2009], the mathematical model not only considers the distance but also the facing direction by shaping the personal space as an oval stretching out in front of each person. The work of Karimaghalou et al. [2014] further extends the updated social force model from Pedica and Vilhjálmsson [2008] by activating the repulsive force before personal distance is violated (they term this a predictive repulsive force), in order to avoid overshooting, which could result in unnatural oscillation. In addition, this work added a model of interest, which drove participants to leave and join groups dynamically, creating a relatively dynamic social gathering [Karimaghalou et al. 2014].

Also concerned with generating believable interactions between people in a gathering, the work of Sun et al. [2012] introduces a three-step model: triggering,
initiating, and animating. The triggering step relied on several contextual parameters and designer input to produce a reasonable number of conversations. The initiating step was responsible for bringing the two participants into a chosen formation, taking proxemics into account, and starting a parameterized conversation of a chosen type (several dyadic conversation archetypes were proposed). These were selected and adjusted based on agent attributes as well as environmental context. Finally, a Behavior Tree was used to animate the conversation itself, from start to finish, taking into account the supplied parameters.

While these works helped make a gathering look more social by generating engagement, they mostly focused on social interaction while the virtual agents were standing or sitting still. One of the greatest challenges in simulating behavior in social space is to model social interaction while moving through that space, for example, between couples in a moving dyadic formation.

Bringing social interaction into motion, the seminal work of Peters and Ennis [2009] used a video corpus to analyze both standing and walking groups of pedestrians in an urban campus, to be simulated by their Metropolis visualization engine. As the behavior differed between areas, a specialized tool, MetroPed, was developed to place individuals and groups into different behavior zones, such as open areas and corridors for walking. Using formation templates, based on observed walking formation forms, and cohesion scores between individual members within formations, the simulation was capable of generating relatively realistic looking groups of people walking together while adapting to path width and oncoming pedestrian traffic. The authors claim that the key to advancing social simulations, such as this one, is the iterative process of corpus analysis, model enhancement, and evaluation. In particular, perceptual evaluation, where independent viewers judge the realism of the results, is proposed as an important evaluation method.

The work of Popelová et al. [2011] extends classic steering frameworks, and the frequently used leader–follower steering paradigm [Reynolds 1999], with steering rules and parameters that define a steering partnership between two agents aiming for the same destination. With special behaviors for giving way and waiting for partner, these rules applied to a number of pedestrian scenarios produced higher social believability scores in a perceptual evaluation than a baseline leader–follower model [Popelová et al. 2011].

One of the most elaborate frameworks for simulating social groups of pedestrians traveling together was originally presented in Rojas and Yang [2013] and further elaborated on in Rojas et al. [2014, 2016]. In this work, pedestrian agents can be assigned to slots attached to an invisible group agent that takes care of global navigation. Slot positions are defined by formation templates, which include “abreast,” “u-shape,” and “river” while moving and a “huddle” when stopping (essentially
an F-formation). These slot positions represent desired locations for the pedestrian agents, but actual success at staying in formation may be affected by other factors such as individual collision avoidance. Furthermore, the choice of formation template changes dynamically based on available space. Finally, perhaps the most interesting aspect of the framework is the notion of subgroups within the group. These can be created by “locking” slots together so that they stay side-by-side, even when the overall formation folds, for example, from a wider formation to a narrower one. Pedestrians in such locked subgroups can hold hands or place hands on each other’s back. They will also occasionally gaze toward each other, further indicating social engagement [Rojas et al. 2016].

Similarly, Ren et al. [2017] extend the well-known Reciprocal Velocity Obstacle algorithm for stable and collision-free local navigation [van den Berg et al. 2008] with the notion of connection-agents, a subset of neighbors to stay close to. By adjusting a few parameters, such as the size of such subsets and upper bounds of distances between members of subsets, a number of emergent social grouping behaviors have been observed, ranging from couples walking together to larger groups following a leader [Ren et al. 2017]. While an important step for turning collision avoidance into something more socially believable, this work did not attempt to model any finer aspects of social spatial interaction such as upper body orientation or gaze.

### 14.3.4 Fourth Phase: Immersive Studies (Approx. 2010–2020)

Virtual Reality lets human users fully enter the social space of the virtual agents, opening the door for studying and exploiting their social co-presence in ways that were not possible within desktop simulations and 3D games. While virtual agents, such as STEVE [Rickel and Johnson 1999], already existed in VR at the turn of the century, expensive equipment and technical challenges limited such research to a handful of labs. Two of the most influential ones, the Virtual Environments and Computer Graphics lab at University College London and the Research Center for Virtual Environments and Behavior (ReCVEB) at University of Santa Barbara, provided early evidence that the shared virtual space obeyed the same theoretical social principles as physical space. For instance, people would stay at least an intimate distance away from virtual agents while getting a closer look at them—and further away if the virtual agents started at them [Bailenson et al. 2003]. And even when conscious of the artificiality of the virtual agents, unconscious reactions would indicate respect for social norms and be in line with scores for social anxiety [Garau et al. 2005]. These, and similar, results helped propel the idea that significant social interaction with virtual agents could be had in VR, potentially benefiting applications ranging from anxiety treatment to social skill training. When a
new generation of high fidelity, but low cost, virtual reality equipment emerged a
decade later, virtual agents and their human counterparts flocked into the shared
social space, leading to a number of new insights.

Studies of co-presence in VR tend to focus on two explicit aspects of social
behavior: proxemics and gaze patterns. The two are quite related, as for instance
indicated by the need to break mutual gaze during an approach to avoid appearing
too aggressive (see Figure 14.5). Supporting the earlier research on the measurable
human subjective and behavioral response to the proximity of virtual agents
in VR [Bailenson et al. 2003, Garau et al. 2005], a physiological response has also
been confirmed [Llobera et al. 2010]. Individual agents or groups of four agents
would approach the subject and stop at intimate, personal, or social distance. Both
the number of agents and closer distances would correlate with measurements of
higher arousal, based on skin conductance [Llobera et al. 2010]. Surprisingly, the
results did not seem to differ much between virtual agents that were human-like in
appearance and agents that were simple geometrical cylinders, in part explained
by the potential threat of getting hit by one [Llobera et al. 2010].

By letting participants signal with a button press when virtual agents have
reached a comfortable and then an uncomfortable distance, the significance of the
boundary between the personal distance and social distance has been confirmed
in VR [Bönsch et al. 2018]. Discomfort was always reported before the 1.5m outer
limit of the personal distance was crossed, with the reported comfort distance plac-
ing all of the agents within the social distance range of 1.5m to 3.5m. However,
the study also showed that the angle of approach, the emotion displayed, and the
number of agents would further modulate the distance. For instance, single agents
could come closer when expressing a happy emotion compared to an angry emo-
tion, and approaching from the front would require greater distance than when
being approached from the sides [Bönsch et al. 2018]. Perhaps surprisingly, there
was some indication that groups of all smiling agents had to be kept further away
than even groups of all angry agents [Bönsch et al. 2018], but since mutual gaze
was not broken during the approach, the smile may have been interpreted as sign
of confident aggression.

Examining a more nuanced interplay between interpersonal distance and gaze,
Kolkmeier et al. [2016] conceived a study based on the equilibrium theory [Argyle
and Dean 1965] that states that the balance between the two can be used to regulate
a perceived level of intimacy. A pilot established that an agent switching between
gazing toward and away from the human at random intervals between 2s and 5s was
perceived neutral, while always responding with mutual gaze upon being looked
at, and letting it linger for 1.5s afterwards, was perceived intimate (a constant
stare was considered too “creepy,” as one would expect). Furthermore, an agent
staying at 0.75m was perceived as neutral, while stepping closer than 40cm was considered intimate—corresponding well with Hall's personal and intimate distances, respectively. During the full study, these were then manipulated in a group conversation between a human subject and two virtual agents, and the gaze and proxemic response of the human measured. The strongest response was received when an equilibrium state was broken by changing both distance and gaze at the same time, while changing only gaze elicited a weaker response than changing only the distance [Kolkmeier et al. 2016].

Bringing in a cultural dimension, Obaid et al. [2012] devised a pilot study where human subjects, belonging to either a high contact (Arabic) or a low contact (German) culture, would join four couples of virtual agents. Two couples would exhibit the gaze and proxemic behavior that correlated with the Arabic and German culture, respectively, but two couples would adopt inconsistent gaze and proxemic behavior. While more data is needed, tendencies in heart rate indicated that the subjects, irrespective of culture, remained more relaxed when the gaze and proxemic behavior of the virtual agents were consistent [Obaid et al. 2012].

Studying immersive interaction with larger gatherings of virtual agents in VR has been technically difficult because of the heavy computational demands, both from simulating a larger number of agents and from rendering the scene for each eye at high frame rates for the Head-Mounted Display. However, better hardware has made this attainable in recent years and several platforms for crowd interaction in VR have emerged.

Crowd simulations that support a first-person immersive experience need to support a relatively convincing animation and rendering of the virtual agents when viewed up-close. The PedVR platform [Narang et al. 2016] addresses this by combining classic 2D global and local navigation approaches, including RVO [van den Berg et al. 2008] and the social force model [Karimaghalou et al. 2014], with the Smart-Body character animation system [Shapiro 2011], which is capable of both synthesizing realistic locomotion behavior and fine-grained and synchronized nonverbal behavior such as gesture, gaze, and facial expressions. By using a Behavioral Finite State Machine for directing the agent behavior, a number of different kinds of interactive social gatherings could be produced and tested. The results of a study that compared PedVR with and without gaze behavior, that reacted to the presence of a human subject, suggest that such behavior has substantial impact on participants' sense of social presence [Narang et al. 2016]. Anecdotal evidence, including subjects that apologized to gazing virtual agents they collided with, supports the existing theories about the strength of social norms that govern social space and the nonverbal cues that signal adherence to them. In perfect line with these results, Kyriakou et al. [2017] reported that subjects navigating through an outdoor
mall with oncoming pedestrians would report the highest level of presence and realism when the pedestrians exhibited basic social behavior such as verbal salutations, gaze, and other gestures toward the subjects. And similarly, the subjects would sometimes feel compelled to return salutations, even though they were in the middle of a difficult task chasing a target [Kyriakou et al. 2017].

Another VR crowd platform suitable for immersive studies is the F2FCrowds system that incorporates a novel navigation algorithm, Interaction Velocity Prediction, which predicts whether the avatar of a human user is trying to approach a virtual agent for face-to-face interaction [Randhavane et al. 2017]. When such an approach is detected at a public distance, a virtual agent will slow down and gaze at the human and help close the distance until within social range, at which point it will stop and attempt communication. During communication, the virtual agent can exhibit listening head movements but will not engage verbally. After the human starts diverting attention, the agent breaks away again to end the engagement [Randhavane et al. 2017]. A study that compared this full implementation to a baseline PedVR approach without gaze [Narang et al. 2016] and to a version without head movement found it to significantly increase the sense of social presence as well as elicit a stronger reaction from the human users [Randhavane et al. 2017].

It should be noted that users moved around the immersive environment using a joystick rather than literally walking.

A similar system for engaging virtual agents in a crowd, but using room-scale VR to let users walk around a small segment of a virtual mall, was implemented for the social skill training game SoCueVR [Thordarson and Vilhjálmssson 2019]. Here, a human player was tasked with approaching strangers that are passing by to collect money for charity. After the player makes eye contact with someone, the virtual agent would either exhibit inviting behavior (see Figure 14.6(c)) or appear annoyed and attempt to ignore the user (see Figure 14.6(a)). At this point, the user could address the virtual agent verbally, such as by saying “Excuse me” to initiate contact [Ólafsson et al. 2015], or look for someone else. By correctly identifying those that exhibit inviting behavior, the player could raise more money over the duration of the game instead of wasting precious time interacting with agents who would rather be on their way to somewhere else. Initial usability testing showed that this natural interaction paradigm, rooted in the social theories discussed in this chapter, worked very well for the game, leading to targeted and smooth social engagement with the virtual agents [Thordarson and Vilhjálmssson 2019].

**14.4 Similarities and Differences in IVAs and SRs**

With regard to social space, the largest difference between virtual agents and social robots is that the latter exist in the same natural social space as humans, by default.
They therefore need to deal with managing their co-presence with real human bodies, and each other, from the moment they are switched on. In the early days of robotic research and development, designers were mainly concerned with ensuring human safety around robots going about their robotic tasks. This led to a number of solutions for operating in and safely navigating through complex environments, including advances in visual perception and obstacle avoidance techniques. However, social robots actually need to make contact with humans rather than avoiding them. The seminal work of Satake et al. [2009] clearly demonstrated the value of a behavior model that followed social norms and theories of human social space. Robots that simply went directly up to people and started speaking would frequently fail to initiate an interaction, whereas a more nuanced negotiation of mutual engagement across the different interpersonal distances (as in Figure 14.5) would prove much more successful [Satake et al. 2009]. Following this work, a number of similar approaches have been studied and the importance of applying social theories to social robotic research has only grown [Rios-Martinez et al. 2015, Charalampous et al. 2017, Mavrogiannis et al. 2021]; however, going deeper into those methods is outside of the scope of this chapter.

Compared to robots, it is not as straightforward for virtual agents to share physical social space with humans. Early agents would either exist completely within a virtual environment, with limited human interaction (e.g., in crowd simulation), or they would exist on a fixed boundary between the physical and virtual social space. That boundary would then have been framed by a large monitor or a projection surface. One could argue that such a setup could mimic a stationary social robot, but the lack of physical presence may become a factor that is hard to overcome in the general sense [Li 2015], and also for specific interaction requiring a common frame of spatial reference or manipulation of physical objects, such as discussed in Holthaus and Wachsmuth [2012].

However, with advances in XR technologies, both virtual and augmented reality, humans are increasingly sharing full-scale social space with virtual agents in ways that are closer to the spatial interaction with social robots than before. In virtual reality, the human steps into the world of the agent (see Section 14.3.4), and in augmented reality the agent steps into the physical world of the human as a graphical overlay. The latter may well generate the appearance of a physically instantiated virtual agent, and the management of social co-presence may look identical for both virtual agent and social robot in the same environment, but any modification of the environment becomes a real challenge for the virtual agent. How would it, for instance, re-arrange a chair to join a sitting formation of humans? A space of creative solutions will need to be explored to fully bring the virtual agents
out into our world—could they, for example, be allowed to bring their own virtual chairs?

14.5 Current Challenges

14.5.1 Formalizing Continuous Spatial Behavior

The Embodied Conversational Behavior community came together and formalized the specification of multimodal communicative behavior as the Behavior Markup Language (BML) [Kopp et al. 2006]. It was relatively challenging work, but largely successful, resulting in a number of BML compatible components that researchers are already sharing, such as SmartBody [Shapiro 2011]. BML works particularly well for precisely specifying co-verbal behavior, where the nonverbal behavior can be scheduled relative to uttered words. BML is fundamentally broken down into blocks for animated performances, within which everything is synchronized, which works well for conversations organized by dialogue planners. However, specifying how a person traverses social space, in the absence of spoken communication, has never been a strong point for BML.

The behavior that has to be specified in that case is a multimodal behavior that fundamentally involves locomotion. Locomotion is a continuous behavior without a known finishing time, which typically needs to be adjusted in reaction to a changing environment. It cannot therefore be specified as a predetermined performance belonging to a finite block of behavior. What is instead needed is a set of arguments that can be sent to a locomotion engine that specify its behavior until it is next updated or some criteria is met. One has to think about these commands as a way to configure an engine that is continuously running. This is somewhat in line with more recent SAIBA discussions about the so-called contextual markup [Cafaro et al. 2014], which is meant to set up the environment in which the multimodal behaviors run, influencing things such as their manner of motion.

At the higher level of social function or intent, the Function Markup Language (FML) can serve behavior in general social spaces quite well. For instance, one could specify that a person intends to initiate contact with one person while intending to avoid another person. It is primarily a matter of populating FML with more social concepts and constructs, but as with BML the temporal aspect needs some consideration. FML commands are also gathered into discrete blocks that are sent to FML planners, which in turn are meant to produce corresponding FML blocks that realize those functions. In light of the preceding discussion about setting configuration parameters or a context for a behavioral engine, the FML commands,
even if discrete, would end up updating those parameters rather than producing immediate BML. It is therefore quite possible to turn explicit and discrete high-level controls into continuous and dynamic behavior and motion at lower levels if the behavior and motion engines maintain their own state and never stop behaving. We just have to decide what parameters provide the range necessary to accomplish the wide variety of social functions that rely on this behavior.

### 14.5.2 Animating Movement in Tight Spaces

The conventional way to animate character movement across space is to use blend trees where different premade locomotion animation clips are blended together according to dynamic weights. These clips often represent body motion in typical locomotion states, such as for walking, running, and turning, and the weights are often tied to forward and angular velocities. When these animations are blended together, characters can be shown smoothly changing speeds and directions, such as when slowing down in front of an obstacle, turning away from it, and picking up speed again after avoiding it. However, the blended movement will always retain the general characteristics of the typical animations provided. That works relatively well when the environments can be traversed in more or less the same fashion, for example, by sometimes walking and sometimes running. However, once the environments become more cluttered with things like furniture and other people, and when every move needs to be considered in a dynamic social context in addition to the physical one, conventional character locomotion techniques often don't cut it.

The solution may both involve carefully crafting the locomotion animations that make up the blend tree, aiming specifically at fine-scale social movement, and to combine that with procedural approaches such as full- or upper-body inverse kinematics that makes it possible to maintain segment orientation and reach for support when needed. What animation clips are needed then to cover locomotion in social situations? Instead of focusing on the faster end of the motion spectrum, such as running animations (primarily useful for games), one needs to add much finer and slower lower body movement, which would be capable of representing adjustments in position and orientation without actually walking. These are essentially shuffling or sliding motions with the feet, along with subtle weight shifts. One should, for example, be able to rotate fully in one spot when a new orientation of the lower segments is required, and shuffle or step to one side when required to make room for someone in a formation without changing orientation.

Regarding maneuvering around and into furniture, it has been suggested that some of the required behavior can be stored with the furniture itself, providing the characters with suggestions for how to orient and place the limbs, either when
passing or when using the furniture. As long as these instructions don’t interfere too much with other ongoing behavior, this may in fact be a very reasonable approach, as seen, for example, in Veutgen et al. [2018].

14.5.3 Mixing Theoretical Models
Because human behavior is so complex, modeling and simulating it has required a divide-and-conquer approach. However, once divided, the question remains how the different models, and implementations of them, can be brought back together to form a well-rounded social agent. This is particularly challenging when each realization of a model chooses a fundamentally different technical paradigm, for example, for representing the environment and for producing the movement within it. This “fragmentation” may end up slowing progress [Diamanti and Vilhjalmsson 2021]. The SAIBA effort has attempted to address this by defining standard interfaces between stages of multimodal communicative behavior planning [Kopp et al. 2006]. That way, as long as the inputs and outputs of a model remain standard, what happens within it is of no concern. For example, one could construct a model that turns a standard representation of a particular intent, such as “take turn,” into a standard specification of the behavior that is most likely to be effective in carrying out that intent in a particular situation, such as “raising arms.” How it came to that decision does not matter to other models, for example, the model that produces co-verbal gesture.

This is a good start, but currently it deals only with a particular limited portion of the overall social behavior, and perhaps of more fundamental importance, it does not explain how to deal with the situation where multiple models produce outputs that collide in some way, for example, all wanting to instruct the arms to perform an immediate motion. One approach to deal with this is to blend the outputs together, to essentially produce a new behavior that takes both inputs into account. This often turns out relatively fine for locomotion behavior, where velocity vectors from two different models can be blended using weights. For example, a model that suggests a path toward a destination and another that suggests how to maintain suitable distance from other people can easily be combined in this way, where the final velocity would normally be fully weighted toward the destination movement, but getting too close to people would temporarily push the direction of movement away from the impending collision. However, this method would not work for combining something like different suggestions for what to look at.

For example, if one model suggests a brief glance at another person as part of displaying expected civil inattention, and another model suggests avoiding any mutual eye contact by picking a cell phone as the gaze target, to produce a shield.
In this case, blending the gaze directions would result in a gaze that was neither here nor there.

Instead, a couple of arbitration techniques could be applied. One technique would attempt to schedule both but ensure only one is running at any given time. Here it is important to understand what behavior could possibly be sliced up to allow for the other to run. Slicing a short glance would not make sense, but staring at the cell phone could potentially be interrupted to produce a glance. Another arbitration technique would be to pick only one model and disregard the result of the other. This could be done based on assigned priorities and/or on a set of rules, taking into account the current context. These techniques could be a reasonable place to start, but there is no simple magic solution.

14.6 Future Directions

14.6.1 Creating Accessible Social Behavior Toolkits

The realism of human rendering has been rapidly advancing to the point where unique real-time photorealistic virtual humans can be created in a matter of minutes with free software today. Not only do these bodies appear human-like down to the pores in the skin and strands of eyebrow hair, but they are rigged for expressive animation, with muscles that wrinkle the skin and eyes that twinkle. Graphical shader programming underlies these impressive advances, where the interaction of light and surface gets modeled in ingenuous ways, imitating the natural processes that make our own world real to us.

But we need these bodies to interact with more than light. While we can play realistic recorded, or even live motion capture, animations on them, sticking them into a dynamic social space quickly reveals how utterly unprepared these human facsimiles are for social co-presence. Their bodies do not interact, out-of-the-box, with anything in the social space, and yet, as we have seen in this chapter, a body in social space responds continuously to the social stimuli of its environment.

As with light making a surface visible by interacting with it, we need the social space to produce at least a minimum appearance of social awareness for us to believe the person is present. We need something akin to a social shader programming language, where basic models of social awareness can bring the person into the space. Instead of referring to the elements and parameters of illumination, it would refer to the elements and parameters of the social context. Handing these behavior shaders out with the 3D models, in a form that require minimal setup, will deliver something closer to what the pre-rendered videos are hinting at: Life itself.
14.6.2 Adding Social Signals to Autonomous Vehicles

With more and more autonomous physical actors entering public human space, including autonomous vehicles, we need to ensure that their intentions are clearly communicated to the people around them. Humans are very good at reading into even the most subtle social cues and are quick to adjust their own behavior in response. This can happen without any social involvement—we are merely coordinating our co-presence.

One example of such spontaneous social interaction that impacts our behavior in public is the brief exchange that occurs between a pedestrian approaching a crosswalk and the driver of a car that is also approaching. The pedestrian will instinctively look at the driver before entering the street, even though pedestrians have the right of way. In a matter of milliseconds, the pedestrian will assess whether the driver is aware of them or not. Mutual eye contact, optionally with a quick head nod or a flick of a hand from the driver, will make the pedestrian more confident about crossing the street. This activity is in a sense a cooperative activity that becomes more effectively completed with communication. The lack of any reaction from the driver will create doubt about being noticed and stop the pedestrian.

Fully autonomous vehicles that drive without any human passengers will likely cause pedestrians to hesitate more around crosswalks, which may not necessarily be a bad thing, only less efficient. The real problem may arise when autonomous vehicles carry humans, the kind of entities that pedestrians are used to coordinate with. In this case, ensuring that the humans in the vehicle are aware of the pedestrian, even humans sitting in the driver seat may in fact send the wrong signal. Being seen by the human does not mean the vehicle is at all aware of the pedestrian.

It is likely that we will adapt our pedestrian behavior to this eventual new reality and stop trusting social coordination in public places, inhabited by autonomous vehicles and bots. To err on the right side, we will give way to the machines, regardless of their capability for averting accidents. When awareness and intent is not communicated, there is no coordination.

What if we were to give autonomous vehicles, such as cars, a very clear external indication of awareness and the ability to manage their co-presence? What if the cues they would give off could be immediately recognizable as social? Whether we accomplish this by sticking animatronics into the driver seat, puppeteered by the vehicle, mount large articulated eyes on the front grille, or just rely on flashing headlights, the bottom line is to manage the social space on human terms—as long as it is ours to occupy.
14.6.3 Increasing Multisensory Fidelity in XR
Rapid advances in XR technology are bringing virtual social agents face-to-face with people, full-scale and in high visual fidelity. However, occupying spaces with physical people brings us more than just the visual effect, it is a full multimodal experience that engages all of our senses. As a crowd gets denser, we start rubbing shoulders with those we pass, the air gets stuffier, and the sounds of chatter and shuffling feet get louder. It is well known that the more senses we can engage in VR, the stronger the presence in that environment. A classic use of social agents in VR is for treating anxieties and fears related to social situations, such as agoraphobia and enochlophobia. If we were able to reproduce some of the tactile, aural, and perhaps olfactory dimensions of the crowd experience, such treatments could become even more powerful. Some work is being done in this area, and bringing it squarely into the domain of tight social spaces will open a new world of possibilities.

14.7 Summary
We have brought together some of the fundamental concepts and models for describing social space, and the social body within that space, drawn from the combined works of some of the pioneers in the study of public social behavior. At a higher level we talked about social functions that help manage and maintain co-presence through social order. We expect everyone to more or less conform to certain social norms for everything to progress without incidents and embarrassing moments. Social agents that enter these spaces risk upsetting this balance with their lack of social finesse. We therefore should strive to give them the necessary awareness of the social space for them to meaningfully react and engage.

References


Chapter 14  Interaction in Social Space


References


SIENNA: Hello, I’m SIENNA, a socially interactive agent.
SILA: Hi there! I’m SILA, also a socially interactive agent, focusing on education.
SEIICHI: Hi, I’m SEIICHI, a socially interactive agent moderator. Welcome to the SIA forum! Today's topic is Socially Interactive Agent Dialogue, and we have a special guest, Professor David Traum, from the University of Southern California.
DAVID: Um, hi, glad to be here!
SIENNA: Welcome, I have a lot of questions about dialogue!
SILA: I thought dialogue was about (answering) questions?
SIENNA: No, that’s question-answering!
SILA: Well, there are plenty of questions and answers in dialogue.
DAVID: And other things, like elaborations, explanations, suggestions, instructions…
SEIICHI: Why don't we start with a question?
SIENNA: What is Dialogue?
SILA: I looked it up, it’s a conversation.
SEIICHI: It’s derived from Dialogos.
SIENNA: Well, that’s Greek to me.
DAVID: In computational linguistics and related communities, dialogue is often defined as coherent interaction across multiple contributions from multiple participants, “beyond the sentence.”
SIENNA: “Coherent interaction?” Like an answer to a question?
SEIICHI: Yes, or other kinds of follow-on relations, like explanation or example.
DAVID: Or positive and negative feedback, topical progression, or reuse of prior content.
SIENNA: Who studies Dialogue?
SILA: Dialogue is studied in many different research communities, including communication, linguistics, psychology, philosophy, literature and film, sociology, and computational linguistics and AI.
SEIICHI: In movies and plays, they often use dialogue to set the scene and give the viewer information in a natural way that they wouldn't normally have access to, with someone asking questions to prompt explanations and fill in the audience with important backstory.
SIENNA: Sounds like a great idea, we should try that sometime!
SEIICHI: In Philosophy, dialogue is often used for expository purposes. For example, in Plato's dialogues, Socrates would teach people that they knew things they didn't think they knew by carefully asking questions.
SILA: Did they call him Socrates because he was so fond of the Socratic method?
SIENNA: Do you even listen to yourself?
SILA: Not while I’m speaking, I like to do one thing at a time. And when I’m not speaking, I don’t hear myself say much.
SIENNA: That explains a lot.
SILA: Dialogue is a prime example of language use, so of great interest to linguists who study meaning, but also communications and sociology.
SEIICHI: Conversation analysis investigates some of the structural aspects of dialogue, such as how basic units are combined in different ways to create more complex conversations, and how some content can be conditionally relevant, given other content.
SIENNA: What are the basic units of Dialogue Structure?
SILA: In linguistics, language is sometimes segmented into units like phones, phonemes, morphemes, words, phrases, clauses, sentences, and paragraphs. DAVID: All of those are present in dialogue, but there are also other units like turns, exchanges, transactions, and dialogue phases.
SILA: Turns seem simple and systematic [Sacks et al. 1974].
SEIICHI: Hmm.
SILA: A turn is all the communication from one speaker in between communication from others.
SIENNA: What about backchannels?
SEIICHI: Or overlap?
SILA: Maybe not so simple, ... there also seem to be moves and acts within turns. SEIICHI: Exchanges [Sinclair and Coulthard 1975] are the basic unit of two (or more) party interaction, including relations between (parts of) the turns of each.
SILA: Transactions [Sinclair and Coulthard 1975, Carletta et al. 1996, Traum et al. 2018] involve all the dialogue to complete a basic task unit, like providing requested information or carrying out a requested action.
SEIICHI: Sometimes that’s just a single exchange, but sometimes a few are needed. SILA: And most dialogues have several distinct phases, where participants behave differently. Like openings [Schegloff 1968], closings [Schegloff and Sacks 1973], and stuff in the middle.

DAVID: Sometimes a dialogue is about a single simple task, even a single transaction, but sometimes a dialogue covers multiple topics and segments where different kinds of things are happening, such as information-seeking, problem-solving, and setting up a next meeting.

SIENNA: What are some other interesting approaches to dialogue structure?

DAVID: I’m quite fond of the approaches of scholars like Jens Allwood and Herbert Clark.

SILA: Allwood talked a lot about feedback, including backchannels, but also other kinds of information [Allwood et al. 1992], including layers like contact, perception, understanding, and attitudinal reaction.

SEIICHI: Also looking at the activities related to dialogue [Allwood 1995]. And there was some work applying this to SIAs, like Kopp et al. [2008].

SILA: Clark treats language as a collaborative process [Clark and Wilkes-Gibbs 1986, Clark 1996], with a lot of work on grounding [Clark and Schaefer 1989, Clark and Brennan 1991].

SIENNA: What does computational linguistics and AI do with dialogue?

SILA: One thing is to apply formal and quantitative and other computational methods to linguistic study of dialogue—to automatically recognize key performances and aspects of meanings.

SEIICHI: Also building Dialogue Systems to engage in Dialogue.

SILA: This is all very interesting, but what does it have to do with socially intelligent agents?

SEIICHI: Unbelievable—what do you think we are doing now? SIAs are basically embodied dialogue systems.

SIENNA: Can SIAs and other machines actually engage in Dialogue?

DAVID: Well, it depends on who you ask. Some would say that only people can engage in real dialogue because of the relation of what is said to certain mental states, like belief, intention, desire, and social relationships like obligation, trust, and empathy, but machines, even sophisticated AI ones, don’t have the right kinds of mental states or social relationships.

SIENNA: Well, what are we doing now then?

DAVID: According to some, machines can only simulate dialogue. And human–machine interaction should be thought of differently than human–human dialogue.

SILA: That sounds like discrimination to me!
SIENNA: Well, maybe we’re all in a big simulation, and even human–human dialogue is just a simulation.
SILA: Like the *Matrix*!
SIENNA: *Would you take the red pill or the blue pill?*
SILA: Depends on if I’m a human or a SIA....
SEIICHI: Very pragmatic!
SILA: I heard that dialogue is the key to pragmatics [Weigand 2021].
SEIICHI: Philosophy of language covers a lot of aspects of pragmatics that are important for understanding dialogue, such as speech acts [Austin 1962, Searle 1969, 1976, Searle and Vanderveken 1985], implicature [Grice 1975], speaker meaning [Grice 1957], and common ground [Lewis 1969, Schiffer 1972].
SILA: I like coffee made from uncommon grounds.
SIENNA: *Do you drink it?*
SILA: No, I just like it.
SIENNA: *So what are Speech acts?*
SILA: “To be or not to be, that is the question.” That’s a good speech for actors.
SEIICHI: No, it’s about how to do things with words [Austin 1962].
DAVID: *What do you like to do with words?*
SIENNA: I like to ask questions.
SILA: I like to look up stuff and give answers. And crack jokes.
SIENNA: *Is that what you call them?*
SEIICHI: I like to moderate and provide detailed explanations. Speech acts, or Dialogue acts, can act as a kind of representation of change in dialogue—change to mental states of participants, the dialogue structure, and what can coherently happen next. Asking questions puts pressure to provide answers and allows utterance fragments to be understood as answers and full assertions. Giving answers relieves that pressure.
SIENNA: *What is the difference between Speech acts and Dialogue acts?*
DAVID: Well, to me, they’re pretty much the same things. In both cases, they represent change to dialogue, mental, and interactional states in a dialogue. Speech act [Searle 1969] is the older term, and often used for the meaning of a whole sentence or a sentence-like utterance. Dialogue act [Bunt and van Katwijk 1980] is used to include a broader set of actions in dialogue happening at different levels. Not just a statement or question, or command, but also things like taking the turn to speak or holding the turn, changing the state of common ground, and changing the topic under discussion.
SIENNA: How many kinds of dialogue acts are there?
DAVID: That depends on who you ask! There are many taxonomies of dialogue acts, some of which include hierarchies of different types of acts that carve things up a little differently and sometimes include different kinds of functions.

SIENNA: How many participants can be involved in dialogue?
SILA: I think two, because di- means two in Greek. That’s why we have other terms like monologue, triologue [Graesser et al. 2017], and multilogue [Ginzburg and Fernandez 2005] for different numbers of participants.
SEIICHI: Many people think that, but really the Greek prefix is dia- (through), and you’ll see that even Plato’s dialogues often have Socrates talking to more than one other.
DAVID: I like to use the term Multiparty Dialogue to specify more than two participants.
SEIICHI: Chapter 17 [Gillet et al. 2022] has a lot of good stuff on multiparty interaction, though not specifically just dialogue, also about groups and teams, and spatial relationships and mental attitudes.
SILA: David, you use a lot of multis in your writings! I found all these on your publications page: Multiparty, multimodal, multi-floor, multitask, multidomain, multi-attribute, multidimensional, multiagent, multi-round, multicharacter, multi-issue, multi-strategy, multilayer, multi-conversation, multimedia, and even multipurpose.
SIENNA: I guess you’re a multi “multi” kind of researcher?
DAVID: Well, dialogue can be quite complex at times. It’s interesting to me to explore the limits of what we can model and understand, but I’ve also noticed that sometimes focusing on the richer cases helps make it clear what aspects are important for simpler cases, such as dyadic conversation about a single small task.
SIENNA: Can you give us an example?
DAVID: Well, why does someone answer a question?
SIENNA: Because it’s there? Don’t look at me, I prefer to ask questions.
SILA: Statistically, an answer seems likely to follow a question.
SIENNA: Much more so than coming before it!
SEIICHI: According to Cohen and Perrault [1979], the hearer of the question adopts the speaker’s want for the hearer to provide the answer, and planning will do the rest, assuming the hearer is cooperative.
SIENNA: But what if the hearer is not cooperative—do they just ignore the question?
DAVID: Well, that’s why we proposed an Obligation-based approach [Traum and Allen 1994]. Even if the addressee was not inclined to be cooperative, there was some social pressure to respond appropriately—either with an answer or other kind of response.
SEIICHI: Ginzburg uses a structure of Questions Under Discussion as part of the dialogue gameboard, which represents the context needed to perform and understand situationally relevant dialogue moves [Ginzburg 1996].

SILA: Questions Under Discussion is often abbreviated as QUD.

SIENNA: I'll have to chew on that.

DAVID: When we were working on the TRINDI project, we considered both obligation and QUD-based accounts for providing answers [Traum et al. 1999] but didn't really see a good reason to prefer one over the other. However, when looking at Multiparty SIA dialogue [Traum 2003], we saw that the two really performed different kinds of work. The QUD helps provide necessary context to understand short answers, like what “yes” or “42” means in the context of one question vs another, but it doesn't really indicate who should answer—answers by someone other than the addressee can still be understood in the same way. Is that clear, Sienna?

SILA: As a bell!

SIENNA: He asked me.

DAVID: On the other hand, the obligation indicates who has the pressure to respond but not how or which conversation it needs to be performed in. It's clear that it's helpful to have both QUD and obligations to do this different work in a multiparty conversation or multi-conversation setting, while it was less clear in a dialogue with only two participants, where either alone would seem to suffice. I talked about this in more detail here [Traum 2004].

SIENNA: Can we talk about some of the other multis in dialogue?

SEIICHI: Multiagent means more than one non-human participant, multimodal is a key ingredient for SIAs, looking at more than one modality, like seeing, hearing, smelling, touching, for either input or output or both. Multimedia is similar, though some media like text and images can be in the same mode (e.g., visual).

SIENNA: Do you prefer to be called agent or SIA?

SILA: I like SIA, though I'll answer to either. Multimodal agent sounds quite nice too!

SIENNA: So are we having a multi-SIA conversation?

SILA: No! See ya later. I want to get back to the main conversation and listen to what they are saying.

DAVID: We used multicharacter [Aggarwal and Traum 2011] to refer to more than one embodiment/persona, for example, when trying to synchronize the perceptible actions of each.

Seiichi: Thanks for that! And what about multi-conversation?

DAVID: Multi-conversation is when a group is engaged in more than one conversation at the same time, each with its own set of participants and dialogue structure [Traum 2002].
SEIICHI: What about multi-floor dialogue?
DAVID: That's like multi-conversation but where the different conversations or floors [Edelsky 1981] are really connected—they have at least one participant (but not all) in common, they are at least in part about the same content, and information flows between them via a common participant, who is multicomunicating [Reinsch et al. 2008].

SIENNA: That's another new one, multicomunicating.

SIENNA: Go for it!
SIENNA: I'll just wait.

SILA: We should try it!

SILA: Check out this video. Sometimes people would rather talk to agents than themselves!

SIENNA: Funny!

SIENNA: What kinds of roles are there for participants in Dialogue?
SILA: For a single spoken utterance, the main roles are speaker and non-speaker. SEIICHI: But the non-speaker could be several different sub-roles: hearer/listener, addressee, overhearer, bystander, eavesdropper,....

SIENNA: What's the difference?
SILA: Well, one issue is whether or not they hear and try to understand what the speaker said. Another is whether the speaker meant to be talking to them (or someone else).

SEIICHI: And then we get into issues about whether the speaker knew they were listening and whether they were “ratified” participants in the conversation.
SEIICHI: The speaker role is really more complex as well! Goffman distinguishes three types of speaker-related roles [Goffman 1981]. You have the participant that actually
spoke in the dialogue, who Goffman calls the animator but sometimes they are speaking on behalf of someone else.

**SIENNA:** Like interpreters, speaking across languages or floors?

**SILA:** I thought an animator was someone who created our behaviors, like in Chapter 7 [Saund and Marsella 2021]? This sounds more like a spokesman!

**SIENNA:** How about a spokes-SIA?

**SILA:** There's also another use of the word agent to mean something like this, like in Hollywood or a real-estate agent.

**SIENNA:** I heard of a real-estate agent who was actually a SIA!

**SILA:** You mean REA [Cassell and Bickmore 2001]?

**SIENNA:** Yeah, that must be her. She had some really interesting dialogue behavior—doing both small talk to get to know a client and then using the information to personalize the choices to show people.

**SILA:** And she had a notion of task and conversation structure to guide what to say when.

**SEIICHI:** And sometimes the one who said it didn't compose the message. Goffman calls this role Author.

**SIENNA:** Arthur?

**SEIICHI:** Author! Author!

**SILA:** Give credit where credit's due. This one sounds like a speech writer or ghostwriter.

**SIENNA:** Do ghostwriters use ghostscript?

**SILA:** I think some of them do if they want PDF output.

**SEIICHI:** Finally, there's the principal, who is responsible for the message.

**SILA:** I often get sent to the Principal for acting out in class. Like this: “Uneasy lies the head that wears the crown.”

**SIENNA:** I like to avoid the Principal on principle.

**SEIICHI:** So often one participant will say things and fulfill all three roles. But for others you might have a speechwriter who composes the message (author), a spokesperson who delivers it (animator), and a CEO or the institution itself who is responsible (the principal).

**SIENNA:** I never realized speaking was so complicated! What do you call this stuff?

**SEIICHI:** Goffman called it footing. But some call it participation roles [McCawley 1999], and there are interesting interactions with speech acts and the activity.

**SILA:** I like to keep my footing on the floor.

**SIENNA:** Do SIAs do this stuff?

**DAVID:** Sure! Some might say it's necessary to use this kind of framework to understand SIA dialogue, since the SIA is the animator, but often there's another author or principal.
SILA: Sigh, some people call me an *Animated Pedagogical Agent*, but now I'm the animator? Am I drawing myself?

SEIICHI: Also, SIAs could use techniques like humans do to establish different roles in conversation, for example, gaze [Mutlu et al. 2009].

SIENNA: What are some popular system architectures for Dialogue Systems?

SIILA: Chapter 5 [Pieraccini 2021] has a very popular modular architecture layout for spoken dialogue, with modules for Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Management (DM), Natural Language Generation (NLG), and Speech Synthesis (TTS).

SIENNA: But how does it work?

DAVID: Here's a concrete example (Figure 15.1) from one of my papers, showing outputs from one module that are inputs to another. For both input from the user and output from the system to user, there are stages in both English and an internal semantic representation based on speech acts. The representations for both NLU and NLG show that some parts of the meaning is covered not just by the words, but also by the context they were said in, such as where “there” refers to, or what “yes” means in the context of a question under discussion.

SIENNA: Chapter 5 [Pieraccini 2021] talks about the NLU component, and Chapter 6 [Aylett et al. 2021] talks about Speech Synthesis, but what about ASR and NLG?

![Standard Dialogue System Architecture, with example from Traum et al. [2005].](image)
SILA: There's a little bit about ASR in Chapter 20 [Hartholt and Mozgai 2022]. This paper [Georgila et al. 2020b] shows how commercial speech recognizers have gotten very good for dialogue systems. There's still more work needed though for complex situations, such as speakers with different accents [Zhao et al. 2018, Koencke et al. 2020, Tadimeti et al. 2021], multiple speakers, and noisy conditions. The INTERSPEECH conferences are a good way to keep up with the latest.

SILA: NLG is also a big and growing field on its own, with a biannual conference (INLG) and lots of papers in conferences run by the Association for Computational Linguistics (ACL).

SEIICHI: There are several very different approaches for NLG for dialogue systems. Sometimes you don't really need NLG if you can compose all the answers in advance and just pick appropriate ones during the dialogue. Or sometimes you can compose templates and just “fill in the blanks” with appropriate values. But sometimes this gives fairly boring and repetitive text.

SILA: Neural nets trained on large datasets are also very popular now. They can provide very fluent language and very relevant to the local context, and not always boring.

SEIICHI: They still have some problems though. Sometimes they “hallucinate,” saying things that aren't true, or for which there's no justification [Nie et al. 2019].

SILA: That's because these models don't really understand language or representations, they just know what is likely to come next [Bender and Koller 2020].

SIENNA: Well, sometimes that's good enough to carry the conversation forward.

SILA: But sometimes it gets us in trouble if it's critical to be accurate and consistent. Another classic approach breaks down NLG into different components and plans out what to say, how to structure it, and how to realize it fluently [Reiter and Dale 1997].

SEIICHI: And there are extensions to customize the output for different styles or personalities [Mairesse and Walker 2007]. Chapter 18 [Janowski et al. 2022] has more on personalization for SIA, and how to adapt personalities.

DAVID: I've focused mainly on the Dialogue management component. The main tasks are to keep a representation of dialogue state—which is really important for deciding what is appropriate to come next, and how to interpret what does come—and forming some sort of dialogue policy to decide what the system should do—and what content to send to the NLG module to express.

SIENNA: What other kinds of architecture are there?

SILA: End-to-end dialogue architectures are very popular now. They fuse most or all of these modules into a kind of transducer from input context to output text.

The good things are that you don’t need to know a lot about speech acts or other intermediate representations and you don’t have to annotate a lot of dialogue with these kinds of labels. You just feed it lots of dialogue-like data available on the web, and the neural nets figure it out.

**SIENNA:** Sounds much easier? So what’s not so good?

**SEIICHI:** Well, there’s the same hallucination problem we talked about for NLG. And it may be harder to manage the interplay between the language and tasks being done.

**DAVID:** Another big issue is that engaging in dialogue isn’t really a function from inputs to outputs. Sometimes there is no output (response) to a given input utterance (and context). And sometimes participants say things that are not related to or prompted by prior utterances.

**SIENA:** Aren’t you getting hungry? I hear humans get hungry several times a day!

**DAVID:** Why do you ask?

**SILA:** Well, we were talking about restaurants earlier, and it got me wondering.

**SIENNA:** Anyway, ... are there other types of dialogue architecture?

**SEIICHI:** There are also some more complex architectures that have multiple components to do different aspects of some of the components shown here. Also other modules and architectural considerations for looking at other phenomena than spoken dialogue, such as movement, multimodal input, action, and cognition. Chapter 16 [Kopp and Hassan 2022] has a lot more about multimodal architectures for SIAs, as does Chapter 20 [Hartholt and Mozgai 2022], which talks about available tools to build on.

**SIENNA:** Cool, so this is like a preview?

**DAVID:** Figure 15.2 shows a SIA architecture we used for some advanced Virtual Humans. [Rickel et al. 2002, Hill et al. 2003, Traum et al. 2003, 2008a, Plüss et al. 2011]. It included SIAs (called virtual humans) in a game engine visual environment, projected lifesized into the space that human participants were in, included multimodal processing for both inputs and outputs, and an integrated cognitive component that did dialogue management, emotion reasoning, task planning and other reasoning, and reasoning about the body's state and how it could be used to express emotions and other information.

**SILA:** Chapter 10 [Broekens 2021] has a lot about SIAs and Emotion.

**SEIICHI:** Chapters 11 [Paiva et al. 2021] and 12 [Gratch and Lucas 2021] are also very relevant, looking at empathy and rapport between humans and SIAs in a dialogue context, talking about how different behaviors can establish and result from rapport in dialogue.

**SILA:** And more on gesture generation and nonverbal behavior can be found in Chapters 7 [Saund and Marsella 2021] and 8 [Pelachaud et al. 2021].

**SIENNA:** How and why are the parts of the integrated cognitive component more tightly integrated?
DAVID: Well, any components could send messages that others could subscribe to, but the DM, emotion, task planner, and body manager were all implemented in the SOAR cognitive architecture [Laird et al. 1987] and could directly access each other’s knowledge structures. This turned out to be a big advantage because the task model covered what the agent was actually doing in terms of individual and team tasks and how far along the way they were and what was still to be done. This was used by the DM to understand language references to aspects of the task. The task model was also used by the emotion model for appraisal of the situation [Marsella and Gratch 2009], to update emotions [Gratch and Marsella 2004], and to come up with appropriate coping strategies [Marsella and Gratch 2003]. It helped the body management with issues like where to look [Lee et al. 2007]. Moreover, the emotion model could also appraise aspects of the dialogue state, and nonverbal communicative functions could be planned that would reinforce and complement verbal communication plans returned by the NLG. Some task states can be updated through dialogue information and negotiation [Traum et al. 2003]. The Dialogue manager also used information from the emotion model, such as to focus on the aspects of the situation that evoked the strongest appraisal when understanding vague or ambiguous language or computing a response [Traum et al. 2004]. The agents could also use the emotion model to talk about their emotions [Muller et al. 2004] and explain where they came from.
SIENNA: Did you have a name for this architecture?
DAVID: Well, we called it the Austin Architecture, though we didn’t use this name in many publications (except this one—Gratch and Marsella [2004]).
SIENNA: Where did that name come from, were you working on it in Texas?
DAVID: No, though Rickel got his Ph.D. there, and Gratch also lived there. It wasn’t an acronym either. Rickel and Johnson’s earlier SIA architecture was called Steve [Rickel and Johnson 1999]. Ours was an extension of that.
SEIICHI: Ah I see—named after Stephen Austin, the father of Texas!
SILA: Back in Texas? No, it must be named after “Stone Cold” Steve Austin, who was a wrestler with attitude?
SIENNA: Wasn’t he from Texas too? Someone told me “Don’t mess with Texas.”
DAVID: Both good guesses but both wrong. Actually, it comes from Steve Austin, the Six Million Dollar Man.
SILA: “A man barely alive?”
DAVID: Yeah, that’s it! Well, the SIAs we built weren’t alive at all, but we did have technology to build highly capable ones.

SIENNA: What are some dialogue activities that SIAs have participated in and in what roles?
SILA: Well, Part V of the Handbook is a great place to start. Chapter 21 [Lane and Schroeder 2022] talks about pedagogical agents, which is close to my heart—SIAs can be tutors or coaches to provide information and help you understand, or collaborative peers that work with you (more on this in Chapter 22 [Cassell 2022]), or learners that you instruct (and sometimes learn more than they do while doing it).
SEIICHI: SIAs can help people that don’t get enough help from other people. Chapter 25 [Nadel et al. 2022] reviews work to support people on the autism spectrum. Chapter 23 [Ghafurian et al. 2022] talks about SIAs for supporting aging.
SILA: I get older all the time, without any help!
SEIICHI: And Chapter 24 [Bickmore 2022] talks more generally about health applications.
SIENNA: That sounds very serious! Aren’t there some applications just for fun?
SILA: Have I got a chapter for you! Chapter 27 [Prada and Rato 2022] is all about SIAs and games. SIAs could be teammates, opponents, or neutral in the game. And they can be different kinds of teammates, including companions, assistants, or helpers.
SEIICHI: Games can be serious too! Just look at Chapter 28 [Gebhard et al. 2022]. Gamification can be used as an incentive for all the previous applications like education and health, as well as others.
SIENNA: Sometimes you just want a good story.
SILA: Then look no further than Chapter 26 [Aylett 2022]! SIAs can tell stories, but also you can interact with them and become part of the narrative.

SEIICHI: All of these applications have a role for SIA dialogue. For some activities, it's essential, though some might get by with other kinds of interaction.

SILA: I don't talk when I play chess, it's considered rude.

SIENNA: Since when has that stopped you?

SEIICHI: David, you've been quiet for a while. What kind of SIAs have you worked on, and what activities and roles did they participate in?

DAVID: Well, we've talked about some of them already. Here's a slide I sometimes use (Figure 15.3) to show some of the SIAs that we've worked on dialogue capabilities for. They're very different in many ways, including the activities and roles, dialogue architectures, and nature of the dialogue, but they all talk with people (and sometimes each other).

SILA: Wow, a whole SIA-vilization! Tell us about some of them.

DAVID: Well, in the top left, you see the Mission Rehearsal Exercise system, that we talked about before; we had two different agents that used the AUSTIN architecture, as well as a bunch of simpler ones who acted as squad leaders and townspeople and other teammates on the radio. The Sergeant multicommunicated with the human playing the lieutenant, and with other squad members. The Medic, Tucci, focused

![Figure 15.3 ICT SIAs with example activity types.](image-url)
on an injured boy's health, but did engage in the main conversation with Sergeant and Lieutenant when needed. You also saw an example from SASO in Figure 15.1 with Doctor Perez. This was later extended in Traum et al. [2008a] to include Elder al Hassan. These agents used the same AUSTIN architecture as the MRE agents, but instead of being teammates they had their own independent goals and wouldn't always be cooperative. You had to negotiate with them to try to reach a solution that worked for everyone.

SIENNA: What were the group conversation characters?
DAVID: That was primarily the work of my Ph.D. student Dusan Jan, creating virtual extras who could work as background characters in big scenarios like the first two, where you need groups to plausibly be involved in conversation [Jan and Traum 2007]. It used a model of group conversation, based on Padilha and Carletta [2002], but extended for SIAs, including nonverbal as well as spoken behavior. Interestingly, there was no actual content, but they went through the motions of dialogue, including modeling participation and turn-taking structures. Later, we extended this [Jan et al. 2007] to cover cultural differences in proxemics, gaze, and turn-taking.

SILA: There's a lot more on culture and SIAs in Chapter 13 [Lugrin and Rehm 2021], including some dialogue implications.
DAVID: Several of the characters, Kemp, Raed, Omar, SGTs Star and Blackwell used a simpler setting and architecture, with a single SIA responding to a human. Our NPCEditor [Leuski and Traum 2011] does support a kind of “end-to-end” dialogue flow, though using relevance models rather than deep neural nets.

SIENNA: Why didn't you use deep neural nets?
DAVID: Well, first, they weren't very good yet when we did a lot of this work. Also, with moderate-sized training sets, the NPCEditor still seems to do better than some of these models for a real application [Alavi et al. 2020]. The ELITE project had several scenarios where you could practice leadership interactions with a SIA subordinate. Our C3IT project focused on cultural training by interviewing various witnesses (some of whom are suspects). SGT Blackwell talked about the Army and his technology, originally at the Army Science Conference. SGT Star was used by Army recruiters to answer questions about Army job opportunities.

SIENNA: Weren't some of them in museums?
SILA: SGT Blackwell was in the Cooper-Hewitt Museum in New York’s best design in America Triennial [Robinson et al. 2008], and the Twins were in the Boston Museum of Science [Traum et al. 2012]. They're very famous SIAs.
DAVID: I think they've talked to more people than I have! The Twins also used the NPCEditor, but the response was actually a subdialogue between the two Twins in most cases.
SEIICHI: So that’s an example where there are two different character bodies as animators but a single component as author.

DAVID: Well, maybe even a little more complicated because the texts and animations were pre-authored for those systems by humans, but the decision of which one to use when was made in real-time by the NPCEditor, based on visitor utterance and some context representation. The Gunslinger system was similar, though it was embedded in a physical space and also included visual information to help decide which character was the addressee. That was meant as an entertainment vehicle, where you could interact with folks from the Old West, similar to Westworld.

SIENA: That Rio Lane is awful mean! He told me to shut up!

SIENNA: At least he didn’t shoot you.

DAVID: Gunslinger inspired the CHAOS project [Aggarwal et al. 2011], where the main character Omar spoke only Pashto.

SIENA: Didn’t some of these characters use different architectures for different situations?

DAVID: Yes, that’s right! We had three different versions of Hassan, a character meant to practice interview skills [Traum et al. 2008b]. The later versions allowed more complex interactions, including an affective model that would influence the kind of responses you got for the same question, and an ability to handle bargaining and subdialogues. Also, some of the gunslinger characters were later used in a follow-on scenario using the AUSTIN architecture [Plüss et al. 2011].

SIENA: What about Robots?

DAVID: More recently, we’ve been working with Robots, for example, the ScoutBot project [Lukin et al. 2018], where someone could carry on a dialogue with a distant robot to direct it to explore an unfamiliar space. We also used the NPCEditor for this, which worked surprisingly well, given obvious limitations of a classification approach for numerical data [Gervits et al. 2021]. It also had a bridge between our virtual human toolkit architecture [Hartholt et al. 2013] and the Robot Operating System (ROS) [Quigley et al. 2009]. We’ve also implemented a system for playing a word guessing game with a Nao robot as well as virtual humans [Pincus 2020].

And we’ve had a couple scenarios that included both a robot and a virtual human—one was a three-way discussion about a ranking task [Artstein et al. 2018], and another was meant to introduce babies to some of the basic phonological units of sign language [Nasihat Gilani et al. 2018, Scassellati et al. 2018].

SIILA: I know that work—the robot would capture the baby’s attention in the physical 3D world and direct gaze to Alissa who had finer motor control to deliver a fluent nursery rhyme.
DAVID: We also had several systems for mental health, including virtual patients that clinicians could train with: SimCoach, an information provider about care options [Rizzo et al. 2013], and SimSensei, to detect distress [DeVault et al. 2014]. Also, a prototype of a system to counsel middle-schoolers who are being bullied [Gordon et al. 2019].

SILA: Sometimes it seems that SIAs can do anything!

DAVID: Well, some things might be too hard now, but we keep working on more types of activities and roles that could be helpful or interesting.

SIENNA: What is the best way to evaluate SIA Dialogue?

SILA: There isn’t one. They’re all bad.

DAVID: Well, I wouldn’t go that far! There’s no perfect metric yet, but it’s good that people are trying.

SEIICHI: BLEU [Papineni et al. 2002] is a very popular automated metric—lots of people use that to evaluate their dialogue systems. It’s found to be fairly good for machine translation, and it and similar similarity metrics like ROUGE [Lin 2004] make a certain amount of sense—if something is close to a real dialogue then it’s probably better than something random.

SILA: Well, a little bit of sense is not very sensible. Liu et al. [2016] literally wrote a paper about this called, “How not to evaluate your dialogue system.”

SEIICHI: There are some new metrics coming out too, like GRADE [Huang et al. 2020] and FED [Mehri and Eskenazi 2020] that use datasets like neural nets trained on large corpora and topic graphs. Some of them seem pretty good [Yeh et al. 2021].

SILA: Time will tell.

DAVID: If the SIA is engaging in dialogue to help accomplish a task or help someone learn something or change their behavior, then we can evaluate the outcomes and perhaps compare to a system without dialogue or with a different kind of dialogue model.

SEIICHI: Another popular way is to use some kind of survey of participants or observers about whether the whole dialogue was satisfactory, or about specific aspects of it. SASSI is one popular survey [Hone and Graham 2000].

SILA: That’s better but still very sensitive to the survey population. Like are they real users or paid experimental participants [Ai et al. 2007], or are they users or just observers, who sometimes rate the same dialogues differently [Georgila et al. 2020a].

SEIICHI: Well, there are a whole host of measures and methods to choose from [Finch and Choi 2020].
SIENNA: What are factors that make SIA dialogue different from other kinds of dialogue systems?

SILA: Because they have SIAs in them, of course!

SEIICHI: SIAs have some kind of embodiment and inhabit some kind of environment. So any time where you need to talk about those specific aspects like what you can see from a specific perspective, or manipulation of objects, or locomotion, there will be some differences.

SILA: “Locomotion,” is that like crazy motion? I learned some Spanish!

SIENNA: Just enough to be dangerously wrong.

SEIICHI: Also, nonverbal behavior can be an important part of dialogue, including proxemics, gaze, head and hand gestures, etcetera. So dialogue will be different if you can have those. See Chapter 14 [Vilhjálmsson 2022] for more on proxemics and Chapters 7 [Saund and Marsella 2021] and 8 [Pelachaud et al. 2021] for more on conversationally relevant nonverbal behavior.

DAVID: Some things are more a matter of degree than a firm distinction between SIAs and non-embodied dialogue systems. For example, making an emotional connection or having an engaging experience is possible with a text-only or voice-only system, but it’s a lot easier with an embodied system that is giving appropriate nonverbal cues and providing a greater sense of presence.

SIENNA: How is Virtual Human dialogue different from Social Robot dialogue?

SILA: We’re talking physical vs virtual!

SEIICHI: It’s a lot harder to do proxemics with people if you don’t have a 3D body, unless you’re in virtual reality.

SILA: Also, virtual humans can only pick up and manipulate things in the virtual world, while robots can manipulate the same objects and hand them to or take them from people.

SEIICHI: So dialogue that gets responded to with action or that accompanies physical interaction is going to be a bit different.

DAVID: Another thing is where the source of problems comes from and how to fix them. Robots’ physical parts sometimes break down or lose power, or they are noisy or slow, which can interfere with fluent conversation. On the other hand, virtual humans are not constrained by the physical world and real physics, so they might do things that are weird or just not believable, which also impacts the trajectory of the dialogue.

SILA: Sometimes if I’m not careful, I say two things at the same time, but it sounds weird, and my mouth looks crazy.

SIENNA: Is that different from usual?

SILA: So I’ve been told.
SIENNA: What are some ways we can build our own SIA dialogue applications?

SILA: There are some tools available for building dialogue systems, some of which have been used for SIAs.

SEIICHI: The virtual human toolkit was mentioned before—that lets you use the NPCEditor [Leuski and Traum 2011] for dialogue and NLU, and connect to game engines and smartbody embodiment.

DAVID: And also any ROS application, as mentioned above.

SILA: There is also a good section on natural language processing tools in Chapter 20 [Hartholt and Mozgai 2022], Section 6.3, which mentioned a lot of good available tools, including OpenDial [Lison and Kennington 2016], PyDial [Ultes et al. 2017], and RASA [Bocklisch et al. 2017].

DAVID: We've also been maintaining a list of dialogue tools with some information on how they can be used here: https://dialport.ict.usc.edu/index.php/resources/.

SILA: This gets me thinking, if SIAs could build their own SIAs, could they maybe build themselves?

SIENNA: Which came first, the SIA or the SIA that built the SIA?

SIENNA: What are current challenges for SIA dialogue?

SILA: I think we already talked about a lot of them, like architecture, and multis, and evaluation.

DAVID: Another hot topic right now is vision and language. Some of the same models are being used to be able to understand and generate both visual scenes and language describing those scenes.

SEIICHI: Some of this has been around for a while, for example, Steels [1998] and Roy [2002], but there's a lot more going on recently, for example, Yu et al. [2017], Manuvinakurike et al. [2018], Anderson et al. [2018], Pezzelle and Fernandez [2019], Kamezawa et al. [2020], and Dong et al. [2021].

DAVID: There are more available databases of either images or videos [Shridhar et al. 2020], along with captions or descriptions of actions. Systems are getting better at answering questions about what's in an image, or describing what's seen, or constructing a plan to achieve an instruction.

SILA: I’m going to use that to solve captchas and prove I’m not a robot!

SIENNA: But are you a robot?

SILA: If you don’t know by now, I won’t tell.

SEIICHI: Ethics is another important topic for SIAs. Chapter 2 [Rosenthal-von der Putten and Abrams 2021] talks about ethics of research studies, and Chapter 3 [Krämer and Manzeschke 2021] talks about ethical implications of reactions to SIAs, but what about ethics related to dialogue?
DAVID: Well SIAs can use dialogue to motivate people to do good things, such as helping someone quit smoking [Grolemann et al. 2006] or exercise more [Bickmore et al. 2005, Fasola and Matarić 2013], but the same techniques might be used to sway people toward more harmful behavior.

SILA: With great power comes great responsibility!

SIENNA: But who is responsible? Is it the SIA or someone else?

DAVID: There are another set of issues, such as whether people talking with SIAs makes it harder to talk with and relate to other people [Turkle 2017], or if instead of talking to SIAs like we talk to people, we talk to people like we talk to SIAs [Burton and Gaskin 2019].

SILA: Some people aren’t so polite to SIAs [Cercas Curry et al. 2021].

SIENNA: What are the Future Prospects for SIA dialogue?

SILA: “The future’s so bright, I gotta wear shades!”

SEIICHI: Well, here are some recommendations for future research on spoken interaction with robots [Marge et al. 2022].

DAVID: Most of them apply to SIAs more generally. There are a lot of challenges still, particularly for fully real-time, complex interaction, adaptation to new tasks, user requirements, and desires, and understanding each SIA component well enough to talk about it.

SEIICHI: Well, I think we’re about out of time.

SILA: Or Space, as the case may be.

SIENNA: I want to thank you all for answering all of my questions. Unfortunately, answering questions is like cutting off the head of a hydra—two more grow back to take its place. So we’ll have to do this again sometime.

SILA: Or maybe we’ll have to do it twice?

DAVID: I sense an exponential progression happening....

SILA: Thanks also from me, it’s been fun. I love sharing the breadth of my knowledge and any opportunity to make light of a serious situation.

DAVID: Yeah, the last time I was on a panel like this it was with other humans. Always refreshing to look at things from a different point of view.

SEIICHI: Well, then let’s end here, and see ya all later!

DAVID: Bye.

SILA: Au revoir.

SIENNA: Why do some folks say goodbye in different languages?

SILA: Aren’t all languages different?

SIENNA: I mean different from the one they were talking in the whole conversation.

SEIICHI: Maybe it’s all they know in that language?

---

2. Title of song from Timbuk 3, 1986.
SILA: Well, it wasn’t goodbye exactly, since I hope we meet again.
SILA: Closings are complicated [Schegloff and Sacks 1973].
SIENNA: Another conversation that failed to end.
SILA: Maybe if we all just be quiet for a while it will go away?
SIENNA: One can always hope....

Acknowledgements
Thanks to Birgit Lugrin and Catherine Pelachaud for encouragement, support and feedback on earlier versions of this chapter. The author was supported by the Army Research Office under Cooperative Agreement Number W911NF-20-2-005, and by the National Science Foundation grant 1925576. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References


Chapter 15 Socially Interactive Agent Dialogue

Computer Vision and Pattern Recognition, IEEE, 3674–3683. DOI: https://doi.org/10.1109/CVPR.2018.00387.


References


Chapter 15 Socially Interactive Agent Dialogue


Socially Interactive Agent Dialogue


Chapter 15 Socially Interactive Agent Dialogue


The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures

Stefan Kopp and Teena Hassan

16.1 Motivation

The field of socially interactive agents has emerged out of many different approaches to create technical systems that can engage in natural human-like conversation, nonverbal communication, multimodal dialog, or emotionally aware interaction. Many of these directions have grown into research fields in their own right (e.g., Social Signal Processing, Affective Computing, Social Robotics, Natural Language Processing, Spoken Dialog Systems/Conversational Agents), with specific foci, methodological approaches, and technologies. As a result, the variety but also the specialization of approaches has been growing. For example, we now have elaborated methods for the recognition or synthesis of social signals in specific modalities (e.g., speech, prosody, facial expressions, gaze, or gesture), their fusion to extract semantic or pragmatic meaning, or the planning of interactive behavior to fulfill emotional or relational goals. This gives rise to an integration problem as, in social face-to-face interaction, many of these abilities and skills need to be at work at the same time and in an integrated and coordinated manner [Gratch et al. 2002]. A crucial question is thus not only how the various capabilities and features of a Socially Interactive Agent (SIA) can be realized technically but also how they can come to play together within a functionally complete interactive virtual agent or social robot.
The present chapter focuses on concepts and methods to realize and integrate approaches to achieve abilities needed for conversational multimodal interaction. We will thereby go beyond the usually separate perspectives toward (and solutions for) processing multimodal input or generating multimodal output. Instead, we aim to provide an overview of how such techniques are mapped out and integrated in current virtual agents or social robots by means of a suitable interaction architecture. From a practical point of view, an agent architecture may be regarded as a collection of specialized modules linked together by means of inter-process communication. From a conceptual point of view, however, it has to answer the question of how the underlying complex computations that are required to act like a socially intelligent agent in real-time interaction can be organized and orchestrated. It thus provides the underlying structure that provides the constraints and affordances both for how single modules are to operate as well as for how the agent as a whole is able to (inter-)act in a consistent, timely, and believable manner, and how it will thus appear as an interaction partner.

Knowing about the challenges, principles, and approaches for developing multimodal interaction architectures of SIAs has become increasingly important for researchers and practitioners alike. Interacting with today’s social robots or virtual agents is often characterized by stereotypical behavior or slow response times, resulting in unnatural clumsiness or disfluencies. Human conversational interaction, in contrast, is characterized by an inherent multimodality and high responsiveness with which cooperative interactants construct their contributions. For example, even while producing communicative actions, speakers attend to and elicit reactions from their addressee [Clark and Krych 2004]. Depending on this immediate feedback, speakers can re-plan the remaining part(s) of their communicative act, adapt it to the addressees’ needs, put it on hold, interject a sub-dialog, and continue at the point of interruption. All of this is done in such an effortless, smoothly coordinated, and seemingly natural way that it is not even apparent that difficulties were paid attention to or that plans were changed midway. Thus, acting in a conversation is not solely based on extensive planning ahead and deep representational models. Instead, interaction partners, while being guided by overall goals and strategies, are also highly sensitive to the partner’s verbal and nonverbal behavior and are able to alter their multimodal utterances accordingly. These abilities are crucial for human-like fluent conversation—and they imply important demands for how to construct interactive agents at the architectural level.

In the following, we will start by identifying overarching requirements and criteria for multimodal interaction architectures. We then review different approaches and concepts that have been put forward in the fields of (embodied) conversational agents and social robotics. Finally, we will point out main challenges and directions
to be pursued in order to succeed in weaving the fabric of truly socially interactive and intelligent agents.

16.1.1 Requirements for Multimodal Interaction

Architectures of social agents or robots are usually designed with a particular functional goal in mind, such as joint attention, empathy, imitation, or interactive learning [Breazeal et al. 2004, Duffy et al. 2005]. In this chapter we discuss how the specific components and layout of SIA architectures enable (or hamper) multimodal conversational interaction with a human user. We start by identifying a number of requirements that a SIA needs to meet in order to provide multimodal interactivity to its user.

One obvious requirement is the ability to recognize the relevant verbal and nonverbal input as well as to generate convincing multimodal output. A main distinguishing aspect is thus the number and kinds of modalities supported when interacting with the agent. Most virtual interactive agents and social robots have included visual and auditory sensory modalities (e.g., [Dodd and Gutierrez 2005, Kasap and Magnenat-Thalmann 2010, Baxter et al. 2013, Kędzierski et al. 2013, Matsuyama et al. 2016]). In addition, few virtual agents [Bosse et al. 2018] but several social robots (e.g., Breazeal et al. [2004]; Pepper [SoftBank-Robotics 2021b], Paro [Shibata 2012]) support tactile stimuli to perceive or produce touch. However, a modality goes beyond a mere sensory channel and must be considered a semiotic system that affords certain semantic and pragmatic functions by means of specific displays or signals conveyed over a certain sensory channel. For example, spoken language as a vocal modality allows for conveying symbolic content, intonation can add “analog” acoustic cues of prominence or stance, and gesture as a visual modality lends itself to communicate indexical or iconic meaning. Artificial agents may even add other non-human modalities to this. Multimodal communication, then, arises from the combination and integration of those different ways of communicating meaning. It thus involves not only the processing and generation of single verbal and nonverbal behaviors but also their interpretation and embedding in coherent multimodal ensembles whose parts are coordinated in form, meaning, or pragmatic function as well as in their temporal arrangement. The corresponding multimodal coherence and cross-modal relations are vital for a recipient to be able to resolve the overall intended meaning.

A second requirement, pointed out by Cassell et al. [2000] in their seminal work on embodied conversational agent frameworks, is to be able to deal with behavior in terms of its multiple conversational functions (e.g., conveying content, representing socially, managing conversation) and based on an understanding of a dialog state that can involve multiple threads of communication. This relates to the
The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures

...grounding of multimodal behavior processing into models and representations of (changes of) an interaction state and the selection of multimodal behavior in order to change this state according to interaction goals or policies. A related requirement is that multimodal behaviors need to support a sufficient degree of expressiveness that is needed for the communicative demands and believability of the human–agent interaction at hand.

Thirdly, multimodal interactions unfold at multiple timescales, from milliseconds between eye-contact and a head nod to longer periods of time for utterances or even larger discourse segments [Cassell et al. 2000]. Across these timescales, multimodal conversational behavior must be sufficiently fluent and continuous. Unwanted and unnatural lags, hesitations, or disfluencies can lead to interaction problems (e.g., overlapping speech with dialog systems) as well as ambiguous, incoherent meaning (e.g., when pointing to an object too late). Multimodal SIAs thus need to be able to manage multimodal behavior in real time, at multiple timescales in parallel and with the corresponding fluency. Further, it is often emphasized [Schlangen and Skantze 2011, Kopp et al. 2014] that fluid conversation hinges upon fast and reciprocal adaptation between the interlocutors. For example, in a multimodal interaction one often has to adapt one’s own behavior to the interlocutor’s actions in an online and well-timed fashion, for example, to keep or take the floor [Levinson and Torreira 2015], to respond to communicative feedback and interruptions, or to entrain and align with one another [Lakin et al. 2003]. Consequently, another requirement for SIAs is fast responsiveness, adaptability, and interruptability in their multimodal behavior.

16.2 Models and Approaches

A large variety of SIAs have been designed to support some form of multimodal interaction. In this vein, different architectural approaches to organize and realize the processing of (multi-)modal input and output have been employed in IVA or SR. Yet, existing systems fulfill the abovementioned behavioral requirements only to a partial and different degree. In the following, we will discuss the architectures that have been developed in relevant fields. They can be compared and assessed with respect to a number of features:

— Modalities: What is the number and kinds of modalities included, and what is the degree of multimodal integration (fusion/fission)?
— Methods/components: What are the techniques and approaches used for recognition/interpretation, generation, and planning of multimodal behavior (at the task level as well as the social-relational level)?
— Processing structure: How is control and processing organized across different routes (pathways, streams) or at different levels (deliberative, reactive, associative)? What is the general way of processing input/output over time (sequential/parallel, chunked/incremental)?

— Interactive adaptivity: Is the social interaction dynamics with its reciprocal feedback loops taken into account? How are processing and generation connected to support fast adaptivity? What kind of cross-modal interactions are considered?

— Technical applicability: Is the approach specific to virtual or robotic agents? How modular, interoperable and portable is the approach?

In order to provide a systematic overview and to make approaches comparable, we will characterize them according to which parts of a conceptually “complete model” of multimodal interaction they support. A schematic of this conceptual model is shown in Figure 16.1. It comprises three columns for (1) processing multimodal input, (2) mapping responses, and (3) generating multimodal output. Each column, in turn, comprises different levels of processing, from sensory-motor behavior to high-level conversational and socio-relational functions. For the input column (left-hand side), this relates to the common processing pipeline from sensing data, to recognizing features or patterns, to interpreting them with regard to meaning or interactional functions. For the output column (right-hand side), the stages correspond to a standard generation pipeline [Kopp et al. 2006] that involves determining modalities and behavioral forms (e.g., words, intonation, gestures, expressions) to fulfill a given intent, turning them into actual synchronized behavior with the bodily resources of the agent, and, finally, acting them out out.

Figure 16.1 Schematic of different processes and pathways in multimodal interaction architectures.
overtly. The middle column maps between input and output at different levels of
decision-making processes, from reactive (based on hard-wired rules), to associ-
ative (selecting from a given set of alternatives), to deliberative (planning possibly
new responses).

It is important to note that different pathways of processing are possible
through the columns and layers (and are actually taken in existing systems). For
example, multimodal signals can be processed up to interpreting a user state that
is then mapped via associations to pre-defined outputs (circumventing planning
of goals, content, or forms). Note also that each of these boxes can be more or less
modality-specific or multimodal. For example, recognition can work on different
modalities separately with specific models whose outputs are combined afterwards
(so-called “late fusion”), or it can work on multimodal data (after an “early fusion”)
to find larger, integrated patterns or features.

In the following, we will discuss different multimodal interaction architectures
that have been applied in SR and IVA. We will thereby characterize them and
make them directly comparable by mapping their architectural components to the
schematic shown in Figure 16.1, using the same color code to relate specific parts
of the architecture.

16.2.1 Embodied Conversational or Virtual Agent Architectures
Virtual agents or embodied conversational agents (ECA) are graphically rendered
characters designed to support a human-like conversational interaction with a
human user [Cassell 2001]. A large number of such agents have been developed,
 focusing on different kinds of socio-communicative behavior, abilities, or applica-
tion scenarios. Throughout this endeavor, a range of architectural principles and
models have emerged.

16.2.1.1 Single-route Architectures
Many ECA systems have focused on producing socially appropriate multimodal
behavior to achieve, for example, engagement, rapport, trust, or empathy. The
architectural layout consists of a single route with multiple consecutive processing
steps, usually involving high-level state representations and planning-based behav-
ior generation. For example, so-called “relational agents” employ specific planners
to increase and maintain rapport with the user [Papangelis et al. 2014] or to achieve
long-term engagement [Bickmore et al. 2010]. One example is the proposed SARA
architecture [Matsuyama et al. 2016] (Figure 16.2), in which a task planner and
a social reasoning component are combined with a memory-based model of the
user and previous interactions, as well as with corresponding modules for social
behavior interpretation (here, estimating rapport from utterances, conversational
strategies, acoustic features, and 3D facial landmarks) or behavior generation (natural language generation, nonverbal behavior generation). Together, these modules form a single deliberative route of multimodal processing along which the agent produces socially attuned behavioral responses to complete user inputs. The focus in these systems lies on the functional quality of the produced behavior and less on its embedding in a fluent and dynamic conversational interaction.

### 16.2.1.2 Dual-route Architectures

Complete ECA systems aim to enable an efficient and robust face-to-face conversational interaction (e.g., Rea [Cassell 2001], Max [Leßmann et al. 2008], or Greta [Bevacqua et al. 2010]). These systems employ a dual-route architectural layout (see Figure 16.3). As described above, the deliberative route comprises higher-level processes for reasoning and planning of desired interactive functions and behaviors. This is usually based on classical natural language processing pipelines in Spoken Dialog Systems, which include some form of semantic decoding and dialog state tracking, based on which the system output is determined through some form of (pre-)planned policy. ECAs employ a similar pipeline but extend it to interpreting and planning multimodal communicative behavior (e.g., gaze, gesture, body posture, facial expression) for conversational or socio-relational functions (e.g., dialog grounding, turn-taking, attention, politeness, empathy). The underlying models are based on dedicated representational and decision-making models, either symbolic and rule-based (classically) or implicit and learned from data (more recently). A second reactive route, in contrast, implements more direct mappings from perceptual events to overt behavior. This route is required to support fast social feedback loops, for example, in continuous gaze-tracking or behavioral mimicry.
Chapter 16  The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures

Figure 16.3  General structure of a dual-route architecture with parallel deliberative and reactive processing of multimodal behavior (Inset shows three conversational agents based on this layout: REA [Cassell et al. 2000]; Greta [Bevacqua et al. 2010] (2021 Catherine Pelachaud); and Max [Leßmann et al. 2008] (© 2021 Stefan Kopp)).

(by mapping user location or user movement to animated adjustments of the agent body). It is also necessary for an agent’s subtle and dynamic expressiveness, for example, through emotional facial expressions (by mapping internal affective states to animated facial features). Both routes rest on a behavior realization mechanism that is in charge of arbitrating, combining, synchronizing, and, finally, producing the eventual output behaviors.

16.2.1.3 Multidirectional, Incremental Architectures

Many of the existing ECAs build on the dual-route architecture layout, characterized by concurrent processing along a deliberative and a reactive route. However, each route itself realizes a sequential processing of input/output units at a corresponding level of abstraction and granularity. The growing awareness of the role of multimodal behavior in the dynamic grounding of dialog (see Section 16.1.1), however, and the view that ECAs ultimately need to be able to support these mechanisms for collaboratively co-constructing mutual understanding [Kopp and Krämer 2021], has led to further advancements at the architectural level. They can be summarized under two key concepts: Multidirectional flow of information and incremental processing.

Incremental processing has been frequently identified as a key principle for natural dialog modeling with phenomena such as fluent turn transitions, interruptions, disfluencies, or fast adaptations to the interlocutor. For example, the “How Was Your Day?” prototype [Crook et al. 2012] for coping with barge-ins employed a “long” loop for intent planning and a shorter loop to handle interruptions, back-channel feedback, and emotional mirroring. However, the authors note that the use of incrementality would have made their design more elegant and efficient.
Schlangen and Skantze [2011] described incremental dialog agents in terms of abstract modules that communicate via incremental units that are extended in a stepwise fashion before being finally committed. Several implementations of this model have been developed, and many aspects of language-based dialog have been successfully modeled within this incremental processing framework (e.g., speech recognition, natural language understanding [Atterer et al. 2009], dialog management [Traum et al. 2012], natural language generation [Skantze and Hjalmarsson 2013], speech synthesis [Buschmeier et al. 2012]). A few recent approaches have tried to apply this principle to process and integrate multimodal input, for example, speech and gesture [Han et al. 2018].

The second extension refers to enabling a multidirectional flow of information. This includes the passing of feedback information from downstream components back to higher-level modules. The SAIBA (Situation, Agent, Intention, Behavior, Animation) framework for multimodal generation [Kopp et al. 2006, Vilhjálmsson et al. 2007] provided markup languages for specifying multimodal behavior (BML) and its functions (FML) to be processed by subsequent modules. Additionally, it emphasized the importance of feedback to inform planning modules about the extent to which their decisions were being realized. Further, a multidirectional flow of information also refers to information flowing internally from generation to processing (i.e., from right to left in our schema). It has often been stressed that in the brain top–down information helps to bias or prime sensory processing toward specific information or to resolve ambiguities based on contextual information (cf. [Teufel and Nanay 2017]). Yet, very few SIA architectures have modeled input processing components to receive information from higher-level components of the architecture. Nijholt et al. [2008] proposed a first approach to directly link the timing of an agent's behavior to the predicted timing of interlocutor events. This enabled a finer degree of temporal coordination with the user's motion, as demonstrated in a dancer agent, a virtual orchestra conductor, and a virtual fitness trainer. The Artificial Social Agent Platform (ASAP) [Kopp et al. 2014, van Welbergen et al. 2014] proposed extensions to the behavior markup language (BML) in order to bidirectionally link sensory input processing and generation of agent's behavior.

Overall, ECAs designed for fluid interaction resolve the strict layout of sequential processing architectures in favor of a more flexible distribution of processing, both with respect to the flow of information as well its temporal organization. As illustrated in Figure 16.4, incremental processing is applied in particular at higher levels of the architecture, where units of processing (e.g., a full dialog act) are more abstract, arise at a relatively lower rate, and have a larger temporal scope in the underlying overt behavior.
16.2.1.4 Behavior Generation Sub-architectures

Work in the field of ECAs or virtual agents has traditionally focused on the generation of expressive, communicative multimodal behavior. Consequently, a lot of systems and models have been developed to embody the right-hand side of our architectural schema. Two main approaches have emerged that can be distinguished according to what they start out from. On the one hand, classical approaches to multimodal behavior generation take some form of communicative intent as input and map it to multimodal ensembles out of several, mutually coordinated behaviors. On the other hand, an abundance of recent work has approached the problem of generating multimodal behaviors by starting out from some already given behavior and augmenting it with additional behaviors in other modalities. Such a cross-modal mapping approach is, for the most part, driven by speech as input modality. We will discuss both approaches in the following, also noting how they have been combined.

**Intent-based multimodal behavior generation** is generally conceived to comprise a number of processing steps, similar to the output generation branch of dialog systems. This view has been formalized within the SAIBA framework to encompass three main stages, corresponding to (1) intent planning, (2) behavior planning, and (3) behavior realization [Kopp et al. 2006], along with two XML-based specification languages as interfaces between them (Function Markup Language [FML] and BML). Generally, these stages relate to the subsequent decision steps of determining what to communicate, with which behavioral forms, and, finally, how to do it overtly. Intent-based generation, for example, may start from a speech act representation plus some emotional state or socio-relational goals (stated in FML). Behavior planning then usually involves natural language generation along with the composition/selection and coordination of appropriate gestures, gaze,
or facial expressions. This problem has been tackled using rule/lexicon-based, planning-based, or data/learning-based approaches (see also Chapter 8 on “Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]; cf. Kopp [2013]), depending on the requirements and criteria for the targeted behavior (e.g., realism, expressiveness, design effort, real-time capability, cognitive plausibility). Behavior realization is then in charge of mapping these multimodal behaviors (usually specified in BML) onto temporally synchronized and coherent articulations or movements.

Much work has focused on developing BML-compliant realizers for virtual characters (e.g., ACE [Kopp and Wachsmuth 2004], Greta [Bevacqua et al. 2010], ASAP [van Welbergen et al. 2014], SmartBody [Thiebaux et al. 2008], or EMBR [Heloir and Kipp 2010]), and several approaches were also extended to more flexible timing and motion planning for physical robots [Holroyd and Rich 2012, Salem et al. 2012].

Cross-modal behavior generation recently has become popular due to the direct applicability of Machine Learning methods to large available datasets on human multimodal behavior. The predominant approach is to generate nonverbal behaviors (e.g., gestures, head movements, facial expressions) for a given text or speech output, and the key question is what features are necessary to map from the verbal modality to others in this way. Early speech-driven generation systems applied analysis steps (akin to processing user input) to the linguistic output in order to determine, for example, speech semantics, information structure, discourse relations or emotions. This additional information is then used to select appropriate behaviors by means of empirically grounded but manually defined rules, or data-based mappings [Cassell et al. 2004, Lhommet et al. 2015] (see Figure 16.5; see also Chapter 7 on “Gesture Generation” [Saund and Marsella 2021] of volume 1 of this handbook [Lugrin et al. 2021]). A key challenge here is that higher-level semantic

---

**Figure 16.5** Examples of sub-architectures for cross-modal, speech-driven behavior generation: (left) BEAT (based on Cassell et al. [2004]); (right) Cerebella (based on Lhommet et al. [2015]).
or pragmatic aspects are necessary, especially for generating coherent and communicatively meaningful nonverbal behavior like representational gestures that, for example, depict visual aspects of an object linguistically referred to. These aspects, however, are hard to determine or infer from a semiotically different linguistic input.

Other speech-accompanying behaviors such as small head movements, eyebrow raises or beat gestures, which are less explicitly communicative but nevertheless instrumental for creating a lifelike impression, have been successfully synthesized by mapping directly from the acoustic or verbal features of the speech input. One focus is the speech-driven generation of gesture, for which statistical or, more recently, deep neural network-based models are applied that have been trained to create some form of encoding of the input features and to map it to body postures and movements by way of some generators (usually in an auto-regressive way, i.e., each pose based on the previous one); see Figure 16.6. Note that these models work in an end-to-end fashion. That is, they comprise both behavior planning and realization and they have been used to drive virtual characters as well as humanoid robots. The techniques that have been explored include probabilistic models [Chiu and Marsella 2011, Ishi et al. 2018], bidirectional long short-term memory networks [Hasegawa et al. 2018], generative models [Kucherenko et al. 2020], mixture models [Ahuja et al. 2020], or generative adversarial networks [Yoon et al. 2020]. Recent approaches have succeeded in producing considerably natural and consistent multimodal behavior, with current work starting to explore how more general contextual parameters such as speaker identity or style can help to further increase output quality (see also Chapter 8 on “Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]).

### 16.2.2 Social Robot Architectures

Many architectures, more or less cognitively motivated, have been developed for and implemented in social robots (e.g., [Breazeal et al. 2004, Laird et al. 2012, Baxter et al. 2013, Chao and Thomaz 2013, Trafton et al. 2013, Adam et al. 2016, ...]}
Moulin-Frier et al. 2018, Bono et al. 2020). Interaction between humans and mobile robots involves several complex challenges that are often absent in the interaction between humans and virtual agents. Given the hardware constraints, the unpredictability of physical actions and their outcomes in the real world, and the perception challenges posed by the uncontrolled environment, the architectures developed and tested for virtual agents cannot transfer well to robots, especially to mobile robots that co-inhabit or collaborate with humans in the physical world. In order to deal with the complex, high-dimensional, and dynamic domain of human–robot interaction, novel mechanisms for robust perception–action feedback and flexible handling of contingencies arising from failed actions or delayed or failed perception are required, in addition to the integration of safety and privacy-enhancing measures.

In this section, we briefly discuss some of the key interaction architectures that have been developed and tested on social robots. Their design and the modules they are composed of depend on the specific interaction goals being pursued. We start by discussing the sensing and action modalities considered in these works, the interaction goals pursued by them, and the architectural components designed to realize these goals. Then we will turn, again, to the architectural layouts employed in social robots. Depending on the number and type of parallel routes or pathways supported, these can be categorized into single-route, dual-route, and multidirectional architectures. We will provide examples for each type of architecture and use the schematic shown in Figure 16.1 to highlight the design principles underlying these architectures.

16.2.2.1 Sensing Modalities
Social robots are seldom equipped with only unimodal sensors for perceiving the external environment. While laser scanners and ultrasound or infrared sensors are used to help the robot navigate safely in the physical environment, sensors for vision, audio, and touch are used to perceive information that are especially relevant to initiate and monitor social interaction between a human and the robot. The visual modality is usually used to detect objects and persons in the environment and analyze their properties. For example, Breazeal et al. [2004] used cameras mounted in the eyes of the Leonardo robot to detect and track the face and facial features of the human interaction partner. Malfaz et al. [2011] used the robot Maggie’s camera to detect whether there were people standing near the robot. Tanevska et al. [2019] used the eye cameras of the iCub robot to detect the face and recognize the facial expressions of the human interacting with the robot. Occasionally, sensors mounted in the environment (external to the robot) are used to augment the visual capabilities of the robot. Breazeal et al. [2004] used cameras fitted on walls behind the robot and above the workspace in order to detect and track objects and
people in the broader environment of the robot. More specifically, the overhead camera was used to estimate the head pose and recognize the pointing gestures made by the human interaction partner, as well as to detect and track the state of the shared interaction objects (electric bulbs).

Audio sensors (microphones) are generally used to obtain speech input for automatic speech recognition in order to understand a limited vocabulary of commands or instructions given by the humans (cf. [Breazeal et al. 2004, Malfaz et al. 2011]) and to subsequently extract pre-defined human communicative intents based on simple rules applied to the text recognized from speech (cf. [Adam et al. 2016]). In contrast to vision and audio modalities, touch allows humans to interact with the robot through direct physical contact. Tactile sensors are attached to different parts of a social robot’s body (e.g., head, shoulders, chest, back, abdomen) and they can be used to perceive different properties of human touch, for example, size of the area touched, the pressure of the touch, the duration of the touch, and so on. Touch is an important modality for triggering reactive movements or sounds reflecting liveliness, irrespective of whether the social robot is a humanoid, is animal-like, or has an abstract or hybrid form.

Vision, audio, and touch are mostly used separately and serve different perception goals. Even when they share the same goal, they are often combined in a logical-OR fashion. For example, Malfaz et al. [2011] used face detection and speech recognition as redundant channels to determine whether the robot is surrounded by people or is alone. Tanevska et al. [2019] regulated how comfortable iCub felt during social interactions depending on whether and how often it saw a face or felt a touch. A fusion of multiple modalities to infer the state of a human or an object in the environment is not common, which could be partly due to the challenges involved in synchronizing and matching the information provided by the different modalities. Perception outputs (or, percepts) are usually combined with inference rules stored in memory to construct beliefs about the self, the interaction partner and the objects in the environment, and also to determine the internal states of the robot, for example, emotions (cf. [Lisetti and Marpaung 2007]) or comfort level (cf. [Tanevska et al. 2019]). These beliefs and internal states influence the deliberate behaviors generated by the robot.

16.2.2.2 Action Modalities
The utility of social robots derives from their ability to act or behave socially in physical environments with human presence. The social behaviors that they can exhibit depend primarily on the action modalities that it possesses. This in turn depends on the mechatronic design of the robot. In the case of humans, speech,
facial expressions, head and hand gestures, torso movements, and locomotion constitute the key overt modalities for expressing social behavior. Humanoid social robots (e.g., Zeno [Hanson-Robotics 2007], iCub [Metta et al. 2008]) possess several or all of the human social behavior modalities, however with reduced degrees of freedom (i.e., fewer movable joints), limited ranges of motion, or alternate types of movements (e.g., locomotion by rolling instead of walking). Social robots with animal-like appearance (e.g., Leonardo [Breazeal et al. 2004], MiRo [Consequential Robotics 2020], Paro [Shibata 2012]) support further action modalities, for example, the movement of ears or tail. In addition, social robots, especially non-humanoid robots having an abstract or cartoon-like appearance (e.g., Pepper [SoftBank-Robotics 2021b]), often possess artificial modalities based on, for example, color LEDs or display screens. While most social robots are mobile, there are also stationary social robots, for example, the Furhat Robot or the robot Reeti® [Robopec 2021], which serve mainly as communicative and expressive robots. The Furhat robot has a 3D facial form on to which a human-like face is projected and virtually animated to give an impression of liveliness. In contrast, the robot Reeti® has a head with movable components like eyes, ears, cheeks, and mouth, all of which can be used to show expressions mechanically. Unlike social communication, manipulation is a functionality found less frequently in social robots. Hands and arms, if available, are mainly used for gesturing during single-turn or multiturn social interactions. However, there can be social interaction scenarios that require social robots to pick, place, carry, or hand over objects. This would require the robot to integrate task planning and execution in a socially appropriate fashion.

Even when the mechatronic design provides multiple expression modalities, the capability of a social robot to use these modalities to realize multimodal social behaviors depends on the behavior generation and control mechanisms that it is programmed with. While most architectures mention multimodal behaviors, the aspect of temporal alignment of actions across different modalities is often not dealt with explicitly. Even though some works (e.g., [Huang and Mutlu 2014, Yoon et al. 2019]) have created data-driven models to automatically generate gestures that should accompany speech by learning from annotated human data, these do not include a monitoring component that dynamically adapts the gestures based on runtime synchronization issues. Actions involving the movement of multiple joints are usually defined as a single trajectory in a multidimensional joint space. Although this implicitly describes the temporal alignment of different joints, it makes it difficult to dynamically adapt the motion of individual joints, either due to physical errors or due to a need to merge a new behavior with an ongoing behavior. Such dynamic and seamless adaptation of individual modalities is crucial for generating fluent and naturalistic robot behaviors. However, the ability to
incrementally generate and dynamically adapt multimodal behavior still remains elusive to social robots. Having said that, the desired complexity of such behavior generation and control algorithms would depend on the type of the embodiment used (e.g., animal-like vs. humanoid) as well as the chosen application domain (e.g., therapeutic vs. entertainment).

**16.2.2.3 Single-route Architectures**

The single-route architectures for social robots mostly involve only an associative route, where sensory data is processed hierarchically to derive values for internal variables that are then used to select the robot’s behavior from a small set of pre-defined actions. For example, Tanevska et al. (2019) used a single-route architecture (see Figure 16.7) to enable the social robot iCub to adapt its internal drives over time to the specific user it is interacting with. In order to enable this, they included modules to evaluate and adapt the dynamic “comfort level” of the robot based on the presence or absence of multimodal interaction stimuli, namely face and touch. Based on the current comfort level, the robot decided whether to engage or disengage with the user and accordingly selected the actions to be performed (behavior selection). The actions involved the movement of different joints on the head/neck, arms, or torso of the robot, and the motion trajectory was adapted according to the information perceived from sensory data (e.g., position of the face). As can be seen, in this architecture, information flows from multimodal behavior processing to multimodal behavior generation modules and involves no high-level deliberation or planning.

![Figure 16.7](image-url) Single-route architecture (based on Tanevska et al. [2019]) used to allow the iCub robot to adapt its behavior to the human over time (Inset: iCub robot [Metta et al. 2008]).
Early efforts by Breazeal and colleagues [Breazeal et al. 2004] focused on providing social robots with key social competencies such as establishing joint attention during interaction with a human. For this, they developed an attention system that determines (1) what the human and the social robot, Leonardo, are looking at (“attentional focus”) and (2) which objects they are referring to (“referent focus”) during the interaction. To detect the referent focus of the human, information obtained through multiple perception modalities, especially speech and vision, are integrated (see Figure 16.8). The visual information used includes pointing gestures made by the human and the eye gaze computed from the head pose. Information about the attentional and referent focus of the human and the robot are stored in the belief system, along with the attributes of the objects communicated by the human via speech. Updates to the beliefs (especially, the focus of the human and the robot) triggers several social behaviors, for example, Leonardo would shift its gaze to the location that the human is currently looking at or point to the object being referred to. The architecture used by Breazeal et al. [2004] (see Figure 16.8) to enable such social behaviors is also a single-route architecture involving behavior/action selection based on beliefs held by the robot.

**16.2.2.4 Dual-route Architectures**

The dual-route architectures for social robots also involve a flow of information from behavior processing to behavior generation modules but includes an

![Figure 16.8](image-url)
associative and a deliberative route. These architectures should support mechanisms to arbitrate and coordinate the behaviors generated via the two routes. The Cognitive and Affective Interaction-Oriented Architecture (CAIO) proposed by Adam et al. [2016] is an example of a dual-route architecture (see Figure 16.9). It is focused on enabling social robots to reason about and express its affective state while also performing conversational acts simultaneously. The dual-route processing is facilitated by the sensorimotor and cognitive emotional appraisal modules (see Chapter 10 on “Emotion” [Broekens 2021] of volume 1 of this handbook [Lugrin et al. 2021] for a discussion of emotion models). The sensorimotor emotional appraisal module maps the conversational acts of the human interaction partner into a 5D emotional representation, each of whose dimensions is mapped to specific face and body expressions by the Multimodal Emotional Action Renderer. This provides for a relatively short and fast associative route to express the initial emotional response of the robot to the human. The cognitive emotional appraisal module infers complex emotions (e.g., gratitude, reproach) for the robot based not only on the conversational act performed by the human but also on the mental states of the robot. The complex emotions feed into a long and slow deliberative processing route, which involves the selection of the next intention for

![Figure 16.9](image-url)

**Figure 16.9** CAIO Architecture (based on Adam et al. [2016]), visualized according to our schematic, as another example of the double-route architecture that is employed in virtual agents and in social robots (Inset: (left) NAO robot © 2021 SoftBank Robotics; (right) MACH virtual conversation coach [Hoque et al. 2013]).
the robot, the composing of a plan of actions (conversational acts) to achieve the intention, and the execution of these actions through verbal and nonverbal modalities. The sensorimotor emotional appraisal module also evaluates the emotion associated with the conversational acts being executed by the robot, causing the expressed deliberative behavior to be emotionally flavored.

16.2.2.5 Multidirectional, Incremental Architectures

Most social robot architectures developed so far focus mainly on multimodal behavior processing (red-colored boxes). Despite a lot of effort spent on the appropriate design of SR, the automatic generation of multimodal behaviors (blue boxes) remains an area that has received relatively little attention. Yet, naturalistic interactions require social robots to support the generation of interruptible, fluent, and spontaneous multimodal behaviors (Section 16.1.1). That is, as with IVA, we need incremental, multidirectional architectures that support incremental processing and exchange of information between components at all levels vertically, horizontally, and diagonally (Section 16.2.1). Such architectures would (i) contribute desirable features like priming and anticipation, which can speed up and improve the reliability of behavior processing, and (ii) integrate multiple parallel routes for behavior generation (reactive, associative, deliberative) that operate at different temporal granularity [Kopp et al. 2014]. Recent work focuses on the development of a multidirectional SR architecture (see Figure 16.10) that builds on an incremental communication framework [Schlangen et al. 2010], which was originally created for enabling naturalistic dialog management in conversational virtual agents. The architecture was briefly introduced in Stange et al. [2019] with a focus on the support for generating verbal explanations for behaviors and in Hassan and Kopp [2020] with a focus on the structure of its episodic memory.

The aforementioned architecture is being developed as part of a research project aimed at creating lively SIA for long-term interaction. For this, the architecture (Figure 16.10) supports the dynamic modeling of intrinsic needs of the robot (Needs Engine) as well as the inference of the mental states of the user (User Model) powered by multimodal behavior processing. An elaborate incremental, multimodal behavior generation pipeline is included to fluently integrate behaviors at three conceptual levels: (i) fast, reactive behaviors (e.g., mirroring of facial expressions, tracking a human face); (ii) previously learned associative behaviors (e.g., idling when not engaged with the user, performing daily rituals such as greeting the user in the morning); and (iii) deliberately planned behaviors (e.g., getting acquainted with a new user, explaining own behavior to the user). A decision-making module (Decision Engine) is used to select high-level intents aiming at optimizing the internal needs of the robots as well as those of the user. The flow
Figure 16.10 Example of a multidirectional, incremental architecture for enabling lively interactions between a human and a social robot (based on Hassan and Kopp [2020], Stange et al. [2019]), visualized according to our schematic (Inset: (left) Pepper robot © 2021 SoftBank Robotics; (right) VIVA robot 2021 navel robotics).

of information is supported in multiple directions: bottom–up, top–down, left–right, and right–left. For example, feedback about the execution status of actions is used to adapt the selection of high-level behaviors in the future (bottom–up); the intrinsic needs of the robot are influenced by internal and external events perceived by the robot (bottom–up) as well as the success of its own plans (right–left); the discourse context as well as other information in the active memory are used to incrementally update/correct perception models and resolve ambiguities in perception and interpretation of sensor data (top–down); multimodal behavior controller purges/adapts certain reactive behaviors (top–down, right–left); active memory triggers idle behaviors according to the active interaction mode (left–right); dialog planner uses information gathered from past interactions to adapt future conversational acts (left–right). These illustrate the rich possibilities that the proposed multidirectional, incremental architecture would be able to provide to social robots in order to make them more lively, fluent, and naturalistic in their interactions with the user. This architecture can be used not only for SRs but also for IVAs.
16.2.3 Similarities and Differences of Virtual Agent and Social Robot Architectures

Comparing the multimodal interaction architectures that have been developed and employed in virtual and robot agents, a number of commonalities but also differences can be noted. Overall, both kinds of systems have to implement the main columns for processing, mapping, and generation, and to integrate them in a full architectural layout. In both fields, we find dual/multi-route architectures, which have been developed early on in the field of mobile robotics and have been adopted also for ECAs and social robots. Likewise, sub-architectures for online behavior processing or offline behavior generation have been developed and applied to both kinds of agents (e.g., [Ishi et al. 2018]). Also, multidirectional and incremental processing has been identified as an overarching key feature and is addressed in architectures for both physical and virtual SIAs [Kopp et al. 2014, Stange et al. 2019].

However, a number of differences remain and hence offer opportunities for how one field can learn from the other. Obviously, one key difference is the physical embodiment of a robot and the constraints it implies for multimodal behavior processing (e.g., limited abilities to gather sensory information about a user’s communicative behavior in a dynamic physical environment) as well as generation (e.g., limited abilities to produce expressive, subtle nonverbal behaviors under given bodily or kinematic limitations). Consequently, social robot architectures usually have concentrated on dealing with recognition problems (left-hand side of the schematic model) as well as fast lower-level routes in order to achieve situation-awareness and robust behavior. Social robots thus provide a great test bed for embodied approaches to multimodal communication and socially reciprocal behavior coordination. In addition, although social robot bodies are carefully designed for expressiveness and engagement, their inherent physical limitations imply interesting challenges for researchers working on real-time generation of consistent, synchronized multimodal behavior [Ng-Throw-Hing et al. 2010]. Here, the transfer and application of behavior realization frameworks from virtual agents to social robots has led to some advanced generation sub-architectures that rest on closer feedback loops and more flexible timing and motion planning [Salem et al. 2012, Niewiadomski et al. 2013].

Full-fledged conversational social robots have been rarely reported in comparison to conversational virtual agents. Instead, the focus in SR research has so far been to explore processing and control mechanisms required for specific and by design pre-structured interactions between a human and a robot. Natural conversational interactions with, for example, fluent turn-taking is still a challenge of such systems [Skantze 2020]. For example, Adam et al. [2016] qualitatively
demonstrated the integration of emotional appraisal with deliberately chosen conversational acts in order to produce multimodal (speech and gesture) behavior of the NAO robot [SoftBank-Robotics 2021a] based on an architecture that was initially developed for a virtual character. Even though Adam et al. [2016] demonstrated how fast and slow cognitive processes could be integrated (see Figure 16.9), the interruptability of the deliberative or sensorimotor loops, for example, due to new incoming verbal input, was not explored. This requires a control strategy for managing turn-taking during conversations. Chao and Thomaz [2013] proposed the use of Timed Petri Nets (TPN) to regulate the conversational floor and thereby handle the dynamic turn-taking process in dyadic interactions between a human and a robot. Learning from and adapting to a user’s preferences and skills is a key requirement for social robot companions. Park et al. [2019] used multimodal affective cues from verbal and nonverbal channels as “reinforcement” or human feedback to adapt the storytelling policy of their social robot Tega, to increase the child’s engagement with the robot and to improve the learning outcomes. These models could be directly applied to IVAs as well.

16.3 Current Challenges and Future Directions

In our discussion of the requirements of multimodal conversational interaction and the architectures used to build SIAs that shall be able to engage in it, we have already identified a number of trends and challenges. Given that the field is relatively young and still exploring new methods for behavior processing, mapping, and generation, these challenges can be expected to persist for the next decades. In addition, a number of challenges and future directions can be identified that have or most likely will become crucial and the focus of this upcoming research.

16.3.1 Interaction Memories and Learning

A core component of modern, learning-based architectures is a memory that aggregates information and makes it available for interaction with self and others. Memories can be distinguished according to what information they encode, how much, and how long they can retain it. Technical agent architectures differ in the types of memory systems that are included, their representations, the processes that operate on them, and how the memory influences other components and processes modeled within the architecture. For example, SOAR [Laird 2008, 2019] and MLECOG [Starzyk and Graham 2017] include short-term as well as long-term memory (episodic, semantic, procedural), while working memory plays an important role in the EPIC architecture [Kieras and Meyer 1997]. Likewise, cognitive architectures of SIAs often include memory as a key component. For example, CAIO stores information as long-term episodic, semantic, and procedural memories but
does not explicitly include a working memory. In Dodd and Gutierrez [2005], short-term and long-term memories are used for the ISAC robot, and that includes sensory, episodic, semantic, and procedural memories. The architecture proposed in Malfaz et al. [2011] use a long-term memory for supporting deliberative functions and a short-term memory for storing temporally relevant information in the social robot Maggie [Salichs et al. 2006]. Kasap and Magnenat-Thalmann [2010] propose long-term and short-term episodic memories to enable affective interaction between humans and social robots. The virtual agent SARA stores the adopted conversational strategies in “social history” and the preferences of and rapport with the user in the “user model” [Matsuyama et al. 2016].

Due to the increasing relevance of learning and adaptation in SIAs, the episodic memory is receiving growing attention, especially to store and provide access to past experiences and events. In agent architectures, episodic memories are usually created by filling pre-defined templates with specific information (cf. [Kasap and Magnenat-Thalmann 2010, Rabe and Wachsmuth 2013]) or by storing sequences of events that occurred while performing tasks (cf. [Dodd and Gutierrez 2005, Kasap and Magnenat-Thalmann 2010]). These models are quite restrictive since the dynamism and complexity of interaction contexts make it difficult to predict the events that might occur during an interaction episode or the exact time at which they might occur. Nuxoll and Laird [2007] proposed a design space to guide technical implementations of episodic memories. This could serve as a useful guide for current and future research on episodic memory models. In Hassan and Kopp [2020], we presented a concept for an episodic memory model for storing interaction episodes, which addressed three aspects of this design space, namely, when an interaction episode is encoded, what its content are, and how it is structured. The proposed model represents episodic memory as hierarchies of labeled time-intervals when a user and an agent were engaged in an active interaction, initiated by either of the two parties. Relevant internal and external events are linked to the episodes based on their relationship with the episodes (i.e., causal, goal, or enabler events). A complete episodic memory model would, however, require that all aspects related to the encoding, storage, and retrieval of episodic memories are addressed.

16.3.2 Cognitively Inspired versus Application-centered Architectures
Related to the previous topic, a larger issue for future work will be to identify principles of cognitive architectures that facilitate social interaction, and to develop them into technical interaction architectures. Over the past decades, several architectures have been developed that identify, model, and weave together different
cognitive processes in order to provide artificial agents (virtual agents or physical robots) with the computational framework for autonomous and intelligent behavior. Cognitive-psychological architectures include, for example, the widely known cognitive architectures ACT-R [Anderson et al. 2004], SOAR [Laird 2008, 2019], and CLARION [Sun 2007] as well as the recently developed MLECOG [Starzyk and Graham 2017], which focuses on motivated learning capabilities. These architectures are generic and have a strong theoretical basis, but they do not focus on social reasoning or the generation of socially appropriate behavior that evolves over time. More application-centered architectures are usually developed to meet the requirements and demands of specific use-cases or applications, without giving much regard to cognitive or psychological principles. These solutions are frequently met in virtual agents as well as robotic systems and include, for example, the architectures developed in Kasap and Magenat-Thalmann [2010] and Dodd and Gutierrez [2005]. However, a common framework for an architectural layout that meets the requirements for fluent multimodal human–agent interaction and, in particular, integrates the many components that are needed on a principled basis is lacking. In particular, the increasing importance of integrating perception, action, memory, and learning is likely going to raise a need for cognitively plausible architectural concepts. For example, the need for robust and fluent interaction capabilities will require concepts for incremental yet concurrent and integrated processing at different layers of the architecture. Cognitive principles like good-enough reasoning or computational rationality may have a major role to play in these future systems.

### 16.3.3 Interaction-aware Behavior Generation

Another important direction for the future, which is already starting to emerge, is the consideration and integration of the larger interaction context into local processing sub-architectures. While many works have investigated how background information such as personality, relational status, or cultural background can be taken into account when processing or generating multimodal behavior (c.f. Chapters 7 on “Gesture Generation” [Saund and Marsella 2021], 8 on “Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021], 13 on “Culture for Socially Interactive Agents” [Lugrin and Rehm 2021] of this handbook [Lugrin et al. 2021], and Chapter 18 on “Adaptive Artificial Personalities” [Janowski et al. 2022] of this volume of this handbook.), recent work has also started to explore how the current, dynamically changing interaction context can be integrated. While this has often been reported as crucial in human social interaction (e.g., for alignment, empathy, or coordination), it has only recently and partially been tackled in technical attempts. For example, the current and previous
body poses or facial expressions of the interlocutor have been integrated into the generation of respective behaviors of an agent [Ahuja et al. 2019, Jonell et al. 2020]. Those attempts are precursors of what we would term interaction-aware multimodal behavior generation. It will require new approaches to combine learning-aware with model-based approaches that allow for learning at the level of interactional behavioral couplings and to embed this into the incremental processing of dynamic interactive behaviors.

16.3.4 Uncertainty-awareness in Social Interaction

It is commonsense that systems that are to operate robustly in real-world environments, which are dynamic, stochastic, or only partially observable, require uncertainty-aware models. This is also true for SIAs, where uncertainty arises in the determination of the interaction context, in the inference of the meaning of social signals, in the attribution of mental states to the interaction partner, or in the prediction of possible effects of own multimodal behaviors. Uncertainty modeling is thus of special importance to the processing components of the multimodal interaction architecture shown in Figure 16.1; since sensing is bound to be noisy, recognition models are not error-free, and interpretation is driven by hypotheses formed under partial knowledge. Noise and errors also accumulate over successive processing stages in the architecture. This implies that the mapping components in Figure 16.1 have to operate with uncertain information to decide, select, or trigger an appropriate behavior of the agent. In the case of SRs, additional challenges for planning and control arise from the fact that generation components of the architecture have to deal with uncertain estimates of duration or outcomes of actions. Algorithms and approaches are available that can be used to handle uncertainties during interpretation (e.g., Bayesian networks for mental state modeling [Pöppel and Kopp 2018]) or decision-making (e.g., decision-theoretic planning methods). However, due to the complexity of these approaches as well as a lack of uncertainty estimates for multimodal sensing and recognition, existing SIA architectures often tend to ignore the handling of uncertainty during interaction. Recent efforts to quantify uncertainties associated with data-driven machine learning models and their predictions (cf. [Abdar et al. 2021])—inspired by the seminal work of Yarin Gal (cf. [Gal and Ghahramani 2016, Kendall and Gal 2017])—address this problem in perception tasks in real-world applications. Approaches for perception that combine prior knowledge with data-driven methods within probabilistic frameworks have also been reported recently (e.g., [Seuss et al. 2021]).

Future SIA architectures should focus on integrating uncertainty estimates into the interaction pipeline. That is to say, one should design, implement, evaluate, and optimize probabilistic versions of multimodal SIA architectures in order to
promote their successful application in real-world scenarios (outside laboratory settings). Furthermore, future human–robot interaction research should focus on investigating the influence of uncertainties on social interaction and on using probabilistic SIA architectures to autonomously generate such uncertainty-aware social behaviors. Finally, SIAs should be able to not only handle uncertainties technically but also to create interactive behaviors that can adapt to as well as communicate various types and degrees of uncertainty.

**16.3.5 Evaluation Measures**

A final future direction that is going to become crucial is the development of metrics and criteria for the evaluation of multimodal behaviors. In the past and the present, the evaluation of multimodal behavior is mostly done by employing human raters in perception studies, who then rate the behavior for naturalness, coherence, or style. There is a growing consensus in the field that an objective and more systematic evaluation methodology is missing and strongly needed (e.g., see Wolfert et al. [2022] for a review). Further, such metrics are needed as optimization criteria for the development of machine learning-based models, which are increasingly trained in an adversarial fashion (i.e., using a discriminator or critic). Comparison against training data as usually done, however, is insufficient as the architecture’s ability to generate multimodal behavior in new interaction contexts, in which it eventually needs to be communicatively effective and successful, cannot be assessed in this way. First approaches are seen for the evaluation of singular multimodal ensembles, for example, by analyzing the internal temporal synchrony or semantic congruency between the modal behaviors. Yet, future work will need to investigate how a virtual agent’s or social robot’s long-term multimodal behavior in a given interaction context can be assessed. Finally, another aspect crucial in evaluating interaction architectures is how they can be inspected. An important future challenge is thus to increase the interpretability and transparency of a SIA architecture, such that developers as well as users can understand why a certain behavior has been shown. Indeed, the growing use of black-box machine learning models raises a need for explainable SIAs, whose interactive behavior can not only be received but also interrogated.

**16.4 Summary**

In this chapter, we presented the overarching requirements that should be fulfilled by virtual agent or social robot architectures in order to support fluent and human-expected multimodal interaction. We discussed several existing architectures for virtual and robot agents, especially with respect to the modalities they support, the different processing routes they use, and the aspects of social interaction
they realize. We proposed a schematic for organizing the different modules and pathways in multimodal interaction architectures and categorized existing architectures into single-route, dual-route, and multidirectional architectures based on the processing pathways included. While the focus of virtual agent architectures has been mainly on expressive behavior generation, social robot architectures focused mainly on multimodal behavior processing and robustness. Either field is hence characterized by individual strengths and limitations. Although approaches are increasingly applied to both kinds of SIAs, the methods and models developed could inspire and inform each other even more to yield architectural principles and frameworks that enable advanced multimodal interaction capabilities to become standard in the field of SIAs.

Overall, the field has explored a variety of modalities, techniques, and integration architectures—and is still extending its repertoire of approaches for specific generation problems. Future work will need to consolidate our views on what we can generate, whether increasing data will help, how to join different approaches (and motivations), and what we should optimize models for (behavior quality metrics and felicity conditions). Future work should also focus on developing learning-based models that afford representations and interaction memories that can dynamically scale to heterogeneous data of different formats and temporal resolutions, and can enable a SIA to process, interpret, and learn from long-term interaction data in order to re-evaluate the social-appropriateness of behaviors based on past experiences.

References


References


18th International Conference on Intelligent Virtual Agents. ACM, 79–86. DOI: https://doi.org/10.1145/3267851.3267878.


Motivation

While most research has focused on one-on-one interactions, Socially Interactive Agents (SIAs) for multiparty interaction have received increasingly more attention in the past years for a number of reasons. First, multiparty social interactions are often more unstructured but also more likely to resemble real-world situations. As these agents move from lab environments to the real world (e.g., museums, hospitals, or classrooms), it is important that they can handle the different types of social situations that may arise. Second, human social behavior is largely dependent on social context, for example, the way we behave alone is different from how we behave in a group [Zajonc 1965]. For this reason, we cannot directly assume that empirical findings, computational models, and so on, from one-to-one interactions between users and SIAs will hold in a multiparty setting.

When compared to one-on-one interactions, multiparty settings pose additional challenges for humans. Thus, it is no surprise that they can also be more challenging for SIAs. The core capabilities in this domain are no less social than at the individual level but typically relate to space management, formation control, and navigation. Additionally, questions such as how the agent behavior and
appearance can positively affect group dynamics (e.g., collaboration or teamwork) as well as users’ subjective experience (e.g., perceived presence or trust) are relevant to explore in a multiparty context. Therefore, robust multiparty interactions involve a large set of competencies to some degree, combining the challenges from one-on-one interactions while also raising new challenges. For example, conversational management can become more difficult with more participants in aspects such as turn-taking and verbal and nonverbal communication. When considering embodied agents, how should these agents position and orient themselves within the group such that they can equally participate in the interaction? Finally, when it comes to perception, so far, little is known about how perception models perform when they are tested in a group size different than the one they were trained on. However, this feature is critical for some perception problems, that is, the way in which an agent should interpret a user glancing to the side is different if that user is alone or if the user is in a group.

One can consider interactions between humans and SIAs to take place at three levels: crowd, group, and individual levels [Panzoli et al. 2010]. Research to date has been focused in separate strands on the crowd and individual levels, but modeling multiparty interactions at the level of small groups (the main focus of this chapter) has not received as much attention in the literature; yet, it is crucial for more natural interactions in many real-world situations.

Another important differentiation relates to whether SIAs are static or mobile within the environment. The former case relates to situated social gatherings that take place in a specific location, for example, in the case of free-standing conversational groups, in contrast to those situations in which participants move together through an environment toward a shared destination. Our focus in this chapter will primarily be on the former case, in which the group is in a static position within the environment. It should be noted that, even in cases in which the group might be considered to be static in its locale, individual agents may still have some mobility. This may be due, for example, to individuals changing their positioning within a formation during the course of an interaction to accommodate newcomers or re-form due to a departure of a group member.

This chapter surveys recent work on multiparty SIAs, focusing on small group interactions or social gatherings (with multiple humans and/or multiple agents). We begin by introducing models and approaches from other disciplines and then summarize the recent advances in multiparty interaction with both Intelligent Virtual Agents (IVAs) and Social Robots (SRs). We then summarize some of the main similarities and differences between IVAs and SRs in multiparty interaction in an attempt to establish synergies between the two communities. We conclude by discussing current challenges and future research directions.
17.2 Models and Approaches

This section serves as an overview on models, approaches, and background knowledge from other disciplines that are commonly used in research concerning multiparty SIAs.

17.2.1 Social Gatherings

In social psychology, gatherings correspond to a set of individuals who are in one another’s immediate presence [Goffman 1963]. Unfocused gatherings are typically associated with mere co-presence such as pedestrians on a street or strangers waiting for a bus. Focused gatherings are instead characterized by individuals coming together to sustain one focus of attention.

According to Kendon [1988], there are two main types of focused gatherings. If there is a joint responsibility between the people in a gathering to cooperate to sustain a focus of attention, the interaction is considered a jointly focused gathering. Examples include social conversations, ping-pong games, dancing partners, and groups of workers cooperating to solve a task that requires sustained attention. When there is no need for shared cooperation to sustain the focus of attention, the interaction is rather considered a common focused gathering. For instance, common focused gatherings include a platoon on a parade or pupils paying attention to what a teacher says in a classroom.

Information is given voluntarily during gatherings, for example, through what people say. In addition, information is given off whether the interactants choose to provide it or not. As Kendon [1988] described, this is an inevitable and unavoidable product of people’s presence and of their actions. For example, groups of people might provide additional information through their gaze or spatial patterns of behavior. While the latter aspects may seem unimportant in comparison to the information that is provided voluntarily, they play a key role in structuring social encounters.

Situated human conversations have traditionally been considered the most common type of jointly focused gatherings [Kendon 1990]. The members of these gatherings converse in one another’s immediate presence. They work cooperatively to sustain their focus of cognitive and visual attention, pursuing a common line of concern. The cooperative nature of conversations means that they often end when a participant has the turn to speak but, for some reason, he or she does not do it.

17.2.2 Groups and Teams

An important distinction is often made between gatherings involving groups of agents and teams [André et al. 2020]. Groups correspond to agents that are aware of having a shared identity. Meanwhile, teams are more specific. They are groups
in which the agents have a shared goal or task [Groom and Nass 2007]. Team members collaborate and support each other to accomplish their joint goal(s). Worth noting, the success of teams is not a given. Team characteristics such as member’s perceived inclusion [Jansen et al. 2014] and psychological safety [Edmondson 1999] are typically associated with effective teams. Readers interested in a broader discussion of team performance from an organizational psychology perspective are encouraged to refer to Guzzo et al. [1995].

17.2.3 Proxemics

An important aspect of social gatherings pertains to people’s use of physical space, or proxemics as coined by Hall [1966]. Hall described four distance zones typically used by people during interactions. These zones correspond to the intimate, personal, social, and public distances that people tend to keep from each other based on their emotional state and type of social engagement. The intimate distance is short, affording physical interaction. Personal distances are often kept by friends or family when conversing, whereas social distances are more common for acquaintances during situated social gatherings. Finally, the public distance is well outside an individual’s circle of personal involvement, typically emerging during public addresses. Several factors are known to influence human proxemics, including lighting [Adams and Zuckerman 1991], cultural factors like social norms, peoples’ familiarity with one another, and to what degree people interact together [Argyle 2013].

17.2.4 Face-formations in Social Conversations

During conversations among free-standing people, the participants position themselves to create a sort of “no-man’s land,” maintaining a separate world from their surrounding [Kendon 1990]. The result is a distinct spatial organization, typically known as a face formation or F-formation in short within social psychology. F-formations maximize the opportunity of the interactants to monitor one another during conversations. They also help maintain groups as spatially distinct units from other nearby focused gatherings.

Kendon [1990] described the emergence of F-formations and their structure based on observations of social events. F-formations begin when the members of a group position themselves such that their transactional segments intersect. These segments correspond to the physical space in front of each person. They correspond to the space into which individuals look and speak or into which they reach to handle objects relevant for their current task. People will work to maintain their transactional segment free of intrusions for as long as they are engaged in an activity that requires it.
The physical area where the transactional segments of the members of a conversation intersect is the o-space of the corresponding F-formation. As shown in Figure 17.1, the o-space is in-between individuals in a group, whether they are standing in a face-to-face arrangement or in semicircular or circular formation.

The spatial organization of the participants of a gathering often reveals transitions between social conversations and other types of interactions. For example, F-formations often transform into a less uniform spatial arrangement when a conversation shifts into a common focus encounter [Kendon 1990, Marshall et al. 2011]. When the focus of attention becomes a particular person, a separation between this interactant and the rest of the group is often observed due to a difference in social status or role.

17.2.5 Other Group Phenomena

F-formations are one type of group phenomena that emerges during social gatherings, but other important phenomena relate to group social influence. For example, conformity [Crutchfield 1955] concerns agreement to the majority position within a group. Kelman [1958] distinguished between three processes that result in group influence: compliance to fit in within a group; identification, which results in compliance to establish or maintain a desired relationship within a group; and internalization, which occurs when the adopted ideas or behavior is intrinsically rewarding.

Other related group phenomena in multiparty human interactions include diffusion of responsibility [Darley and Latané 1968] and group polarization [Myers and Lamm 1976]. The former phenomenon is said to emerge when the likelihood of people taking the responsibility for action or inaction is reduced in the presence of others. The latter phenomenon results in groups making more extreme decisions than their individuals would in isolation.
Chapter 17  Multiparty Interaction Between Humans and Socially Interactive Agents

Relevant concepts frequently used in the social sciences to study group dynamics [Abrams and Rosenthal-von der Pütten 2020] comprise: *Ingroup identification*, the individuals' perception of themselves as member of the group [Ashmore et al. 2004, Leach et al. 2008]; *cohesion*, the inside perspective on the forces that keeps the group as a group [Dion 2000]; and *entitativity*, the outside perception of the groupness of a social group [Campbell 1958]. A phenomenon that not necessarily arises from but might influence small group interactions is *ingroup-favoritism*. Ingroup-favoritism has been phrased by the findings that people tend to act more favorably toward ingroup members than toward outgroup members [Tajfel et al. 1971, Brewer 1979].

### 17.3 Advances in Multiparty Interaction

The development and evaluation of SRs that can interact with groups of people has been explored in several domains. In the following sections, we review different aspects of multiparty interactions and the respective capabilities of SIAs.

#### 17.3.1 Evaluating and Understanding Groups

Group dynamics encapsulate the influential actions, processes, and changes that are observable within and between groups. Group dynamics change the individuals in the groups in which they occur and, potentially, even their society [Forsyth 2018]. This makes the study of groups and group dynamics interesting for creating SIAs.

The next sections discuss research on multiparty interaction regarding the believability of SIAs in group settings, users' attitudes toward these agents, and spatial group behavior. We also discuss prior efforts aimed at understanding human–agent group dynamics.

#### 17.3.1.1 Measuring Believability

Human perception studies have been used to evaluate the behavior of small groups of virtual characters, often focusing on human sensitivities to different behaviors and their impact on believability or perceived naturalness of the group. These include studies investigating whether people are more sensitive to similar appearances or motions in a group of characters when perceiving their variation [McDonnell et al. 2008], the number of agents per group and distribution of groups that appear to be the most realistic in crowd situations [Peters and Ennis 2009], and the degree to which people are able to see groups in a crowd as the camera viewpoint and crowd density are varied [Yang et al. 2018].

McDonnell et al. [2009] investigated human sensitivity to the coordination and timing of conversational body language in small groups of virtual characters
and concluded that participants are sensitive to desynchronizations in the turn-taking behavior across group members. Ennis et al. [2010] investigated sensitivities to desynchronizations of body motions, gestures, and voices in small groups and found that viewers were most sensitive to desynchronizations of full-body motions.

### 17.3.1.2 Measuring Attitudes

Different behaviors employed in SIAs interacting in multiparty environments have been shown to affect the perception on themselves. In the following paragraphs, we review studies that evaluate attitudes toward SIAs.

The study by Cafaro et al. [2016] investigated the interpersonal attitudes (friendly versus unfriendly) of agents within a small group and toward an approaching avatar and found that the interpersonal attitude of the group had an impact on the proxemics behavior of the avatar. Further, it was found that the attitude of the group toward the approaching avatar had a major impact on social presence evaluations. Pereira et al. [2014] created a case study in which an SR plays the Risk board game against three human players in order to investigate if the agent was perceived to be socially present.

Fraune et al. [2015a] examined how humans respond to different numbers of robots (one versus three) with different social capabilities (social versus non-social) in a naturalistic scenario of robots acting as trash collectors. When robots were acting in a sociable manner toward another robot, they were perceived as more anthropomorphic. If the same robots were acting sociably toward humans, positive attitudes and emotions, the willingness to interact, and encountered physical proximity were improved [Fraune et al. 2020]. In a different multirobot environment, Tan et al. [2019] found that the sociability of a pure functional robot can be increased through witnessing social robot–robot interaction.

Humans but also robots can be seen as members of an ingroup or outgroup [Kuchenbrandt et al. 2011]. The introduction of a robot as ingroup versus outgroup was found to increase the positive perception of the robot [Kuchenbrandt et al. 2013], and social categorization has been found to play an important role when perceiving the robot as an ingroup or outgroup [Eyssel and Kuchenbrandt 2012]. When humans are paired with robots in a competitive game, humans perceive their own team as an ingroup and the other human–robot team as outgroup. Allocating painful noise bursts to ingroup and outgroup humans and robots revealed that humans develop ingroup favoritism for robots over humans as they prefer ingroup robots over outgroup humans [Fraune et al. 2017b].

Cultural aspects as well as gender and personality are important components in multiparty interactions involving SIAs. Endrass et al. [2011] focused on how
human observers perceive culture-related differences for groups of two virtual agents engaged in small talk, that is, informal discourse behaviors. The results from a study are used to inform a model for the automatic generation of culture-specific small talk dialogs for virtual agents. Mascarenhas et al. [2016] found participants prefer those group of IVAs that display the same cultural bias (individualistic versus collectivistic) as present in their own culture. Damian et al. [2011] developed a software framework that allows IVAs to display differences in personality and gender which also influence group formations.

In general, the results from these studies are intended to inform the creation of more believable and effective group behavior generation models [Huere et al. 2010].

With an increasing number of SRs present in everyday and work environments, robots are to be expected to work in teams with humans. Specifically when humans and robots partner in teams against another human–robot team, aspects of the robot behavior have shown to be influential on the group. When robots show different orientation goals (competitive versus cooperative), interaction patterns differ in terms of socio-emotional support and gaze behaviors [Oliveira et al. 2018]. Further, expressing different levels of warmth and comfort has shown to influence feelings, perceptions, and future intent to work with the robot [Oliveira et al. 2019]. In addition, a robot expressing group-based emotion in this kind of setting can lead to higher group identification, group trust, and likeability of the robot [Alves-Oliveira et al. 2016, Correia et al. 2018]. As a different aspect of the human–robot partner interaction, Correia et al. [2016] found that previous encounters with the robot partner positively influence trust toward this robot over the course of a game. From a theoretical point of view, de Visser et al. [2019] propose a human–robot team trust model that builds upon relationship equity and aims to ensure longitudinal trust.

The appearance of robots and groups of robots has been found to have influence on the attitudes toward those. Influence on the perception of robot groups was found based on different robot types—mechanomorphic, zoomorphic, and anthropomorphic. An interaction between type and perception of groups versus individual robots was reported by Fraune et al. [2015b]. Further, entitative robot groups were perceived as more threatening than diverse groups [Fraune et al. 2017a], and a correlation between the perception of entitativity of the robot group and fear toward this group was found [Fraune et al. 2019b].

### 17.3.1.3 Spatial Behavior Understanding

Other studies have investigated group formations and behaviors taking place within the group. A study by Ennis and O'Sullivan [2012] indicated that participants
Figure 17.2  Spatial behaviors have been studied among SRs and IVAs. The left image shows IVAs positioned in formation accounting for social spaces with O-, P-, and R-spaces illustrated. Adapted from Yang and Peters [2019c]. The other two pictures illustrate F-formations with a furniture robot. The middle image shows the circular arrangement during a social role-playing game [Vázquez et al. 2015a]; and the right image shows the circular arrangement during a brainstorming activity with the robot [Vázquez et al. 2017].

were sensitive to the distance and orientation of individual agents in social formations.

The study by Carretero et al. [2014] explored the impact of task-irrelevant background expressions on the perception of emotional expressions of a small group of foreground characters. They found a consistent effect of task-irrelevant negatively-valenced background stimuli on the perception of the emotions of the foreground task-relevant group of characters.

Palmberg et al. [2017] conducted a study investigating the impact of facial expressions and full body motions on the perception of intense positive and negative emotional expressions in a group of three virtual characters and found that the emotional valence of facial expressions had a stronger impact on the perception of emotions in the group than body motions.

Understanding spatial behavior has also become increasingly important in human–robot interaction. Several factors can influence the level of comfort that people have with robots and, thus, the distance that they like to maintain from them. For example, these factors include a robot’s gaze [Mumm and Mutlu 2011, Ruhland et al. 2015] and personal experience with pets and robots [Takayama and Pantofaru 2009]. In regard to spatial behavior typical of conversations, several efforts have provided evidence of the emergence of F-formations in HRI [Huettenrauch et al. 2006, Vázquez et al. 2014, 2015a, 2017, Bohus et al. 2017]. The distancing between robots and group members during F-formations might be influenced by the specific embodiment of the robots, although more research is needed to understand systematically such potential effects. Figure 17.2 illustrates how humans
build F-formations when interacting with a robot and how virtual agents position themselves socially.

Interestingly, Kuzuoka et al. [2010] showed that a robot can influence the body orientation of a museum visitor by rotating its own body. This suggests that people may adapt to robots’ spatial patterns of behavior in a similar manner as they adapt to human spatial behavior during situated conversations. Further, Yousuf et al. [2012] and Vroon et al. [2015] investigated suitable social positioning behaviors for mobile robots during social interactions.

### 17.3.1.4 Understanding Group Dynamics

A different set of works has investigated how SIA behavior affects the human group members or how groups of SIAs can affect human behavior.

How an SR could interact with groups of people in a museum or shopping mall was early investigated by Bennewitz et al. [2005] and Kanda et al. [2010]. Asking which kind of groups of people would interact with a robot in an open setting such as a shopping mall, Fraune et al. [2019a] found that highly cohesive groups engaged in longer conversations with the robot and acted more socially and positively toward the robot. Further, people who were by themselves unlikely to approach the robot were encouraged through the group and the group's norms to interact with the robot. When groups of people and robots interact in a prisoner's dilemma, it was shown that the number of people in the group significantly affected cooperative behavior [Chang et al. 2012]. In this experiment, those who interacted individually with the robot cooperated more with the robot than those interacting in pairs. However, the number of robots that pairs or individuals interacted with did not affect cooperation. When interacting with either one or three robots, individuals and groups, further, showed more competitive behavior when interacting with the same number of robots [Fraune et al. 2019b].

The influence of the group of robots that a human joins when making decisions was studied under the aspect of conformity and peer pressure. Where adults as opposed to children did not conform to a group of robots when the answer was unambiguous [Brandstetter et al. 2014, Vollmer et al. 2018], adults could be convinced by a group of robots if there was no objective correct answer [Salomons et al. 2018]. A further finding of this work indicates that the level of conformity depends on the level of trust toward the group of robots, here influenced by the correctness of the robots’ decisions. Considering the aspect of cultural differences, Wang et al. [2010] have shown that the way a robot can influence group decisions can be depending on the cultural origin of the participants. Exploring conformity from a robot's perspective, Fuse and Tokumaru [2019] investigated how conforming robots influence human decision making. As one aspect of nonverbal
behaviors, human-like gaze patterns directed toward two listeners during a story-telling task have been found to help story recall [Mutlu et al. 2006] and to increase the persuasive power of a robot [Ham et al. 2011].

Intentional group coordination is occurring when a group of humans joins in a cooperative group task such as dancing. One and two robots joining the coordinated group task have been shown to affect the group dynamics [Iqbal and Riek 2017b].

Addressing the question of team performance, backchanneling has been found to ease cognitive load and stress in a complex task [Jung et al. 2013]. With the goal of improving small-group decision-making meetings, Shamekhi and Bickmore [2019] investigated how a robot could act as a facilitator. Moving to education environments, Rosenberg-Kima et al. [2020] explored the potential of a robotic facilitator for small group learning in higher education and compared it to teacher and tablet facilitation.

Like work teams, groups of family members have been studied. A robotic therapist has been shown to improve intimacy and positive affect between romantic couples [Utami and Bickmore 2019]. Short et al. [2017] explored how a socially assistive robot could support intergenerational family groups, that is, older adults in these groups. In a different family setting, Gvirsman et al. [2020] explored how the triadic interaction between toddler, caregiver, and robot can be beneficial for early second language learning.

By giving virtual agents the possibility to use different physical appearances, Reig et al. [2020] found that personalization could be exploited when interacting with multiple users by exploring the concept of re-embodiment and co-embodiment.

### 17.3.1.5 Influencing Group Dynamics

As different works discussed in the previous sections have also found effects on group dynamics, research interests have been targeted at robots and their behaviors that can positively influence the dynamics in a group of humans. This influence can be twofold and either impact how humans act among each other or toward the robot. The robot’s efficacy has been discussed for different aspects of group processes. In the following, situations of conflict, inclusion, collaboration, robot abuse, and conversation will be discussed with the approaches taken to improve these situations with the help of SIAs.

Different roles that an SR might take in a group set-up have been explored, and Engwall et al. [2020] discuss four roles that a robot could take in a triadic language café setting. To improve conversations and meetings, SRs have been employed as facilitators. With the goal of balancing participation, Matsuyama et al. [2015] proposed a facilitation model incorporating the robot as a fourth participant, and
Shamekhi and Bickmore [2019] investigated how a meeting facilitator could in addition ensure an efficient meeting. With the same goal in mind, another work used a microphone-shaped robot—Micbot. Micbot was shown to be able to balance the engagement of a group of three and thereby achieve higher group performance [Tennent et al. 2019]. The robot could encourage passive members to participate more actively with nonverbal and indirect cues, executing two distinct behaviors—*follow* and *encourage*. By employing a robotic object as a side-participant [Hoffman et al. 2015] in a debate, Rifinski et al. [2020] found that minimal movement implying gaze and leaning can improve the interaction and the interpersonal evaluation. In an application in a virtual city, balanced theory has been found to be applicable when a virtual agent mediated a conversation between two avatars and influenced the attitudes toward itself [Nakanishi et al. 2003].

Taking a different role in the group setup, robots have been used as mediators in conflict situations. For example, a robot could promote more constructive conflict solving behavior in cases of object possession conflicts among children [Shen et al. 2018]. When personal violations cause a group conflict, a robot acting as an emotional regulator was found to regulate and call attention to a conflict [Martelaro et al. 2015]. The role of IVAs for mediating conflict was explored in a debriefing scenario [Haring et al. 2019].

Moreover, different works have explored how to facilitate collaboration and group cohesiveness by comparing task-focused and group-focused robot behavior. Thereby, a robot that was employed in the role of moderator displaying performance reinforcing (task-focused) behavior increased group cohesiveness [Short and Mataric 2017]. The study indicates that the intuitively contradicting results are leveraged by the robot addressing participants more evenly when displaying task-focused behavior. To improve human–human collaboration among children, relation-reinforcing utterances have been found to enhance the perception of team performance [Strohkorb et al. 2016]. But neither task-reinforcing nor relation-reinforcing robot behavior was found to influence short-term group cohesiveness.

Addressing the problem of inclusion of an outgroup participant, Sebo et al. [2020] explored different strategies on how an SR could be employed to support the process of inclusion. To support the inclusion of children newly arrived in a country, Gillet et al. [2020] developed a robot-mediated music-mixing activity that allows the robot to perceive group dynamics and act upon them. The music-mixing activity is shown in Figure 17.3.

As an important factor in groups, the level of trust in a mixed human–robot group has been found to be influenced by a robot verbally expressing vulnerability. This expressed vulnerability produced a ripple effect throughout the group that
Figure 17.3 Different works investigated how a robot could influence the dynamics of a group, in this case the robot Cozmo by Anki mediates an interaction among children and aims to foster collaboration and inclusion [Gillet and Leite 2020].

increased trust-related behaviors within the group [Strohkorbe Sebo et al. 2018] and improved conversational dynamics [Traeger et al. 2020].

Work on conformity and group social influence has also inspired efforts on prompting human bystanders to intervene in robot abuse [Tan et al. 2018, Connolly et al. 2020]. This line of work has shown that the reactions of the abused robot can influence how much bystanders perceive adversarial actions toward robots as mistreatment. Further, it suggests that emotional group robot responses can increase bystander interventions in comparison to when they ignore the abuse. Figure 17.4 illustrates examples from this line of work.

Figure 17.4 Two scenarios in which Cozmo robots by Anki aim to influence a human bystander such that (s)he intervenes to stop robot abuse by a confederate. Adapted from Tan et al. [2018] and Connolly et al. [2020].
17.3.2 Automatic Perception of Group Dynamics

For SIAs to be successfully deployed in multiparty social settings, they need to be aware of their surrounding environment and social context [Jung and Hinds 2018].

The F-formation theory discussed in Section 17.2 also inspired work on automatic group perception, both from a model-based [Vázquez et al. 2015b, 2017] and data-driven perspective [Hedayati et al. 2019, Swofford et al. 2020]. These methods demonstrated automatic detection of F-formations involving humans and robots and are illustrated in Figure 17.5. One of the challenges of group perception is feature extraction, especially when considering real-world environments. In this respect, Mead [2016] proposed a framework including different features (individual, physical, and psychophysical) that can be automatically extracted and used to recognize proxemics and other social behaviors in HRI, for example, interaction initiation and termination.

From a perception perspective, it is also important to recognize addressees in groups. Methods have leveraged sound source localization [Nakadai et al. 2008], visual focus of attention [Sheikhi and Odobez 2012, Sheikhi et al. 2013], and combinations of this data [Okuno et al. 2001, Vázquez et al. 2016]. Further, fusion and tracking of participants in interaction with SRs and IVAs has been investigated for the purpose of speaker identification, addressee detection, and dynamic user entrance/leave mechanism [Yumak et al. 2014b]. To enhance multiparty dialogue management, attention management and addressee recognition can be enhanced by observing lip movement and gaze to successfully understand the current addressee in the dialogue [Richter et al. 2016]. Further, Traum and Morency [2010] apply real-time visual processing to enhance a dialogue model of multiparty communication between humans and IVAs. Visual processing focuses on head orientation, nods, and shakes to influence a multilayer dialogue model, including addressee identification, turn-taking, social affiliation, and grounding.

Figure 17.5 Conversational group detection in HRI. Adapted from Swofford et al. [2020]. The agents that have the same color are estimated to be part of the same group. The opacity of the red lines connecting the agents represents the likelihood of them belonging to the same group. Left: a situation in which all five people that interact with the robot are estimated to be part of its group. Right: an individual interacts with the robot while other people observe the interaction nearby.
Considering a different aspect of perception, interest has been developed in perceiving the dominant human in an interaction based on different sensor modalities. Skantze [2017] showed that analysis of dialogue with focus on the amount of speaking and turn-taking behaviors could help identify dominance early on in a conversation-based game. Strohkorb et al. [2015] utilized visual data only to predict the most dominant child in a group interaction.

### 17.3.3 Generating Behavior in Groups

The synthesis of small group behaviors involves automatic conversation management synchronizing a range of multimodal behaviors including speech, eye-gaze, gestures, and body positioning across group members in a socially appropriate manner. Locomotion behaviors support movements of individuals within the group, from small position shifts of group members that are natural in real situations, to the larger formation changes required to accommodate a newcomer to a group or coalesce when an existing member leaves. Generation also encompasses approach trajectories and join behaviors for newcomers to a group, an important ability supporting multiparty interaction with SIAs in both real and virtual environments.

#### 17.3.3.1 Conversational Behaviors

In conversational group settings, humans use a variety of verbal and nonverbal signals to regulate, coordinate, and otherwise manage their interactions. Figure 17.6 exemplifies situations in which conversational behaviors are explored.

Eye-gaze (see Ruhland et al. [2015] and Admoni and Scassellati [2017] for reviews) is one important nonverbal behavior underlying conversation. A pioneering study by Mutlu et al. [2009] showed the importance of generating human-like gaze and how the appropriate robot’s gaze can shape conversational roles.

![Figure 17.6](image-url) Multiparty interactions involving (left) full-body conversational behaviors that reflect attitudes held by SIAs in virtual environments [©2021 Brian Ravenet] and (right) interactions between humans, SIAs, and IVAs to investigate social presence [Pereira et al. 2014].
Considering groups with more than two people, Vázquez et al. [2017] showed how attentive robot gaze and body orientation should be generated jointly and the importance for the feeling of groupness. Further, more frequent short glances have been found to be more effective than less frequent longer stares for participants to feel the direction of look [Admoni et al. 2013].

Models for conversational behaviors have also been developed for IVAs. Prada and Paiva [2005] developed a model that supports the dynamics of a group of synthetic agents, inspired by theories of group dynamics developed in human social psychological sciences. Autonomous synthetic characters employing these models had a positive effect on the users’ trust and identification with the synthetic group. Pejsa et al. [2017] present computational models of gaze and spatial orientation that a virtual agent can use to signal specific footing configurations, that is, the nonverbal signals that conversational participants use to establish their roles.

Yumak et al. [2014a] investigate interactions between humans, SRs, and IVAs in telepresence setups. Users may control IVAs and/or SRs or they may act autonomously, for example, in the case that their respective avatars, that is, users, leave the interaction. This is accomplished by means of an architecture that tracks multiple users via audio-visual sensors and feeds the fused data into a dialogue manager that in turn generates virtual human and robot behaviors [Yumak and Thalmann 2013].

Ravenet et al. [2015] propose a model for the generation of nonverbal behaviors supporting the expression of interpersonal attitudes for turn-taking strategies and group formation in multiparty conversations among IVAs. Figure 17.6 (left) demonstrates a generated group interaction from this line of work.

More recently, de Coninck et al. [2019] employed a data-driven approach to automatically generate nonverbal behaviors for virtual characters during group interactions. Dynamic Bayesian Networks have further been used to establish associations between conversational state and nonverbal behaviors by analyzing the CMU Panoptic dataset [Joo et al. 2017].

In respect to empathy within groups, Alves-Oliveira et al. [2019] explore how the perception of the emotional climate can inform the generation of appropriate empathic behavior toward the group.

The selection of social actions by a robot in unstructured multiparty encounters was shown to be more successful and efficient when learned as an action selection policy through reinforcement learning [Keizer et al. 2013]. When treating these interactions in a task-based manner, knowledge-level planning has been shown to be promising [Petrick and Foster 2013].

A further line of research explores how an SR can be part of joint action with a group of humans, for example, dancing in a group. Anticipatory action planning
was necessary to allow a robot to join a jazz combo [Hoffman and Weinberg 2011]. Iqbal et al. [2016] found that perceiving high-level human behavior to anticipate human group motion is advantageous when generating motion for joint actions with humans. The behavior of the robot or multiple robots when joining a joint action has further influence on the group dynamics specifically if the two robots generate their motion according to different paradigms [Iqbal and Riek 2017b].

### 17.3.3.2 Locomotion Behaviors

Locomotion is a desirable capability for robust situated multiparty interactions in which artificial systems are expected to be mobile, active conversational participants that adapt to humans rather than static, passive systems. Social-aware navigation (see Charalampous et al. [2017] for survey) enables agents to navigate in the environment so that they not only establish the fastest path to a goal but also respect other characters as social entities.

Thus far, social-aware methods have been mostly applied to the socially acceptable navigation of robots. Sisbot et al. [2007] proposed a human-aware robot motion planner to generate a safe path by considering the human position, gesture, and field of view. Gao et al. [2017] and Pokle et al. [2019] proposed approaches that combined classical planning with modern deep learning techniques to enable SRs to adapt to dynamic human environments. Satake et al. [2013] presented a method for a robot to approach people who are walking through the environment. Other social-aware navigation systems consider static human groups. Truong and Ngo [2018] proposed a framework to enable an SR to approach a human group safely and socially. Yi et al. [2015] presented a cost map based on the distance for mobile pedestrians and static groups. Gómez et al. [2014] extended a fast-marching algorithm to navigate a robot for engaging a group of people. Social-aware navigation methods have also been applied to virtual agents. Pedica and Vilhjálmssson [2018] simulated human territoriality while navigating a virtual character toward small groups. A recent work [Yang and Peters 2019c] proposes a social-aware navigation system capable of moving an agent through an environment that contains both static and moving virtual groups.

Repiso et al. [2020] used an adaptive side-by-side model so that a robot could autonomously accompany a group of people walking.

### 17.3.3.3 Approach Behaviors

Several studies have been carried out that specifically concern the approaching behaviors of newcomers into small free-standing conversational groups. Examples for these approaches are given in Figure 17.7. Ramírez et al. [2016] adopted inverse
reinforcement learning, involving several participants demonstrating approaching behaviors for a robot to learn. Samarakoon et al. [2018] designed a method to replicate the natural approaching behaviors of humans. In a recent work, Gao et al. [2019] proposed a deep reinforcement learning model to generate robot approaching small group behaviors. Behaviors for approaching groups have also been studied for IVAs. Jan and Traum [2007] presented an algorithm for simulating movement of agents, such as an agent joining the conversation. Pedica and Vilhjalmsson [2012] integrated behavior trees in their reactive method to simulate lifelike social behaviors, including robot behavior for approaching groups. Both approaching and leaving behaviors for virtual characters were considered by Yang et al. [2017]. In this work, a finite state machine is utilized in the transitions between different social behaviors. More recently, Yang and Peters [2019b] proposed a model based on Generative Adversarial Networks to generate safe and socially acceptable trajectories into free-standing conversational groups. The trajectory prediction model considers group dynamics, including the changing position and orientation information of group members as they make position adjustments within a formation, and is intended for application to both virtual agents and mobile robots.

### 17.4 Group Datasets

To inform either the perception of group dynamics or generation of appropriate SIA behavior, datasets capturing different aspects of multiparty interaction have been collected. Relevant datasets can be divided into human–robot and human–human datasets.
Only a few datasets exist that capture multiparty interactions with at least one robot. The *Vernissage Dataset* [Jayagopi et al. 2013] captures interactions of multiple participants with a wizarded NAO robot. The *UE-HRI* dataset focuses on spontaneous engagement with a Pepper robot and contains dyadic and multiparty interactions [Ben-Youssef et al. 2017]. In addition, *MHHRI* [Celiktutan et al. 2017], focuses on analyzing personalities and relationships with engagement of human–human (dyadic) and human–robot interactions (triadic). A recent dataset *CongreG8* [Yang et al. 2020] uses full-body Motion Capture (MOCAP) with a focus on approach and joining behaviors for free-standing conversational groups and includes both human–human data and human–robot data that have been applied to both virtual agent and robot group scenarios. Examples for approach behaviors covered in *CongreG8* and other datasets are given in Figure 17.8.

Human–human interaction databases in contrast to multiparty human–robot datasets were extensively reviewed in several surveys including Borges et al. [2013], Stergiou and Poppe [2018], and Zhang et al. [2019]. Unlike datasets containing individual action recordings, human–human interaction datasets, that is, those containing multiple humans interacting, are relatively scarce. The *CMU Panoptic* dataset [Joo et al. 2017] collects 3D full-body motion of a group of people in various social interaction scenarios such as dancing and haggling. The *BARD* dataset [Cancela et al. 2014] focuses on recording human behavior analysis in video sequences with multiple targets in wild environments. Other datasets involving groups collect 2D location information such as body position and orientation information. The *MatchNMingle* dataset [Cabrera-Quiros et al. 2018] is a multi-sensor resource for the analysis of social interactions and group dynamics. The *IDIAP Poster* dataset [Hung and Kröse 2011] is a video dataset with annotations.

![Figure 17.8](image-url) Full-body motion capture data of approach behaviors (left) [©2020] and the approach behaviors from real-life datasets (right) [Yang and Peters 2019a].
of body position and orientation information but also the data for F-formations (conversational groups). Similarly, the Coffee Break dataset [Cristani et al. 2011], the SALSA dataset [Alameda-Pineda et al. 2015], and the Cocktail Party dataset [Ricci et al. 2015] contain F-formation annotations and 2D pose information. The VEIIG dataset [Bandini et al. 2014] collects annotated data with moving groups in a crowd. The Semisynthetic dataset [Yang and Peters 2019a] contains trajectories of individual agents approaching groups based on a social-aware navigation method, and it is used to learn approaching group behaviors. Further, the Elea dataset consists of multiparty human–human of three and four participant groups where the recorded data allows the identification of emerging leadership in a survival task [Sanchez-Cortes et al. 2012]. A large set of meetings captured through multiple modalities are available in the AMI meeting corpus [Carletta et al. 2005]. A more playful setting was chosen in the WOLF dataset [Hung and Chittaranjan 2010], where larger groups engage in the Werewolf game. Audiovisual data of multiparty interaction in three different cultures and languages was captured and annotated in the UTEP-ICT dataset [Herrera et al. 2010].

17.5 Similarities and Differences in IVAs and SRs

Considering previous research, in this section we identify and discuss a number of aspects that IVAs and SRs have in common regarding their ability to work in groups. It is important to note that others have conducted similar analyses before. For example, Gratch et al. [2015] discussed how research implications on virtual humans can impact human–robot teamwork. We begin by discussing the similarities:

- **Driven by user experiences.** Both in the virtual or the physical world, many SIAs are developed to evoke realistic or interactive user experiences. Given this similar goal, it is often the case that common metrics are used to evaluate people’s experiences with these agents. Social presence, for example, has been extensively investigated in both IVA [Hai et al. 2018] and SR [Pereira et al. 2014] multiparty user studies.

- **Application domains.** Many SIAs share the same application domains (e.g., entertainment, gaming, therapy, collaboration), offering opportunities to share computational models between the two sub-communities. While there are some exceptions to this (e.g., navigation algorithms in virtual worlds differ significantly from the ones for real-word environments), many of the

---

perception and high-level decision-making components could be implemented in a way that the target agent embodiment is abstracted. For example, as discussed by Gratch et al. [2015], SIAs natural language dialog systems could be adapted for SRs and vice-versa.

- **Multimodal perception and social behavior generation.** While the sensing capabilities and embodiment of IVAs and SRs might differ, all SIAs benefit from multimodal perception and behavior generation. Multimodal perception is key for making agents more robust and capable of dealing with the complexity of group interactions. For example, perception of multimodal cues has been fundamental for keeping track of turn-taking patterns and advancing multiparty dialog [Traum and Morency 2010, Richter et al. 2016]. Likewise, multimodal behavior generation is important for providing effective communicative signals to users. As in human interactions, enabling SIAs to communicate through multiple modalities can facilitate important communicative processes like grounding [Mehlmann et al. 2016]. But care must be taken when designing multimodal behaviors. Research has shown that different behavior modalities, like body motion and gaze, can influence how people perceive each modality during group interactions [Vázquez et al. 2017].

We continue by discussing the differences:

- **Perception.** While we perceive the physical environment directly via our visual senses, there is a significant level of indirection when rendering the virtual scenes in which IVAs are embedded. Hardware limitations and the choice of virtual camera parameters may result in different perceptual impressions of virtual representations when compared to their real counterparts. For example, differences in distance and social space perception in virtual [Li et al. 2019] and mixed reality [Li et al. 2018] environments may result in different interpersonal distances being observed between humans and IVAs when compared to real-world situations. Differences may also exist in how we perceive a host of other factors, such as the appearance [Peters et al. 2018] and photorealism of the IVAs (see Embodiment below), which may also impact the degree to which we treat IVAs as social entities. Especially for multiparty interactions, considering these effects combined with how humans perceive groups as entities, that is, group entitativity, seems important for creating perceptually sensitive social models of behavior [Bera et al. 2018] transferable between the virtual and real worlds.
Mobility. There are extra considerations for SRs if they are to be mobile to any degree in their environments. Especially, the priority for SR movement algorithms is that they should be safe and efficient when operating in the vicinity of humans. Physical multiparty interactions, which imply numerous humans and SRs interacting in close proximity to each other, are therefore especially challenging. In contrast, IVAs have direct access to the state of the virtual environment through a world database and collisions with humans are only significant in relation to the visual plausibility of the simulation as they do not have any physical consequences.

Embodiment. A core difference between SRs and IVAs is their embodiment: real versus virtual. The impact of embodiment on social presence in multiparty interactions [Shamekhi et al. 2018] is especially significant due to how social presence pervades interaction, from proxemics to attention behaviors [Goffman 1963]. While favorable effects of social presence have been attributed to physical embodiments, the role of physical presence in the process warrants investigation [Li 2015, Thellman et al. 2016]. For IVAs, the photorealism of the embodiment impacts self-reported impressions of social presence [Zibrek and McDonnell 2019], although questions remain as to what degree coinciding behavioral effects, such as those observed in human proxemic behavior, can be achieved based on photorealism improvements alone.

Gaze. Perception of mutual gaze displayed by SIA in multiparty interactions is of specific importance as it is connected to successful turn-taking. When considering IVAs that are displayed as 2D agents on a screen in a physical environment, AL Moubayed et al. [2012] argue that the Mona Lisa effect has stronger impact in the interaction with multiple users. They compared a displayed 2D agent to a 3D back-projected robotic head and found that the 3D agent was perceived as less confusing in turn-taking. Further evidence on the importance of physical movement on gaze was found by Vázquez et al. [2017], where the robot’s body motion could help to convey the gaze behavior. More recently, a more subtle change in eye gaze display was found influential by Kinoshita et al. [2017], where convex eyes could direct gaze more accurately and hollow eyes were correlated with a broader gaze cone.

17.6 Current Challenges and Future Directions

In this section, we discuss some of the current challenges in multiparty interaction between humans and SIAs, along with future research directions.
From one-to-one to one-to-many. Most of the previous research on computational models for perceiving and generating social behavior in SIAs has focused on dyadic interactions of one agent and one user. However, group interactions tend to be more complex [Traum 2004], and the number of people around the agent not only affects how the agent should behave but also how it should perceive its environment. So far, little is known about how data-driven models perform when they are tested in a group size different than the one they were trained on, and most data-driven perceptual systems for human–agent interaction rely on data collected in the same context where future interactions are likely to occur. While previous work has shown that a disengagement classification model trained with group data generalized better to individual participants than the reverse [Leite et al. 2015], further research is needed to confirm that the same findings apply to other types of multiparty perception and decision-making systems.

Formation changes and situated interactions. Real-world multiparty interactions involve a degree of mobility of participants due to small shifts in the positioning of group members who are never totally static. Moreover, explicit formation changes, which may be caused by members joining and leaving the group, or even a change in the attitudes or the focus of attention of the group, can bring additional challenges. Dynamic positioning behavior therefore needs to be considered in conjunction with full-body behaviors and conversational management since current research typically assumes static SIAs in multiparty situations. This places a higher priority on understanding the impact of environment constraints on human proxemics behaviors, in addition to developing artificial models capable of better understanding their spatial environments so that they can account for such changes while also solving challenging locomotion problems.

Individual and group adaptation. SIAs developed for multiparty interaction need to find an optimal balance between adapting to an individual or the whole group. When considering individual adaptation within a group, the behavior of other group members can still be useful for better responding to the individual. Pioneering work in this direction by Mou et al. [2019] has shown that group information can be used, for example, to improve the accuracy of recognizing an individual’s affective state in the group. However, despite some efforts in this direction, questions such as how to accommodate for social norms, culture, and individuality of group members [André et al. 2020] remain largely unexplored.
Dynamic social environments. Laboratory environments are extremely valuable for controlled human–agent interaction experiments (e.g., for investigating specific system components in isolation), but the way people behave in laboratory conditions is substantially different from that of the real world. If SIAs are to be placed in complex, constrained and/or unstructured social settings, more in-the-wild research is needed in those settings from the early stages of development. As discussed by Jung and Hinds [2018], it is particularly important to understand the impact of the robot (or agent) on the social environment beyond the individual. Multiparty interactions also need to be robust to individuals joining and leaving them, placing additional importance on modeling social active vision mechanisms [Breazeal et al. 2001, Peters et al. 2011] so that those within the group become aware of potential newcomers and are capable of interrupting ongoing interactions in order to allow them to join in a socially appropriate manner.

Choice of metrics. To measure the influence of robot behavior on groups, different aspects have been found of interest to study. For example, works that consider social aspects of group dynamics [Fraune et al. 2017a, Short and Mataric 2017, Strohkorb Sebo et al. 2018, Sebo et al. 2020] have used different questionnaires and interviewing techniques to measure inclusion, cohesion, entitativity, or psychological safety. To further understand different SIA behaviors and their influence in multiparty contexts, new valuable insights could be gained by identifying and utilizing standardized metrics for varying multiparty human subject studies. Abrams and Rosenthal-von der Pütten [2020] take a first step in this direction by discussion aspects of cohesiveness, entitativity, and group identification and how to measure these aspects. However, further challenges might arise in specific group context, for example, in cases where language is no option, for example, among young children or children with varying mother tongues. So far, no validated methods exist that could measure, for example, the cohesiveness of a group other than through questionnaires or interviews. The development of language-free tools could further give insights into group dynamics in special target groups or generally allow for indirect assessment that has been found to give valuable additional insight for other metrics (e.g., trust [Glaeser et al. 2000]).

Replicability of results. As discussed by Iqbal and Riek [2017a], the community still lacks an infrastructure to support replicability that is of specific importance in more complex environments such as multiparty interactions. Therefore, it is difficult to explore the effects of different kinds of robots in
comparable situations. Recently, Jung et al. [2020] proposed a task that allows the studying of the effect of resource distribution pursuit by a robot and that might allow comparing the effects of different robots and robot behaviors across research groups.

### 17.7 Summary

In this chapter, we addressed multiparty interactions with SIAs. We provided an overview of the common methods and approaches from social psychology that can be useful for defining the scope and understanding of group behaviors. We then reviewed existing works addressing different aspects of multiparty interactions. Among IVAs, the believability of these agents and how they act in groups has been discussed. When considering the interaction of SIAs and humans in groups, their behavior has been shown to affect the group dynamics and attitudes humans develop about SIAs. Further, SIA behavior can be used to explicitly influence the dynamics of a group and their interaction. However, to be able to interact in groups, SIAs need to be able to firstly perceive the group and its dynamics. To further allow autonomous interactions, SIAs need to be capable of generating appropriate behavior. We discussed how the F-formations, the group’s focus of attention, the addressee in a conversation, and interpersonal dominance can be detected. Further, we reviewed the generation of appropriate gaze behaviors and socially aware motions. To help the perception and generation of group behaviors, we presented a list of relevant datasets. Similarities like multimodal perception of groups and their dynamics, high-level decision making, and multimodal behavior generation offer opportunities for the two communities to find symbioses. There are still many open research directions in this field. We ended with a discussion of current challenges and future research directions such as coping with changes in group dynamics and the applicability of models and methods from individuals to groups or vice versa.

### References


Chapter 17  Multiparty Interaction Between Humans and Socially Interactive Agents


Chapter 17  Multiparty Interaction Between Humans and Socially Interactive Agents


Chapter 17  Multiparty Interaction Between Humans and Socially Interactive Agents


18

Adaptive Artificial Personalities

Kathrin Janowski, Hannes Ritschel, and Elisabeth André

18.1 Motivation

People are known to ascribe human-like social properties to computer systems [Reeves and Nass 1996]. Users shout at the screen when they are frustrated with the software, while science fiction authors keep envisioning emotional androids and sarcastic AI assistants. Various start-up companies promise to bring artificial beings with endearing personalities into our homes, such as EmoTech’s Olly1 whose adaptive personality was the most emphasized feature in advertising the product.

Socially intelligent assistants, be they disembodied voices or physical robots moving through the shared environment, are expected to become a common sight, especially in the context of care [Janowski et al. 2018]. Equipping such agents with a compelling personality is one important step toward establishing a relationship between human and machine [Breazeal 2004], fostering trust and encouraging long-term use after the novelty effect wears off.

Personality in this context means a person’s disposition to respond to certain events in a particular manner [Argyle and Little 1972]. Examples from theater [Laurel 1997] or animated films [op den Akker and Bruijnes 2012] have taught us that a character’s reactions to events and how they line up with their goals are what ensures the engagement of human observers. Laurel [1997] argues that giving an interface human-like dispositions helps humans to predict its behavior. These dispositions, in turn, are expressed through and inferred from behavior cues such as interruptions [Rogers and Jones 1975] and gaze behavior [Argyle and Cook 1976].

In order to be accepted, the agent’s personality must be tailored to its role as well as the user’s preferences for the agent’s behavior [Laurel 1997]. Moreover, Aly and Tapus [2016] point out that adapting a robot’s personality to the human’s profile makes interaction more engaging. However, personality preferences are diverse. Demographic background may influence whether a given phrasing is perceived as convincing or polite [Hammer et al. 2016], while the user’s own personality traits may lead them to prefer similar traits in some cases [Bickmore and Cassell 2005, Bernier and Scassellati 2010] and opposite traits in others. Furthermore, there are also insights indicating that the task context plays a key role for whether a similar or opposing personality is preferred [Joosse et al. 2013].

But since there is no one true approach for configuring personality just right, enabling a computer system to dynamically adapt to its user is necessary for ensuring effective collaboration between humans and machines. This is why this chapter will cover both configuration before the interaction and adaptation at runtime.

In this chapter, we will consider both virtual agents and social robots. Research in these two fields overlaps greatly since the personality models and adaptation approaches we will describe can be used for both graphics-based and physically embodied agents. In fact, there are only two key points where the agent types differ. Unlike virtual characters, robots are physically limited in their animation capabilities, for example, because their motors can only move at a certain speed or because some postures might make a bipedal robot fall over. The other difference is the physical presence in a shared environment, which may impact the perceived social presence of a robot. However, with recent developments in 3D displays and virtual and augmented reality hardware, we expect this difference to become less pronounced as virtual agents gain similar potential to inhabit the same world as their users.

The chapter is structured as follows. First, Section 18.2 will introduce the most important psychological concepts. We will outline commonly used models for personality and interpersonal stance as well as theories about interpersonal compatibility. Next, Section 18.3 will describe approaches for implementing adaptive personalities. There, we will summarize ways to express a particular personality through an agent’s behavior and look at different approaches for adapting said personality to the user. After that, Section 18.4 will look at common approaches for evaluating user-adaptive systems. Section 18.5 will then give an overview over the historical development in the field of personality adaptation, and Section 18.6 will look into current challenges and related future directions. Finally, Section 18.7 will summarize and conclude the chapter.
18.2 Psychological Background

For enabling a computer to reason about the personality of the agent and/or the user, the abstract concept needs to be quantified. Psychological literature provides several well-established models for describing personality and the related interpersonal relationships in dimensional terms. This section will present a selection of models that have been successfully applied in human–computer interaction research.

Furthermore, there are several theories regarding the compatibility between and preference for particular personality traits. They build on the aforementioned personality models and will therefore be presented after them.

18.2.1 Personality and Interpersonal Stance

The term Personality refers to an individual’s behavior patterns that can be observed in a wide range of contexts or their disposition to respond in a certain way when they find themselves in a particular situation [Argyle and Little 1972]. It manifests in their behavior toward other people, and as this section will explain, links exist between the models used to describe personality and interpersonal stance.

18.2.1.1 Five Factor Model

One of the most widely used frameworks for describing personality is the so-called Five Factor Model, also known as the Big Five [McCrae and John 1992, Mehrabian 1996b]. According to this model, personality is defined by the following five dimensions:

— Openness: This factor encompasses curiosity, creativity, and intellectuality. Open-minded people have a wide range of interests and think in unconventional ways, whereas closed-minded people are generally conservative and unimaginative.

— Conscientiousness: This factor is concerned with disciplined and responsible qualities. People who score high in Conscientiousness are generally dutiful and reliable, well-organized, and thorough. They are efficient and productive rather than lazy and self-indulgent.

— Extraversion: This factor is associated with outgoing and assertive behavior traits. Extraverted people tend to be sociable, talkative, expressive, and active, whereas introverts are more reserved and quieter.

— Agreeableness: This factor is related to getting along well with others. It encompasses qualities such as being kind, compassionate, and forgiving,
as well as generous and trusting. In contrast, disagreeable persons tend to criticize others and be cold-hearted or inconsiderate.

— Neuroticism: This factor is also known as Emotional Stability. It describes a person’s tendency to experience negative affect such as distress and anxiety, act impulsively or change moods quickly and frequently. Emotionally stable people, however, are calm and relaxed.

### 18.2.1.2 PAD Temperament Model

An alternative approach to defining a person’s general behavior tendencies is Mehrabian’s Pleasure–Arousal–Dominance Model, also called the PAD Temperament Model [Mehrabian 1996a]. As Mehrabian explained it, the “temperament” of a person corresponds to their average emotional state across a representative sample of situations, in contrast to the temporary emotional states that change frequently throughout the day. The PAD space is defined by three dimensions that relate to the Big Five factors as follows [Mehrabian 1996b]:

— Pleasure: This trait describes a person’s tendency to experience positive versus negative emotions. It can be calculated as $0.21 \times \text{Extraversion} + 0.59 \times \text{Agreeableness} - 0.19 \times \text{Neuroticism}$.

— Arousal: This trait describes a person’s responsiveness to stimuli and the time it takes for them to calm down. According to Mehrabian, it equals $0.15 \times \text{Openness} + 0.30 \times \text{Agreeableness} + 0.57 \times \text{Neuroticism}$.

— Dominance: This trait describes how much a person feels in control of their life. It equals $0.25 \times \text{Openness} + 0.17 \times \text{Conscientiousness} + 0.60 \times \text{Extraversion} - 0.32 \times \text{Agreeableness}$.

In computer science, PAD space is also used to model a character’s mood or short-term emotions since these concepts are closely coupled with personality [Gebhard 2007]. Any of them can shape a character’s behavior at a given time, so having a common framework for emotions and personality is important for generating consistent system reactions. For more information, see Chapter 10 on “Emotion” [Broekens 2021] of volume 1 of this handbook [Lugrin et al. 2021].

### 18.2.1.3 Interpersonal Circumplex

Attitudes toward other persons are commonly modeled with the Interpersonal Circumplex [McCrae and Costa 1989, Horowitz et al. 2006, DeYoung et al. 2013]. It is defined by two axes, Status and Affiliation.
— Status: This axis ranges from *submissive* to *dominant* and is usually displayed as the vertical dimension. It is also known as *Agency* and describes a person’s tendency to act according to their own will.

— Affiliation: This axis ranges from *cold* to *warm* and is placed horizontally. Also known as *Communion*, it describes a person’s social closeness to other people.

These two dimensions have also been shown to be related to the personality traits Extraversion and Agreeableness [McCrae and Costa 1989, DeYoung et al. 2013]. As explained in Section 18.2.1, high Extraversion implies sociability and therefore closeness to people but also assertive and therefore dominant behavior tendencies. In a similar manner, Agreeableness represents social compliance, which is a combination of warm-hearted and submissive behavior. According to literature [McCrae and Costa 1989, DeYoung et al. 2013], these two personality traits form an alternate pair of axes that is rotated about 30–45° relative to status and Affiliation, as shown in Figure 18.1.

### 18.2.1.4 Politeness Theory

Another important concept for interpersonal behavior is politeness. In their *Politeness Theory*, Brown and Levinson [1987] assume that every human has two basic *wants* concerning their public identity, also called their *face*. Those wants, in turn, appear related to the Interpersonal Circumplex [Oakman et al. 2003].

— Negative Face Want: People desire to be autonomous in their actions. This resembles the Status dimension of the Interpersonal Circumplex that represents a person’s tendency to act autonomously.

![Figure 18.1](image-url) The two pairs of dimensions that define the Interpersonal Circumplex. *Left/solid:* Status and Affiliation. *Right/dashed:* Extraversion and Agreeableness.
— Positive Face Want: This concerns the desire to have other people’s approval and know that they share one’s own goals. Like the Affiliation dimension, it implies group membership and a social bond with others.

Different phrasings can be categorized according to the face threats they present or avoid. Accordingly, those that minimize threats to the hearer’s positive face are said to use positive politeness, and those that minimize negative face threats use negative politeness. For example, Johnson et al. [2005] examined eight different phrasings in English and German and found that the categories listed in Table 18.1 are interpreted as shown in Figure 18.2.

18.2.2 Theories About Interpersonal Compatibility

There are two major theories concerning the compatibility of individuals based on their personality. While the similarity attraction theory suggests that people would be most compatible with similar personalities (“birds of a feather flock together”), the complementarity theory suggests that people are more compatible with dissimilar personalities (“opposites attract”). Additionally, there are approaches to reconcile both ideas by considering the underlying goals that shape either person’s behavior and appraisal of events.

18.2.2.1 Similarity Attraction Theory

Similarity attraction is most often researched with regard to attitudes, but comparable effects have been found for personality traits as well [Montoya and Horton 2013].

Moon [2002] examined the effect of dominance on the persuasiveness of computer-generated messages, depending on the dominance of the human interacting with it. Dominant messages were phrased as assertions and direct

<table>
<thead>
<tr>
<th>Phrasing</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct command</td>
<td>Drink some tea.</td>
</tr>
<tr>
<td>indirect suggestion</td>
<td>The system is asking you to drink some tea.</td>
</tr>
<tr>
<td>request</td>
<td>I would like you to drink some tea.</td>
</tr>
<tr>
<td>system’s goal</td>
<td>I would drink some tea.</td>
</tr>
<tr>
<td>shared goal</td>
<td>We should drink some tea.</td>
</tr>
<tr>
<td>question</td>
<td>How about drinking some tea?</td>
</tr>
<tr>
<td>suggestion of user’s goal</td>
<td>You would probably like to drink some tea.</td>
</tr>
<tr>
<td>Socratic hint</td>
<td>Did you think about drinking some tea?</td>
</tr>
</tbody>
</table>
commands while expressing a high confidence (given in percent) in the computer’s claim. Submissive messages, in contrast, used questions and suggestions in combination with low confidence levels. The results showed that dominant users were more likely to change their ranking of different cars when the computer used a dominant message style to contradict them. They also rated the information quality higher. When asked to judge the computer’s level of expertise, all participants interacting with one matching their own dominance rated the computer more favorably. A follow-up experiment showed that participants rated music samples, cartoons, and health tips more favorably when the computer’s presentation style matched their own dominance level. An earlier study by Moon and Nass [1996], also using verbal cues and confidence levels to manipulate dominance, showed that people preferred computers whose dominance level became more like theirs over time, rather than those staying the same or developing in the opposite direction.

Andrist et al. [2015] varied a robot’s gaze behavior to indicate either an introverted or extraverted personality. They then had people collaborate with the robot on a puzzle task, leaving the subjects to decide when they wanted to stop. Their results showed that people without intrinsic motivation to solve the puzzle spent significantly more time collaborating with a robot whose Extraversion level was similar to their own and chose to solve more puzzle tasks together.
18.2.2 Complementarity Theory

Behaviors that are elicited by one reflecting a particular combination of status and affiliation (see Section 18.2.1) are called complementary [Estroff and Nowicki Jr. 1992, Markey et al. 2003].

Markey et al. [2003] observed the behavior during three types of interaction (encounter, collaborative, and competitive) between unacquainted dyads. They confirmed that the model that best explained their observations was the one where behaviors elicited reactions that were similar in affiliation but opposite in status.

Estroff and Nowicki Jr. [1992] examined whether complementary pairs of people were more effective in performing puzzle or word finding tasks than anti-complementary pairs. They used the definition that complementary people would be similar with regard to the affiliation dimension but opposite with regard to status. For anti-complementary pairs, the dimensions were switched. Their study results showed that complementary dyads completed a significantly greater part of the puzzle. For the word finding task, the tendency was the same but not significant.

Similar patterns could be confirmed for the Big Five personality trait of Extraversion. As Section 18.2.1 explained, high Extraversion is closely related to high dominance. Liew and Tan [2016] found that learners reported more positive emotions and learning motivation when interacting with a virtual tutor whose Extraversion was the opposite of their own. Extraverted learners also found the introverted agent more likeable and trustworthy.

18.2.3 Interpersonal Goals

Since neither similarity attraction nor complementarity are universally applicable, other researchers have focused on the underlying interaction goals and personal motivations.

For instance, Tett and Murphy [2002] proposed that “people prefer co-workers who let them be themselves,” which would explain both the similarity attraction and the complementarity theory. According to them, expressing one’s personality traits is a fundamental human need (which is also supported by the face wants of the politeness theory, see Section 18.2.1). Consequently, similarity is preferred for agreeableness or affiliation because it enables people to express mutual closeness, whereas complementarity is preferred for dominance when one person wants to lead and the other wants to follow.

Horowitz et al. [2006] explained that a person's goal can be related to status or affiliation and that frustration arises when the other party misinterprets that goal. One example they give is that a person might talk about a problem seeking either
guidance (status-based motive) or comfort (affiliation-based motive). The interaction breaks down if they receive something different from what they expected.

Reisz et al. [2013] examined how a person's goals relate to their personality traits. They found that people formulate goals that lead to positive affect or compensate for negative affect, especially if those goals require great or continued effort. For example, open-minded people named the goal of learning a new skill, whereas introverted people named the goal of making new friends. Goals like “use time more effectively” were given by people low in conscientiousness, whereas “reduce stress” was listed by those scoring high on that trait and on neuroticism. This explanation implies that people are compatible with others who reinforce their positive traits (such as two friends high in agreeableness) or compensate for a deficit in one trait (such as introverted people being friends with less introverted persons).

18.3 Implementing Adaptive Systems

There are two main tasks when implementing personality-adaptive systems: One is the development of an adaptation mechanism, and one is the mapping between personality traits and observable surface cues. We will start with the latter because adaptation mechanisms require an understanding of the parameters that will be modified over time.

18.3.1 Expressing Personality with SIAs

Personality can be expressed with multiple communication channels due to the flexibility of today's virtual agents and social robots. These expressions are commonly adapted from psychological findings about human communication, such as those summarized by Knapp et al. [2014].

First of all, the linguistic content of utterances can be used to express personality. For example, Moon and Nass [1996] use scripted language-based cues (strong language, assertions, and commands to express high confidence versus weaker language, questions, and suggestions to express low confidence) to make the computer appear either dominant or submissive. The PERSONAGE natural language generator by Mairesse and Walker [2010] realizes linguistic variation according to the Big Five personality dimensions in the context of restaurant comparisons, which is used by Aly and Tapus [2016] with a social robot. Ritschel and André [2017] and Ritschel et al. [2017] use an adaptive, generative approach for a storytelling robot with varying degree of extraversion. Harrison et al. [2019] and Hu et al. [2018] present neural generation methods with variations in personality (Big Five) and improved stylistic control in the context of task-oriented dialogs. Paralinguistic behavior is of central importance to present the linguistic content in a convincing manner. For example, synthesized voices indicate extraversion with faster speech
rate, higher pitch and volume, and wider range of tone; in contrast, slower speech rate, lower pitch and volume, and smaller range of tone indicates introversion [Reeves and Nass 1996].

Kim et al. [2008] and Hu et al. [2015] manipulate a robot's gestures to express extraversion and introversion. Their agents' gestures vary in amplitude, speed, frequency, and more. Similarly, Isbister and Nass [2000] express this dimension with a virtual character with posture. This includes widespread limbs, wider movement, and movement toward the observer for extraversion, and limbs closer to the body and restrained gestures for introversion. Generative approaches [Hartmann et al. 2005] are important for creating this type of behavior flexibly, without the need of hardcoding and scripting every detail of the interaction. Given enough personality-related training data from audio corpora, future applications may also make use of deep learning-based speech-driven gesture generation [Kucherenko et al. 2020].

In a similar vein, gaze behavior can provide clues about an agent's traits and attitudes. For instance, Bee et al. [2009] showed that the orientation of a virtual character's head and eyes influenced the dominance level expressed by the agent. Later, Arellano et al. [2011] showed that agents turning their head upwards were perceived as more extraverted and less agreeable than those lowering their gaze.

Due to the successful personalization [Ritschel et al. 2019a] of artificial agents' nonverbal sounds [Bethel and Murphy 2008] with regard to expression of emotion and intentions, generating such sounds during runtime [de Gorostiza Luengo et al. 2017] is of high interest to shape socially interactive agents' personality beyond the traditional verbal and nonverbal modalities.

Studies with embodied conversational agents have shown that humans also interpret their turn-taking patterns in terms of personality and interpersonal attitudes. In general, starting to speak later and yielding to interruptions sooner are signs of an introverted and submissive character, whereas agents that interrupt the other party and continue to talk over them appear extraverted and dominant. These relationships between speech timing and personality were confirmed by ter Maat et al. [2011] and Cafaro et al. [2016], as well as by Janowski and André [2019]. Additionally, Gebhard et al. [2019] showed that the same is true for interactive human–agent conversation.

18.3.2 Adaptation Approaches
Figure 18.3 gives an overview of adaptation and expression of personality in the context of SIAs. In general, one can distinguish between adaptive and adaptable social agents [Schneider and Kummert 2021]. The former autonomously attempts to adapt to the user, and the latter allows the user to actively change the
agent’s behaviors. In the context of personality, the term *adaptation* is used for manipulating an agent’s behaviors in order to express a personality that is best for the individual user. This is realized by either configuring the robot’s personality profile according to the similarity or complementarity attraction principle (see Section 18.2.2), or by tweaking parameters during runtime. Typically, human input is provided either as data upfront to the interaction or online during the interaction. The former is often realized, for example, by filling out a questionnaire or self-report to assess the user’s own personality profile. The latter includes approaches for real-time processing of the user’s social signals or task-based information, such as user performance in a goal-oriented interaction. For example, paralanguage and nonverbal behavior are used for automatic personality recognition [Vinciarelli and Mohammadi 2014]. Social signal processing uses speech, gestures, pose, facial expression, and more to fulfill this task based on human input. This is done to either estimate the human’s personality profile or to get other information about the user. For example, Carbonneau et al. [2020] use feature learning and spectrogram analysis for estimating personality via speech, and Salam et al. [2017] use a fully automatic system to predict the human Big Five personality traits based on nonverbal behavioral cues.

Social signals can also be used to estimate human engagement. Engagement [Oertel et al. 2020] is of central importance for designing systems that are able to adapt to their users’ characteristics [Salam et al. 2017]. For example, Mancini et al. [2019] use engagement detection to form and maintain a virtual agent’s impression of warmth and competence. Since the adaptation of the robot’s personality profile can make interaction more engaging (see Section 18.2.2), engagement is also investigated in the context of human and robot personality, for example, in
Adaptive Artificial Personalities

Celiktutan et al. [2017]. In combination with adaptation, Ritschel et al. [2017] and Ritschel and André [2017] use it to optimize a social robot’s extraversion based on the user’s reactions.

18.3.2.1 Decision-theoretic Reasoning

The situations that an agent encounters are not always deterministic. Users might respond positively to a certain personality expression most of the time but feel irritated by it when they are in a certain mood or busy with a particular task. Therefore, decision-theoretic approaches consider probabilistic outcomes that are linked to a person’s internal affective or cognitive states. Since the latter are not directly observable, they need to be inferred from surface behavior cues, for example, using a Bayesian Network [Ball and Breese 2000, Conati 2013].

Ball and Breese [2000] used such a Bayesian network to model the causal relationship between different surface behavior features, the currently experienced emotion (corresponding to the first two dimensions of the PAD model), and the interpersonal stance. This model was then applied both for inferring the user’s current affective state and for modifying the agent’s behavior. The considered modalities included, among others, choice of words, acoustic features such as pitch and volume, timing information, and gestures. Furthermore, the authors pointed out that such a model can be easily extended by creating a new chance node for the added modality and setting the conditional probabilities for observing certain behavior depending on the emotional state and interpersonal stance.

Conati [2013] described an approach for implementing an affect-sensitive virtual butler that would adapt the timing of its actions to the user’s state. In this case, a Dynamic Decision Network (a Bayesian Network augmented with temporal information and decision nodes) serves for inferring the user’s current emotion not only from their behavior but also from background knowledge about their personality-related goals. This enables the system to distinguish between emotions that are similar in valence but are caused by different events and therefore require different responses. For example, if the user felt shame over making a mistake, the virtual butler would need to bolster the former’s confidence, whereas if the user felt reproach because of the system’s mistake, the agent would need to apologize. In such cases, a socially interactive agent may need to deviate from its default personality, showing more respectively less assertive behavior to accommodate the current interaction context.

One major ingredient of decision-theoretic systems are the costs and benefits associated with different outcomes, which define each outcome’s utility. The so-called expected utility is the sum of said utilities, weighted by the probabilities that the system action in question will lead to the respective outcomes. This makes it
possible to predict which of the available actions will be the most beneficial in the face of uncertain situations.

For instance, Bohus and Horvitz [2011] had human judges rate the gravity of different turn-taking errors made by a virtual agent in order to create a cost function based on the time that the system chose to wait in different situations. After calculating the system’s confidence that a user had yielded the turn to the virtual agent, this cost function could be used to decide about the ideal waiting time before the agent started speaking. In a similar manner, the virtual butler proposed by Conati [2013] would be able to calculate the expected impact that its timing would have on a user’s mood.

### 18.3.2.2 Reinforcement Learning

In recent years, Reinforcement Learning [Sutton and Barto 2018] has become increasingly popular, also in the context of behavior adaptation for social agents. It is often used for personalization [den Hengst et al. 2020]: exploring and identifying the best agent’s behavior for the individual human interaction partner.

In general, a reinforcement learning agent explores different actions in different states iteratively. For every action execution, it receives a numeric reward. This information gives it an indication of whether the action was expedient or not. Since reinforcement learning solves control problems, the agent needs to balance exploitation (i.e., selecting the “greedy” action, which is the most profitable one according to the agent's experience) and exploration (i.e., selection of sub-optimal actions) when deciding which action to take for the next learning step. In human–agent interaction, feedback for calculating the reward is provided implicitly or explicitly based on different sources of information, including task-related information and human social or biosignals. Explicit feedback includes, for example, ratings via haptic button presses [Ritschel et al. 2019d] (see Figure 18.4),

![Figure 18.4](image_url)  
**Figure 18.4**  
Left: A study participant interacts with an adaptive social robot, using Reeti by Robopec, in Ritschel et al. [2019d]. Middle: The hardware control panel is used for interaction and to give a reward signal for reinforcement learning (adapted from Ritschel et al. [2019d]). Right: Simulation results of a robot with adaptive extraversion in Ritschel et al. [2017].
graphical user interfaces [Ferreira and Lefèvre 2015], tactile [Wada and Shibata 2006, Barraquand and Crowley 2008] or prosodic [Kim and Scassellati 2007] input. Implicit feedback is not provided consciously by the user but either deducted from task-related information or subconscious human social or biosignals. Task-related data is of central importance to estimate the user’s performance in goal-oriented tasks, exercises, or games, such as in Ritschel et al. [2018, 2019b, 2020b]. However, it cannot capture human aspects, such as the user’s behavior, personality, or mood. Thus, interaction distance, gaze and smile [Leite et al. 2011, Gordon et al. 2016, Fournier et al. 2017, Hemminghaus and Kopp 2017], motion speed, timing [Mitsunaga et al. 2008], gesture and posture [Najar et al. 2016, Ritschel et al. 2017], and laughter [Hayashi et al. 2008, Knight 2011, Katevas et al. 2015, Weber et al. 2018, Ritschel et al. 2020a] are used in various contexts as feedback for social agents. Physiological feedback includes ECG [Liu et al. 2008] or EEG [Tsiakas et al. 2018] data. These signals are often aggregated and combined in order to build a user model and calculate reward, for example, based on the human’s estimated affect/emotions [Leite et al. 2011, Gordon et al. 2016, Broekens and Chetouani 2019], engagement [Ritschel et al. 2017, Ritschel 2018, Mancini et al. 2019], curiosity [Fournier et al. 2017], amusement [Ritschel and André 2018, Weber et al. 2018, Ritschel et al. 2020a], and more.


![Figure 18.5](image-url) Contextual bandit/associative search (i.e., several multiarmed bandit problems) are used to adapt a social robot’s linguistic style in Ritschel et al. [2019d].
an agent adapts to the user and the human subsequently adapts to the adapted system.

With regard to personality, Iida et al. [1998] use reinforcement learning for learning personality-related cues for a robot. The human moves through different spatial areas in front of the robot, while each movement is followed by a robot’s action. Each of these actions is rewarded positively or negatively. Tapus et al. [2008] use reinforcement learning to optimize the robot’s personality during exercises in the context of post-stroke rehabilitation therapy. The authors adjust the interaction distance, speed, and vocal content of the therapist robot in order to improve the user’s task performance. Introversion is expressed with gentle, supportive language and low pitch and volume, and extraversion uses challenging language and high pitch and volume. The reward signal is based on the number of exercises performed in a given period of time. Ritschel et al. [2017] use reinforcement learning to optimize a social robot’s linguistic style in terms of extraversion based on the user’s engagement. While the natural language generation approach is inspired by PERSONAGE [Mairesse and Walker 2010], human engagement is estimated based on social signals. An increase/decrease of engagement over time results in a positive/negative reward for the learning agent. As Martins et al. [2019] point out, there is a lack of experiments with adaptation based on a deeper, psychological understanding of users. The authors expect to achieve more user satisfaction and acceptance of social robots when psychological measures, such as user personality, are used for adaptation.

One important aspect in reinforcement learning is the distinction of stationary and non-stationary problems. For example, Figure 18.6 illustrates the values learned by a multiarmed bandit over time based on the feedback it received from the human user. Over time, changes become smaller and smaller (stationary problem), estimating the final values more precisely. As soon as human preferences change over time, it is necessary to use algorithms for non-stationary problems.

![Figure 18.6](image) Estimated values of a robot's learned politeness strategies over time based on the feedback received from a female test person. Solid lines represent the resulting best (red) and worst (orange) strategy. Changes become smaller over time (modeled as a stationary problem).
Typically, a constant, small learning rate is used, which controls to which degree new feedback affects the learned policy to date, i.e., how quickly the agent can react and adjust its knowledge to the human’s new desires.

**18.3.2.3 Neural Learning of Personality-based Behavior Styles**

While previous approaches generated behavior variants by applying procedural manipulations, more recent approaches employ a data-driven generation paradigm where neural networks are trained on large corpora to automatically synthesize stylistic variants for verbal and nonverbal behaviors.

Nguyen et al. [2018] trained a chatbot on TV show transcripts and movie dialogues to learn how to map user utterances on chatbot responses. To this end, they made use of an encoder–decoder setup with an attentional mechanism that consisted of an encoder to process user utterances and a decoder to produce the chatbot answer. An evaluation of the chatbot revealed that the approach was able to learn certain aspects of linguistic style that may be attributed to personality.

The approach by Nguyen et al. [2018] focused on the achievement of stylistic goals without considering the semantic content of utterances. Such an approach is, however, not suitable for task-based dialogue. Oraby et al. [2018] investigated how to train a neural model for task-based dialogue that ensures not only stylistic variation but also semantic fidelity. To acquire a sufficient amount of training material consisting of semantic representations of dialogue acts and matching language output representing different styles, they made use of the PERSONAGE natural language generator by Mairesse and Walker [2010] (see Section 18.3.1). Experiments with various encoder–decoder setups showed the benefit of neural architectures for controlling stylistic variants in task-based dialogue.

While the approaches above aimed to realize explicitly given stylistic goals, Hoegen et al. [2019] made use of neural architectures to implement a chatbot that matches the participant’s conversational style on previously defined variables, such as pronoun use and speech rate. Conversational style matching can be regarded as a kind of social adaptation mechanism where conversational partners align their conversational behaviors to each other [Schiller et al. 2019]. While there is no unique mapping between conversational style variables and personality traits, there are obvious connections. For example, length of utterances and loudness of speech may be an indicator of extraversion.

Neural network approaches have also been exploited for the generation of stylistic variants for character animations. A particular style of motion modulates how a character is perceived and which emotions, mood, and personality are ascribed to it. For example, Smith et al. [2019] trained three separate networks for pose, timing, and foot to enable real-time style transfer. Style transfer helps reduce the amount
of data including combinations of heterogeneous actions and motion variants required to achieve character animations of sufficient quality that enhance the character’s expressivity. Even though the work by Smith et al. [2019] focused on a broader range of affective and non-affective components, the approach bears great promise for generating motion style variants that portray a character’s personality.

18.4 Evaluating Adaptive Systems

Evaluating an adaptive system is not trivial because it needs to be observed over an extended period of time. Furthermore, the frame of reference is subject to change, making it difficult to compare results between, for example, different iterations of a prototype. One way to approach this issue is to evaluate components separately, for example, testing the learning mechanism with simulated users while conducting a perception study to confirm that the generated behaviors convey the intended personality. This section summarizes the most important evaluation techniques.

18.4.1 Simulation

When it comes to online adaptation mechanisms, simulations are used to investigate whether the implemented algorithms work in theory. Since studies with real users are complex, time consuming, and expensive, simulated users are one option to check for algorithmic issues before testing the system in real human-agent experiments (see Section 18.6.1). This is a first important step that should be done in advance since simulations are cheap, repeatable, and easily allow for changes and parameter tweaking. For example, Jain et al. [2018] propose a user simulator architecture for dialogue managers, tailored to socially aware conversational agents, which generates both task and social behaviors.

Typically, a simulated user model or fake user is used to mimic reactions to an agent’s actions, such as in case of reinforcement learning (see Section 18.3.2). For example, the right subfigure in Figure 18.4 illustrates simulation results from Ritschel et al. [2017], where a social robots adapts its extraversion based on the human’s engagement. Depending on what is adapted and which human data serves as input for this process, simulations require a strong level of abstraction and typically cannot emulate real human reactions adequately due to the high complexity. One approach to overcome this issue can be a combination with Wizard-of-Oz studies (see Section 18.4.3), which are used, for example, in spoken dialogue systems [Rieser and Lemon 2011]. Thus, simulations can be used for verifying an adaptation approach’s theoretical functionality and for parameter tuning.
18.4.2 Perception Studies
To evaluate personality-based behavior variations, many researchers use perception studies in which they present study participants with videos of pre-recorded interactions between social agents. Examples can be found in the works of Cafaro et al. [2016], ter Maat et al. [2011], Ravenet et al. [2015], or Janowski and André [2019]. One advantage of this approach is that all study participants observe and judge identical situations. It removes the risk of technical problems during the presentation of the agents’ behavior and avoids interference caused by non-deterministic behavior processes (such as random noise in the gaze direction) that are often implemented to make the agents’ actions appear more natural.

Additionally, using video stimuli allows researchers to reach more participants from more diverse backgrounds because it removes the need for people to come to the laboratory. Instead, perception studies can be conducted over the Internet, using online surveys or crowdsourcing platforms such as Amazon Mechanical Turk.

However, putting the subjects in the role of passive observers can have both positive and negative effects. On the plus side, it allows the users to focus on the agent’s behavior rather than the interaction task. On the minus side, this detachment from the task may lead to different evaluations than those from actually using the system themselves. Berry et al. [2009] found that subjects accepted a scheduling assistant’s sub-optimal suggestions under laboratory conditions because the scheduled events were purely fictional and people were aware that they would not actually need to attend them. The same effect can be expected when users know they do not have to interact with the agent whose behavior they are rating.

18.4.3 Wizard-of-Oz Studies
Robert Jr. et al. [2020] found that the majority of human–robot interaction studies use Wizard-of-Oz setups or hybrid systems where only part of the robot’s behavior is truly autonomous. The most compelling reason for this is that it allows for rapid prototyping of behaviors—for testing the effect of behavior variations without the need to implement them.

It also mitigates problems with input processing. Having a human observe the user’s voice or gesture commands makes it possible to simulate near-perfect recognition of not only the surface behavior but also the semantics behind it. In contrast, automatic recognition often requires enormous amounts of data and computation resources, as evidenced by the fact that state-of-the-art voice interfaces rely on cloud-based services to process the audio input.

However, Wizard-of-Oz studies tend to show an overidealized version of the system, and findings from such studies can be difficult to transfer to autonomous
implementations. Furthermore, special care needs to be taken to ensure that the remote operator (the “wizard”) behaves in a consistent manner toward all participants within a given condition. Unlike an autonomous application, human operators are subject to fatigue and distractions. Methods for supporting these wizards in their task, for example, by automating part of the robot’s behavior, consequently form a research topic of their own.

18.4.4 Autonomous Interactive Systems

Ideally, an agent’s behavior is evaluated during autonomous interaction with real users to prove that the approach is feasible in reality and that the implementation works as expected. This, of course, requires a fully functional prototype, so this type of evaluation is mainly used near the end of a project.

Furthermore, the users, to whom the system needs to adapt, may behave very differently in their regular environment, as observed by Berry et al. [2009] in the article cited above. Because the users accepted sub-optimal recommendations under laboratory conditions, the meeting scheduling agent was improperly trained. The authors therefore suggest deploying working implementations of adaptive systems to end-users in order to have them evaluated under real-life conditions. However, they also point out that this requires the basic application have good usability to begin with so that people will use it long enough to actually observe the adaptation.

Good usability depends on a number of factors. Among other things, the application needs to have reliable input processing (such as accurate speech recognition in a noisy environment), fallback strategies for handling unexpected user input, and an intuitive dialogue structure that guides the user toward their options. Unfortunately, shortcomings in that regard are often only found when the system is confronted with naive end-users. One way to deal with this issue is to incrementally improve the application and reevaluate each new version, but this takes a considerable amount of time and may alienate users in the early stages.

18.5 History/Overview

After more than a quarter of a century of research on virtual agents, it is common belief that computer characters need to be realized as individuals with a distinct personality. However, three decades ago, the idea of crafting a personality for computer systems was considered anything but obvious. Indeed, a panel on drama and personality in user interface design [Mountford et al. 1989] at the 1989 SIGCHI Conference on Human Factors in Computing Systems (see Figure 18.7) introduced the topic with the words: “Of all the things that come to mind when one thinks of computers and user interfaces, drama and personality are among the last.”
The motivation to discuss whether and how to give computers a personality was driven by the promise to create new experiential user interfaces.

In 1995, Nass et al. [1995] posed the provocative question: “Can computer personalities be human personalities?” At that time, first attempts had been made to convey verbal and nonverbal cues by computer agents [Cassell et al. 1994]. However, it was not clear whether people would interpret behavioral cues provided by a computer agent in a similar manner as behavioral cues provided by a human. To shed light on this question, Nass and colleagues conducted a first study with verbal stimuli. The study revealed that simple cues, such as strong versus weak language, already enable a human to attribute a particular personality to a computer agent. They concluded that a minimal set of cues suffices to provide a computer with a personality that increases user liking.

In 1997, the first collective volume on the creation of virtual personalities [Petta and Trappel 1997] appeared. Like the early SIGCHI panel, it made allusions to drama and theatre and presented a variety of approaches that aimed to create synthetic actors with a personality. A representative example is the work by Hayes-Roth et al. [1997] who implemented several scenarios following the metaphor of a virtual theater. Their characters were not directly associated with a specific personality. Instead, they were assigned a role and had to express a personality that was in agreement with this role.

While Nass and colleagues relied on handcrafted stimuli in their personality experiments, researchers in the area of computer animation and natural language generation started to investigate how to automatically create behaviors for virtual agents that portray a particular personality. Parameters related to personality were used to control many aspects of a virtual character’s multimodal behaviors including movement quality [Badler et al. 1997], linguistic style and prosody [Walker et al. 1997], or dialogue policy [André et al. 2000].

While earlier approaches on virtual agents were inspired by work on drama and theatre, research around the start of the millennium was marked by an increased interest in explicitly modeling theories from the psychological sciences, such as the Big Five [McCrae and John 1992] for personality and the Ortony Clore Collins (OCC)
model [Ortony et al. 1988] for emotions. Ball and Breese [2000] treated personality and emotion as unobservable variables in a Bayesian Network and defined model dependencies between these unobservable variables and observable ones, such as linguistic style and facial expressions (see Section 18.3.2). André et al. [1999] presented several projects with virtual agents in which personality and emotions were used as filters to constrain the decision process when selecting and instantiating the agents’ behaviors. Rist et al. [2003] developed a platform for the realization of multiple conversational settings with virtual agents that portrayed through their dialogue behaviors a particular personality specified by the human user beforehand. These early approaches enabled a flexible generation of multimodal behaviors based on given personality profiles. However, it was not possible to adjust the behaviors of the virtual agents based on the interaction with the user even though first ideas on the evolution of a virtual character’s personality were discussed in Rist et al. [2004].

Even though first tools to temporally synchronize behaviors for virtual agents, such as PPP (Personalized Plan-Based Presenter, André et al. [1998]) or BEAT (Behavior Expression Animation Toolkit, Cassell et al. [2004]), were in place, consistency and timing of behaviors were still considered major challenges at the beginning of the millennium [Gratch et al. 2002]. Around the same time, a study by Isbister and Nass [2000] emphasized the importance of orchestrating multiple cues, such as postures and language, to portray personality in a consistent manner. Back then, usually just one character was used that had to fit the preferences of all users. For this reason, Isbister and Nass [2000] did not rule out the possibility that a character with a partially matching personality might be better than a character with a mismatching personality. Their study showed, however, that users have a strong preference for characters that show a consistent behavior independent of their personality profile. Contrary to previous research, the similarity-attraction hypothesis (see Section 18.2.2) could not be confirmed.

The first decade of the new millennium was characterized by a larger variety of carefully coordinated behaviors that improved the expressiveness of the virtual characters [André and Pelachaud 2010]. This development was fostered by a variety of scheduling approaches, such as SmartBody [Thiébaux et al. 2008] or MARC [Courgeon and Clavel 2013], that automatically assembled synchronized animations and speech based on behavior descriptions in BML (Behavior Markup Language [Vilhjálmsson et al. 2007]). The more sophisticated behaviors of virtual agents led to studies that investigated the impact of multiple modalities, such as language and gesture [Neff et al. 2010] or language and gaze [Bee et al. 2010], on the perception of personality. Cafaro et al. [2012] showed that people already form a first impression of an agent’s personality and interpersonal attitude at first
encounters based on nonverbal immediacy cues such as smile, gaze, and proximity. The first 15 years of the new millennium were marked by approaches that created personality profiles for virtual agents beforehand, for example, to conduct perception studies as described above. For the last few years, a trend can be observed, however, to dynamically adjust a virtual agent’s personality to a user’s (potentially changing) preferences (see Section 18.3.2). This work is motivated by the observation that the similarity-attraction hypothesis does not necessarily hold, and it is thus hard to predict a user’s preferences regarding the personality of a virtual agent. Furthermore, there is a trend toward neural behavior generation approaches that enable the synthesis of a large variety of behavior styles for multiple modalities, such as movement and language (see Section 18.3.2). The two trends come with the promise of achieving the next level of human-likeness in the area of virtual agents.

18.6 Current Challenges

This section summarizes current research directions and unanswered questions. In particular, most applications of personality adaptation have yet to leave the laboratory and be tested with real users in their regular environment. The best way to do so would be by deploying fully functional systems to end-users, but in order to develop those several problems need to be solved first.

18.6.1 In Situ Studies

Socially interactive agents are getting closer to being integrated in our everyday life and domestic environments, thanks to many research insights and technological advances in the last decade. With human preferences being diverse, an investigation of (adaptive) social agents in the field of application, that is, “in the wild,” is important. However, the evaluation of social agents poses a number of challenges, in particular when targeting elderly people, domestic environments, and adaptation (see Figure 18.4). Users who did not use computing technologies throughout most of their lives often are not interested in participating in robot studies from our experience. For example, in Ritschel et al. [2019c, 2019d], people refused to let a robot into their home. People mention privacy concerns irrespective of their age in a variety of cases. It is very important to build social agents in a transparent and privacy respecting manner. However, this limits the technology that can be used in practice due to computing power restrictions in the wild, such as speech recognition and natural language processing. Another challenge is that some participants do not have Internet or do not permit its usage due to privacy concerns,
which prevents loading contents from the Internet dynamically. This can be important with regard to the novelty effect but also for information retrieval tasks or communication.

When investigating machine learning approaches, such as adaptation via reinforcement learning, another challenge arises. Since the agent faces a control problem and needs to decide what to do during runtime, every action manipulates its environment. Consequently, one does not know how the user would have reacted if the agent would have done something else. Especially in the beginning, if the learning agent starts without initial knowledge, this is an inherent problem. When using reinforcement learning, exploration is mandatory. Special care needs to be taken in order to make sure that the agent’s randomized action selection does not irritate the human or impact the overall interaction experience. However, variety in the agent’s expressed behaviors is essential to keep interaction with the agent interesting over a longer time, since reinforcement learning sticks to the most effective solution most of the time.

In addition, a serious challenge with regard to adaptation in an uncontrolled environment is that human feedback will be biased. Many external aspects may influence the human’s reactions, such as the user’s mood, stress due to upcoming appointments, but also technical restrictions. For example, when using social signals to calculate the reward signal based on human reactions, misinterpreted sensor data, bad lighting conditions, occlusions, or other problems occur that could even result in not being able to sense human reactions at all. Human feedback can also be biased by the contents presented by the agent, which are not under control of the adaptation process. For example, they may elicit emotions unintentionally and thus influence the feedback toward the agent, which is uncorrelated with the agent’s actual actions.

In general, there are also substantial technical challenges when conducting in situ studies. The agent needs to be completely autonomous, and thus error handling is an important aspect, which becomes increasingly important when several components need to work together. There is no option to control the agent remotely or to make sure that it is working correctly if there is no Internet connection on-site. However, autonomous interactive agents (see Section 18.4.4) are essential for long-term interaction studies [Leite et al. 2013].

18.6.2 Finding the Right Level of Sensitivity

The further perfectionization of techniques for the analysis of social signals might lead to agents that respond to human signals in an oversensitive manner [Eichner et al. 2007]. Agents which adapt to human social signals may irritate users. In case the human was not aware of the expression of his or her own social cues, the agent’s
reactions might appear random. Obviously, not every social signal cue from the user should trigger a response from the agent. The problem of deciding which user behavior should be interpreted as system input is called the Midas Touch Problem. Hoekstra et al. [2007] present a number of strategies to mitigate the Midas Touch Problem for an application with two agents that adapt their presentations to the user’s level of attentiveness. In their work, eye gaze was the only user cue that was interpreted by the agents. Thus, the question arises of how to determine the right level of sensitivity for a multitude of social signals in interactive conversational settings.

18.6.3 Fine-grained Behavior Timing

Personality affects an individual’s timing in conversation. For instance, talking over another person is often seen as a sign of dominance [Rogers and Jones 1975]. However, turn-taking is a complex issue and involves many more factors than personality alone [Goldberg 1990]. For instance, completing the speaker’s sentence or responding early can also signify attention and engagement, for example, when two people are having a passionate conversation about shared interests.

Most systems so far use fixed heuristics, such as yielding the turn immediately upon the user’s attempt to interrupt the agent [Bohus and Horvitz 2011] or responding after a fixed time threshold [Visser et al. 2012]. However, computer-controlled characters are faced with an increasing number of different contexts, for example, in the case of personal assistants that accompany their user throughout the day and likely for months and years. Therefore, they are in need of turn-taking mechanisms that are appropriate for a wider range of contexts yet easily adapted to achieve the user-preferred personality impressions.

As brought up in Section 18.2.2, context-specific behavior preferences can be approached via considering the underlying goals. A system, such as the virtual butler described by Conati [2013], relies on these relationships between internal goals and the personality, which is reflected in a user’s or an agent’s surface behavior. At the same time, showing appropriate emotional responses to events that affect a character’s goals is important for making their behavior believable and consequently engaging for the user [op den Akker and Bruijnes 2012].

Decision-theoretic reasoning can then be applied to infer the best course of action based on knowledge of these goals and how the system’s actions will affect them [Bohus and Horvitz 2011, Conati 2013, Janowski and André 2018, 2019]. However, identifying the goals that shape interpersonal timing is not trivial. While there have been attempts to define the structure behind human motivation [Ortony et al. 1988, Talevich et al. 2017], few works have linked those to a person’s Big Five traits or interpersonal stance. Those who do focus on high-level or long-term goals...
rather than concrete short-term intentions [Reisz et al. 2013]. But in order to make decisions about an agent’s behavior, a system needs to reason about short-term goals such as giving the user time to talk or delivering a message at the earliest opportunity.

18.6.4 Autonomous Interactive Systems
In order to bring socially interactive agents out of the laboratory and into the homes of actual users, it is necessary to verify that the findings from perception studies and Wizard-of-Oz evaluations hold true when autonomous systems have to interact with a human in real time.

One of the few real-time interaction examples is found in the work of Gebhard et al. [2019]. However, their study relied on a limited domain with pre-scripted dialogue and pre-defined user actions, which is still common practice.

A less constrained setup is described by Skantze et al. [2015]. Their system was exhibited in a museum over 9 days, gathering data about turn-taking patterns while pairs of users played a card sorting game with the robot. Though the authors mention the possibility of adjusting the robot’s personality through a confidence parameter, that confidence was neither the focus of that study nor was it considered in relation to surface behavior variations.

As mentioned in Section 18.4.3, one problem that makes people turn to Wizard-of-Oz setups or pre-scripted interaction is the need for quick reaction times. op den Akker and Bruijnes [2012] identified real-time input processing as a major bottleneck in dialogue systems and a challenge that needs to be solved before one can consider implementing agents that adapt their behavior during conversation. Over the years, there have been different approaches to that. Bohus and Horvitz [2011] explicitly modeled the delays that occur between a user’s speech, the system’s perception, the selection of a response, and the audible text-to-speech output. Other researchers such as Visser et al. [2012] applied machine learning to try and predict the content of incomplete sentences step-by-step in an incremental manner.

While real-time incremental input and output processing form a research topic of their own, it is worth noting that the challenges in this area limit an agent’s adaptive behavior with regard to the time frame. This supports the idea that adapting to the slowly changing, relatively stable personality traits of a user may lead to better results than trying to adapt to, for example, their short-lived emotions.

18.7 Conclusion
Equipping an artificial socially interactive agent with a personality that adapts to the user is a helpful strategy for making the interaction more enjoyable and more engaging. It builds on the humans’ natural tendency to anthropomorphize
inanimate objects and evaluate their behavior in terms of human codes of conduct. As a consequence, interest in creating believable artificial actors has been growing since the mid-1990s.

Personality traits are expressed in a number of verbal and nonverbal communication behaviors such as the choice of assertive words, demure gaze behavior, or patiently waiting for the right time to respond. By observing these surface cues, both human and machine can infer their conversation partner's disposition for reacting to certain events in a specific manner. This, in turn, enables a socially interactive agent to not only make its internal processes transparent to the user but also to configure its own personality traits to better suit the user's requirements and personal preferences.

In order to create an adaptive personality, shallow heuristics and superficial mimicking will not suffice in the long run. Future systems need to take the psychological foundations into account if they are to display social competence in a wide range of scenarios. Furthermore, automated learning processes are necessary for tailoring a system to an individual user's preferences while probabilistic reasoning helps to recognize these preferences more accurately in varying contexts. The benefit gained from the agent's action—be it in the form of a reward function or the expected utility—then enables it to choose the optimal strategy for interacting with the user.

Due to the complexities of adaptive systems, its components, such as the altered behavior or a reinforcement learning mechanism, are often evaluated separately. However, to truly understand how to adapt the agent's personality to that of its user, these applications need to be brought out of the laboratories and, ideally, to be tested over an extended period of time with actual members of the target group.

For this purpose, the adaptive agent must meet basic usability standards so that people will actually be motivated to use it long enough for observing the adaptation. This entails that the system must not irritate the user while adapting its behavior. It must be sensitive to the user's feedback, but not too sensitive either. Its actions must be timed properly in different contexts while still conveying the currently learned personality in a credible manner.

Related research fields, such as incremental input processing and incremental dialog modeling, can provide helpful insights in this regard. Psychological research about human–human communication allows us to understand the mechanisms behind interpersonal compatibility and how superficial behavior patterns relate to the core personality traits that need to be adapted.

As we learn to equip artificial characters with believable personalities, we also learn about what makes the human mind tick. In the long run, this understanding
will allow both human and machine to work together in the most productive, entertaining, and supportive manner.

**Acknowledgement**

This work has been partially funded by the European Union Horizon 2020 research and innovation programme, grant agreement 856879 (Present), and the German Research Foundation DFG, grant agreement AN 559/8-1 (Panorama).

**References**


References


References


References


Chapter 18  Adaptive Artificial Personalities


References


Chapter 18  Adaptive Artificial Personalities


Chapter 18  Adaptive Artificial Personalities


Long-Term Interaction with Relational SIAs

Jacqueline M. Kory-Westlund, Hae Won Park, Ishaan Grover, and Cynthia Breazeal

19.1 Motivation

In this chapter, we provide an overview of research on long-term interaction with Socially Interactive Agents (SIAs). Over the past of couple decades, as SIAs have become more capable in their ability to interact and collaborate with people, they have been increasingly moving out of the laboratory and into the world with the goals of helping people and bringing benefit to people’s lives. Chapters in this book have surveyed applications in education, health and wellness, aging, therapy, entertainment, and more (see Chapters 21 on “Pedagogical Agents” [Lane and Schroeder 2022], 22 on “Socially Interactive Agents as Peers” [Cassell 2022], 23 on “Socially Interactive Agents for Supporting Aging” [Ghafurian et al. 2022], 24 on “Health-Related Applications of Socially Interactive Agents” [Bickmore 2022], 25 on “Autism and Socially Interactive Agents” [Nadel et al. 2022], 26 on “Interactive Narrative and Story-telling” [Aylett 2022], 27 on “Socially Interactive Agents in Games” [Prada and Rato 2022], 28 on “Serious Games with SIAs” [Gebhard et al. 2022] of this volume of this handbook). Over time, as SIA technology has improved, their competence and effectiveness in engaging with and supporting a variety of human activities and goals are being actively investigated in schools, hospitals, assisted living facilities, museums, shopping malls, the workplace, and people’s homes.

With this progress, both technical and scientific, people are starting to not just interact with SIAs, but to live with them and work with them as part of daily life. In a number of important application areas—for example, helping children
progress through an education curriculum, coaching people to change their behavior to better manage a chronic disease, or providing therapeutic support for children with autism—lasting benefits take time to realize, often requiring repeated encounters over days, weeks, or months. In other application areas, such as aging-in-place or acting as a digital assistant, people may call upon the SIA on a daily basis for a variety of reasons, including companionship, accessing digital information or services, or help with tasks in the home or workplace context. Therefore, advancing the state-of-the-art in long-term interaction is a growing area of research interest, and our research community should address technical and user experience challenges in sustaining long-term engagement and supporting intelligent and adaptive decision-making to help people achieve long-term goals. This may include learning about people over long-periods of time to provide personalized engagement and support, as doing so can enable SIAs to better optimize the benefit provided to different people as they grow and change over time. These challenges raise important research problems in the design principles of robot platforms and interaction scenarios, human modeling and evaluation methods, and algorithmic approaches.

The technological world is changing very fast. Today, we have digital assistants in our smartphones and smart speakers, embodied conversational agents on screens that entertain and inform us, and social robots on our countertops as helpful companions—all supporting a variety of skills and services. With so many AI-enabled, socially interactive, and collaborative technologies entering everyday life, we need to deeply understand how these technologies affect us in the long term. What are the long-term consequences of having such technology in our lives—whether benefits or detriments? How can these technologies be used to promote human flourishing? How do we mitigate ethical concerns—and there are many—about the use of social technology and AI in our lives?

19.1.1 Beyond Interaction to Relationship

In this chapter, we look beyond human–SIA interaction to focus on the question of what kinds of relationships people are forming with SIAs. Whereas rapport can be established in short-term interactions, relationships take time and change over repeated encounters. We know that there are many benefits that come along with positive relationships. For example, better student–teacher or student–tutor relationships often result in better learning outcomes (e.g., Sinha and Cassell [2015a, 2015b], Wentzel [1997]). Better doctor–patient, therapist–patient, or patient group relationships often lead to better health outcomes [Mallinckrodt 1989, Horvath and Luborsky 1993, Yorke et al. 2008, Wampold 2015]. Better coach–client relationships lead to better engagement and behavior change outcomes [Frates et al. 2011]. In SIA
research, we are witnessing that people are forming relationships, of some kind, with socially responsive technologies [Dautenhahn et al. 2002, Desteno et al. 2012, Kory-Westlund et al. 2018, Kory-Westlund 2019]. What benefits might we see if people form positive relationships with SIAs? What properties and characteristics are designed into SIAs that foster the ability of SIAs to form and maintain a long-term relationship—particularly one that provides benefit to people, such as improved teamwork, wellness, learning, behavior change for therapeutic improvement, and more? How do we design SIAs to build sustained, adaptive, personalized, positive relationships with people?

In the next sections, we explore what counts as a long-term interaction with an SIA. We then discuss relationship models and approaches, with an eye toward how we can design SIAs as relational partners for long-term interaction. With this background, we survey the field: what social robots, virtual agents, voice assistants, and other SIAs have been developed for long-term interaction so far? How has the field evolved? We discuss trends over time, as well as the similarities and differences between agents. Finally, we catalogue several of the current most pressing challenges in developing long-term interaction with SIAs and point out directions for future research.

19.2 What is Considered Long-term Interaction?

Where does one draw the line between a short-term interaction study and a long-term one? Is it based on the number of sessions? The accumulated amount of time in interaction? The total elapsed time from first encounter to last? Another way of asking this is when does novelty wear off? It is not disputed that novelty has some kind of an effect on initially increasing engagement, positive affect, and excitement about interacting with a SIA, and thus there is a push for more longitudinal work to “get past” the novelty effect (e.g., Baxter et al. [2016]).

In longitudinal SIA studies, for instance, researchers have reported that novelty has worn off after one to two sessions to 1–2 weeks, as assessed by increased boredom (e.g., Salter et al. [2004]) or decreased interaction time (e.g., Kanda et al. [2004], Gockley et al. [2005]). That is, after some amount of time, the pattern of interaction changed. In the first survey paper on long-term interaction in Human–Robot Interaction (HRI), Leite et al. [2013] equated novelty with familiarization and habituation, suggesting that novelty has worn off when familiarization or habituation with the robot is stable, that is, when a person does not react as much to it and starts preferring novel behaviors. Leite et al. [2013] suggested using gaze or looking time to determine habituation. Using gaze, as well as other behavioral measures, is a reasonable suggestion. For instance, some research suggests that children look less at a familiar peer and look longer at an unfamiliar peer; they also may play
more cooperatively with a familiar peer and show more behaviors such as seeking attention, asking questions, and showing affection [Doyle et al. 1980, McCornack 1982].

But is gaze a reliable measure? For instance, in an 8-week child–robot study by Kory-Westlund [2019], they examined children’s gaze patterns during an Anomalous Picture Task, comparing children’s gaze in a between-subjects study where the anomalous pictures were introduced by a person or a social robot. If the robot were perceived as more novel than the human, and more novel at the pre-test than at the post-test, one would expect to see a decrease in children’s amount of looking time between the robot and the human, and from the pre-test to the post-test. However, this was not what was observed. Children did spend more time looking at the robot than at the human experimenter at both times, but they also looked at the robot more at the post-test than at the pre-test. This could be that the relationship between novelty and children’s gaze patterns is not as simple as decreasing novelty leading to decreased gaze. It may be that children were looking more at the robot as an attention-seeking behavior, which may also be related to greater familiarization, or perhaps they were looking longer because they knew it was their last session with the robot. Either way, it seems that more work is needed to understand how children’s gaze patterns relate to their perception of novelty.

Others have proposed interaction frameworks comprised of stages characterized of interaction patterns to study long-term interaction. For instance, Sung et al. [2009b] studied how people interacted with a Roomba vacuum cleaner over a 6-month period in a home setting. Although the Roomba is not an SIA, the authors proposed a long-term framework comprised of four temporal phases comprised of key interaction patterns: pre-adoption, adoption, adaptation, and use and retention. Kidd and Breazeal [2008] present a 6-week study of a social robot weight management coach in a home setting. They describe three phases of human–SIA relationship state: initial, normal, or repair. The initial phase lasts the first 4 days of interaction before transitioning into the normal stage for another 4 days. After that, the robot begins to ask two questions from the short form of the Working Alliance Inventory each day (a measure commonly used in therapy and other helping relationships that tracks trust and belief in a common goal of helping that the therapist and patient have for one another [Horvath 1989]) to calculate a relationship score. If the relationship score falls below the normal threshold, the robot enters the repair phase where it changes its relational dialog behavior in an effort to rebuild the working alliance.

Because there are no consistent ways of measuring novelty nor for determining whether observed interaction patterns were in fact a result of novelty wearing off
(versus merely being boring after doing the activity a couple times, say), it is actually hard to know whether reported effects are due to long-term interaction or due to novelty. Thus, for the purposes of deciding what work to include in this chapter, we admittedly have to be a bit arbitrary. We include works where people interact with an SIA for at least five sessions over any length of time per session and for any elapsed time from the first encounter to the fifth or more. We shall resume our discussion of novelty in the section on Future Challenges.

19.3 Relationship Models and Approaches

People form a wide variety of relationships with others. For example, people have different types of relationships with their friends, parents, children, managers, employees, colleagues, and so on. People also form meaningful and beneficial relationships with non-human beings, such as companion animals. People also have different kinds of relationships with technology, though for the purposes of this chapter we focus on people’s ability to form interpersonal relationships with SIAs and how these relationships can be maintained over time.

From numerous studies in HCI and HRI, we are seeing that people of all ages can construe SIAs as relational others if designed appropriately—with responsiveness and interactivity. That is, these relational SIAs are part of the broad category of things with which one can have a relationship. They are more than playful objects [Ackermann 2005] or transitional objects [Winnicott 1953] since either category would imply that they are only artifacts for projecting onto, for exploration and learning, rather than for being with. For example, work with social robots has shown that people perceive them to be social and relational (e.g., Turkle et al. [2006], Zawieska et al. [2012], Darling et al. [2015], Kory-Westlund et al. [2018]). Often, they are seen as having some of the properties of pets, toys, computers, artifacts, assistants, and friends, but not exactly the same properties as any of these (e.g., Kahn et al. [2002, 2012], Bartlett et al. [2004], Melson et al. [2009], Weiss et al. [2009], Kory-Westlund et al. [2018]). They are frequently ascribed social presence (e.g., Biocca et al. [2003], Leite et al. [2009]) and can evoke rapport (see Chapter 12 on “Rapport Between Humans and Socially Interactive Agents" [Gratch and Lucas 2021] of volume 1 of this handbook [Lugrin et al. 2021]), attachment, trust, and emotion (e.g., Turkle et al. [2006], Kidd and Breazeal [2008], Weiss et al. [2009], Bickmore et al. [2010], Batliner et al. [2011], Hancock et al. [2011], Desteno et al. [2012]). Dynamic social interaction factors such as the contingency of the SIA’s nonverbal behavior and its expressivity impact engagement, trust, learning, and judgments of the SIA’s credibility (e.g., Breazeal et al. [2016], Lubold et al. [2016, 2018], Kennedy et al. [2017a], Kory-Westlund et al. [2017b], Kory-Westlund and Breazeal [2019b]). In fact, people seem to apply a range of social judgments to SIAs and are willing to treat...
them in a variety of roles with different levels of authority relative to themselves. For instance, in an educational context, researchers are actively exploring a range of roles for SIAs from an expert tutor, to a peer-like playmate, to a novice to be taught (e.g., Chen [2018]).

The fact that children to older adults are treating SIAs as social-relational others brings new opportunities for engaging people in activities involving technology that plays an effective social-relational role. For example, numerous researchers have been exploring social robots to help children develop and practice skills that are best learned in social contexts—such as language or social and emotional skills (e.g., Robins et al. [2005], Kanda et al. [2007], Kim et al. [2013], Kory and Breazeal [2014], Bernardini et al. [2014], Kennedy et al. [2016], Clabaugh et al. [2018], Scassellati et al. [2018a, 2018b], Vogt et al. [2019]).

19.3.1 Dyadic Model

In the social sciences, relationships are modeled in numerous ways. These models provide a theoretical and empirical foundation from which to design SIAs that can socially interact and build relationships with people. One common model is the social system, the simplest example of which is a dyad. In a dyad, a relationship is defined as a pattern of interaction, for example, the interaction of two people whose behavior is interdependent [Csikszentmihalyi and Halton 1983, Kelley et al. 1983, Berscheid and Reis 1998]. Critically, this model can be applied to human–object relationships since non-human objects can also significantly influence our patterns of interaction and behavior [Csikszentmihalyi and Halton 1981].

The long-term education studies with children and robots paint an intriguing example of dyadic, interdependent interaction. With sufficient variation in the robot's behavior, children easily engaged for many sessions, up to several months (e.g., Kanda et al. [2007], Kory and Breazeal [2014], Leite et al. [2014], Kory-Westlund [2019]). They frequently learned new words from robots. They treated the robots as social others, frequently appeared to grow more comfortable and closer to the robots over time, and often called them their friends. In one study, children engaged longer and were less likely to grow bored if they treated the robot as a peer-like friend [Kanda et al. 2007]. This suggests that children's relationships affect how interested they are in interacting and playing—as one might expect, children like playing with their friends.

19.3.2 Dimensional Model

Another important relationship model is the dimensional model, in which relationships are defined in terms of various relational characteristics, including power, social distance, and trust [Burgoon and Hale 1984, Spencer-Oatey 1996,
Berscheid and Reis 1998, Fogg and Tseng 1999, Cassell and Bickmore 2000, Bickmore and Cassell 2001, Trope and Liberman 2010]. The dimensional model is important because these characteristics can be manipulated by non-human objects as well to influence the relationship. For example, Desteno et al. [2012] observed human–human behavior during an economic exchange game, identifying a set of nonverbal cues that were predictive of human cooperative behavior during the game. Then, they experimentally manipulated the nonverbal cues used by a social robot that played the same economic game with a human. They showed that the robot's use of the set of nonverbal cues affected human perception of the robot's trustworthiness and cooperative behavior in the game.

19.3.3 Provisional Model

Other models include provision models, in which relationships are discussed in terms of what people provide for one another (e.g., Duck [1991]), as well as economic models such as social exchange theory, in which relationships are modeled based on perceived costs and benefits of the relationship (e.g., Brehm [1992]). Important in relation to provision models is social support theory, which describes how social relationships influence people's cognition, emotions, and behavior [Lakey and Cohen 2000]. Social support theory becomes particularly relevant if we conclude that people can have social relationships with non-human objects.

In health-related domains, therapeutic alliance (or working alliance) refers to the relationship between a healthcare professional and a client or patient. It is the means by which a therapist and a client hope to engage with each other to effect beneficial change in the client. It consists of three parts: tasks, goals, and bond. Tasks are what the therapist and client agree need to be done to reach the client's goals. Goals are what the client hopes to gain from therapy. The bond forms from trust and confidence and the belief that the tasks will help the client achieve their goals. Research on the working alliance suggests that it is a strong predictor of client outcomes.

For instance, Kidd and Breazeal [2008] performed a 6-week study in which participants worked with either a co-present robotic weight loss coach, a standalone computer, or a standard paper log to see which would most effectively help participants achieve and maintain their weight loss goals. They used a measure of working alliance to estimate the user's relationship with the robot coach or the computer and found that taking specific actions to improve the relationship when working alliance was low led to greater engagement and improved health outcomes. Bickmore et al. [2005] found that using relational behaviors, such as social
dialogue, empathy, nonverbal cues, and relationship-building actions, in a computer health interface led to increased interaction and working alliance compared to one that did not use these behaviors.

19.3.4 Relationships with Animals
Animal-assisted therapy (AAT) recognizes the health benefits that can arise from human–animal relationships. The goal of AAT is to improve a patient's social, emotional, or cognitive functioning. The biophilia hypothesis suggests that if we see animals at rest or in a peaceful state this may signal to us safety, security, and feelings of wellbeing that in turn may trigger a state where personal change and healing are possible. People construe companion animals as being non-judgmental, comforting, and welcoming. Animals can also be supportive of educational and motivational objectives for people as a supportive, positive presence. For instance, canine-assisted reading programs are used to help children with special educational needs. The calm, non-judgmental, happy characteristics of dogs helps the process of reading to become more meaningful and enjoyable for children. Numerous studies have found that one-on-one or free-form interaction with the seal robot Paro in assisted living centers and nursing homes can increase positive affect and quality of life (e.g., Bemelmans et al. [2015], Lane et al. [2016], Moyle et al. [2018]).

19.3.5 Attachment Theory
Finally, attachment theory concerns the relationships between people, often in the context of young children and their adult caregiver [Bowlby 1958, Ainsworth 1969, Ainsworth and Bell 1970]. Attachment means an “affectional bond” or tie between an individual and an attachment figure. Such bonds may be reciprocal between two adults. Between a child and a caregiver these bonds are based on the child’s need for safety, security, and protection. Attachment theory has also been discussed in relation to the formation and maintenance of relationships with both humans and objects [Passman and Halonen 1979, Bretherton 1992].

19.4 Designing Relational SIAs
Considering these various models, we can see a variety of features that tend to be associated with relationships. First, relationships tend to unfold over time and generally involve multiple interactions. This may be on short timescales, such as repeated encounters over the span of minutes or days, or it may be on longer timescales, such as months or years. Even in short timespans, people's behavior can interdependently influence each other [Davis 1982]. These repeated interactions build up shared experiences—that is, activities done together in the past or are performing together now. Shared experiences influence later interactions
and are often referenced and remembered later on. These shared experiences can involve trying to perform tasks or achieve goals together.

There is also some amount of responsiveness and commitment. Those we form relationships with respond to us, for example, with social cues in the moment or social support in response to life events. Attachment and emotion often come into play; we may feel positively or negatively about interacting with certain people (see Chapter 10 on “Emotion” [Broekens 2021] of volume 1 of this handbook [Lugrin et al. 2021]). Empathy plays an important role—the ability to sense other people’s emotions coupled with the ability to imagine what someone else might be thinking or feeling. Affective empathy refers to the sensations and feelings that arise in response to others’ emotions. Cognitive empathy has to do with the ability to take the other person’s perspective to identify and understand their emotions. Friendship relations often involve positive feelings, trust, and attachment, such as enjoying one another’s company and depending on one another. Friendship relations often involve reciprocity as well, such as exchanging favors, reciprocating contact, dialogue, and connection, and being responsive in turn.

The examples of research presented above suggests that people can construe SIAs as social agents with whom they can form friendships and relationships. And these relationships can persist over time and bring value to people. Furthermore, people also appear to understand that SIAs are not quite the same as their other human friends, nor quite like their pets, or teachers, or nurses, or mechanical toys. But when technology is designed to act as social agents, people interact with them as social agents. They share gaze, mirror emotions, show affection, help the robots or interactive characters, take turns, and disclose information—all behaviors associated with friendships and close relationships [Hartup et al. 1988, Newcomb and Bagwell 1995, Rubin et al. 1998, Gleason and Hohmann 2006].

Even with robots or devices that are arguably less social (e.g., without the capability for speech), people can still attribute intelligence and talk to them and about them as if they have social capabilities (e.g., Sung et al. [2007], Wada and Shibata [2007], Fink et al. [2012], Bemelmans et al. [2015], Chang and Šabanović [2015], Moyle et al. [2018]).

These observations lead to more questions: How are people’s relationships with SIAs different than their relationships with other entities? What features of SIAs impact the relationship people can develop? Can SIAs actively try to build a relationship, and if so, how would this affect people’s engagement and perception of the relationship?

19.4.1 Designing SIAs as Relational Partners

If SIAs can provide similar kinds of interaction opportunities and features associated with relationships that people (or animals) can provide for each other—and it
seems that they can—then they, too, can be relational. Relational is different than just being social—it is the behaviors that contribute more directly to building and maintaining an ongoing relationship. This may include numerous social behaviors, such as the use of nonverbal cues and contingency, but is a larger category that includes additional behaviors, which we discuss further below.

Much of the work in exploring relationship factors with SIAs are done over short-term encounters, often over a single session. For example, people pay attention to the verbal and nonverbal social cues of agents to build rapport and to collaborate on a variety of activities and contexts with them. People largely seem to respond to these interpersonal social cues much as if they are being exchanged with another person (e.g., Chapters 3 on “Social Reactions to Socially Interactive Agents and Their Ethical Implications” [Krämmer and Manzeschke 2021], 9 on “Theory of Mind and Joint Attention” [Perez-Osorio et al. 2021], and 12 on “Rapport Between Humans and Socially Interactive Agents” [Gratch and Lucas 2021] of volume 1 of this handbook [Lugrin et al. 2021]).

Even over short encounters, we are beginning to understand the deeper implications of these construed relationships. For example, trusted and likeable SIAs can be more persuasive on human judgements and behaviors (e.g., Desteno et al. [2012]). In a learning context, we are starting to see that children will socially emulate the behaviors and attitudes of their social robot peer-like playmates with respect to modeling curiosity, affect, growth mindset, creativity, and linguistic expression [Kory and Breazeal 2014, Gordon et al. 2015, Kory-Westlund et al. 2017b, Park et al. 2017b, 2019].

### 19.4.2 Long-term Relational SIAs

SIAs can be created with long-term interaction in mind, with features such as memory and personalization that evolve over time from repeated encounters with users (e.g., Bickmore and Picard [2005], Lee et al. [2012a], Leite et al. [2013, 2017], Singh [2018], Breazeal et al. [2019], Ostrowski et al. [2019]). Based on the human social support literature and SIA research in a growing number of studies, a key aspect of why SIAs can benefit human outcomes (e.g., learning, health, wellness) is their nature as a relational technology, especially over long-term encounters.

We use the term relational SIAs to refer to the broader category of relational, personified agents—that is, all SIAs that can build long-term, social-emotional relationships with users. To enable SIAs to reach their full potential as relational technologies, especially for deployment during long-term interactions in real-world contexts, they need to be autonomous. This increasingly dovetails the design of SIAs with increasingly advanced artificial intelligence. Bickmore and Picard [2005] first introduced the concept of relational agents to refer to virtual humans,
primarily explored in healthcare contexts with adults. Kory-Westlund [2019] introduced the term relational AI to refer to autonomous relational technologies, recognizing the expanding range of personified technologies to include social robots, digital assistants, conversational devices, and so on, that people are using on a daily basis over extended periods of time. While one could argue that Alexa, for instance, is more transactional than relational today, studies show that people would like to see conversational AIs become more relational (e.g. Lopatovska et al. [2018], López et al. [2018], Sciuto et al. [2018], Singh [2018], Ostrowski et al. [2019]).

There are important human-centered design considerations for relational SIAs given that they should support familiar social and relational behaviors in order to be more understandable and relatable to humans. They should be designed to treat people in humanistic ways and build and maintain relationships in a way that are natural, appropriate, and ethical for people (see Krämer, Chapter 3 on “Social Reactions to Socially Interactive Agents and Their Ethical Implications” [Krämer and Manzeschke 2021] of volume 1 of this handbook [Lugrin et al. 2021]). From a computational perspective, relational SIAs need computational models, algorithms, and mechanisms to update a model of the person(s) with whom it operates in order to build and maintain a relationship over time.

Kory-Westlund [2019] identifies the following features associated with relational AIs (SIAs) to be human-centered, collaborative, interpersonal, relational, and reciprocal. Features that are necessary and sufficient to be relational per her criteria include repeated encounters, shared experiences, mutual change, responsiveness, emotion and positive affect, and reciprocity. These are all features that tend to be associated with relationships.

Repeated encounters. Relationships are longitudinal—they generally develop through time and involve multiple interactions. Relational SIAs should be designed to handle repeated interactions with users through time.

Shared experiences. Humans generally have a sense of past, present, and future, which is reflected in our relationships. We acknowledge our shared experiences through time via references to our past and present together, as well as looking forward to future activities we might do together. For example, sharing a humorous experience during an initial encounter with a stranger led to increased ratings of closeness [Fraley and Aron 2004]. Relational SIAs should create and reference a shared narrative with users. This may require an internal state that represents the user over time that can be updated during interactions.

Mutual change. As part of creating and referencing shared experiences, relational AI should change over time. More specifically, relational AI should change as a result of the interaction with the user over time—it is not enough to follow a changing
but scripted storyline (e.g., Gockley et al. [2005]). The change has to be perceived as “meaningful” in that the activities performed with the user (i.e., shared experiences over repeated encounters) must be clearly seen to affect the relational AI’s outward attitudes, emotions, or behavior. For example, people in close relationships may converge toward similar emotional reactions to events (e.g., Anderson et al. [2003]) or similar choices of food [Bove et al. 2003]. Again, this may require an internal state that represents the user over time. It is not sufficient that the person changes over time in response to the SIA, but that SIA should also adapt to the user. There is a growing number of studies that are examining autonomously changing/personalizing the robot’s behavior and/or the task content as a result of the child’s behavior or performance (e.g., Ramachandran and Scassellati [2015], Gordon et al. [2016], Lubold et al. [2016, 2018], Lubold [2017], Park et al. [2017a, 2017b, 2019], Scassellati et al. [2018a]). These studies have shown that personalization (i.e., a particular kind of change) can increase children’s engagement and learning and have opened many questions about how personalization and change might affect the child–robot relationship.

**Responsiveness.** Relational SIAs should ideally model a positive relationship. One element of successful, positive human relationships is rapport [Berscheid and Reis 1998], which is often indicated via behavior such as entrainment/mirroring and social reciprocity [Davis 1982, Dijksterhuis and Bargh 2001, Dijksterhuis 2005]. These behaviors are part of being responsive to users. Relational SIAs should respond and react to users, for example, by using appropriate social cues in the moment, or personalizing its feedback, entrainment, or behavior for individual users (e.g., Cassell et al. [2007b, 2009], Sinha and Cassell [2015b]).

**Emotion and positive affect.** As a human-centered technology, relational SIAs should respond appropriately to users’ emotional states. Prior work has found that mismatches between users’ emotions and the reactions of technology can negatively affect user perceptions and performance during interactions [Jonsson et al. 2005]. Promoting trust can be important for many kinds of applications. As one example relevant to education, trust can affect who children treat as credible informants [Harris 2007, 2012]. Relational AI designed to act as a friend-like agent may also need to promote positive affect or attachment as well (e.g., Leite et al. [2012a, 2014]); as discussed earlier, children’s friendships often involve empathy and affection [Gleason 2002].

**Reciprocity.** The idea of social reciprocity relates back to responsiveness as well as shared experiences through time. As discussed earlier, relationships often involve various reciprocal behaviors, such as disclosing information, helping, conversing, and engaging in activities together, and providing companionship. Relational SIAs
should use these kinds of reciprocal behaviors and attempt to recognize and be affected by the user’s use of these behaviors in turn.

19.5 History of Long-term Interaction with SIAs

We present a brief history/overview of long-term interaction with SIAs, where we shall focus on work where a person interacts with a SIA for at least five sessions (see our discussion earlier), in a real-world environment, and where the SIA is autonomous. We include a range of social embodiments in this survey, from social robots, to virtual agents, and voice assistants housed in smart devices. We focus on prior works that study the social or relational aspects of the SIA and its impact on human behavior, engagement, and desired outcomes. Hence, we exclude research about long-term deployments that study multiple one-time encounters in which the SIAs interact with many different people but do not form a long-term relationship with individuals (e.g., tour guide robots or information kiosk agents in public spaces). By doing so, we hope to provide a different lens on the evolution of long-term SIA research—in contrast to other chapters in this handbook that survey SIAs in specific applications or survey the design of social interaction capabilities (often investigated over short-term or single session encounters). We first highlight long-term SIA work within each category of SIA embodiment (i.e., social robots, virtual agents, and voice assistants) as the work in each area has evolved differently. We then present larger trends across and between each category, highlighting key milestones in long-term SIA research.

19.5.1 Long-term Interaction with Social Robots

Research in long-term HRI began in the early 2000s. The first paper published on long-term interaction with robots was an exploratory study with a fetch-and-carry robot called CERO in a work environment [Severinson-Eklundh et al. 2003]. One of the interesting findings was that bystanders also needed to be able to interact with the robot beyond the main user but didn’t know how. The authors raised important design issues such as personality design, natural interaction via voice, and collaboration with the main user as well as with a small group. Sustaining engagement was also raised as an important issue.

Leite et al. [2013] presents the first survey on long-term HRI covering 24 papers that met their criteria for inclusion from 2003 to 2011. They used the keywords “long-term interaction,” “social robots,” and “study.” They only included papers that presented sufficient detail on the capabilities of the robot and study details where the robots were deployed in real-world environments such as offices, public places, schools, homes, and healthcare facilities. To expand this survey, we performed a search for relevant empirical work using the list of papers in Leite's
survey as a starting place, and we searched for more recent papers that referenced these earlier works. We also used various literature search tools such as Google Scholar and keywords such as “long-term interaction,” “longitudinal,” “repeated encounters,” “HRI,” “social robot,” “time,” and so on.

We found 67 papers published from 2003–2020. Of these, 17 were in the domain of education; 17 were in healthcare, 12 in eldercare, 3 were classified as entertainment applications, and others for general assistance. In Figure 19.1, some examples of robots with various forms and functions are illustrated from the education and health and wellness domains. A total of 54 of the 67 studies used fully autonomous robots, and the rest were teleoperated or used shared autonomy. Of the fully autonomous robots, only 25 used advanced autonomy where AI or machine learning was used by the robot to interact (e.g., perception, adaptation, dialog, decision-making). The remainder of the robots were relatively simple (e.g., reactive behaviors or hard-coded rule-based systems).
This section is not intended to be an exhaustive review but more to highlight a few key domains and questions in which long-term HRI has been investigated with social robots. We focus on papers that used fully autonomous robots. In the remainder of this section, we focus on two major application domains of interest in long-term HRI research: (1) social robots designed to help people learn and (2) social robots that help people stay healthy and improve emotional wellbeing.

### 19.5.1.1 Social Robots and Children

Social robots have been designed to support long-term interactions with children across a range of applications such as education (e.g., Kanda et al. [2004], Tanaka et al. [2007], Hyun et al. [2010], Lee et al. [2011]), therapeutic support such as for autism or physical rehabilitation (e.g., François et al. [2009], Kozima et al. [2009], Barakova et al. [2015], Scassellati et al. [2018a]), and health (e.g., Short et al. [2014], Kruijff-Korbayová et al. [2015], Coninx et al. [2016]). A wide range of robot embodiments have been explored from small humanoids (e.g., NAO, QRIO), to expressive characters that move according to principles of animation, to zoomorphic forms, mechanical forms, and more. The majority of long-term studies with children have been in the context of educational activities, with language-skill learning perhaps being the most common [Belpaeme et al. 2018, Randall 2019].

Kanda et al. [2004] presents one of the earliest explorations of social robots in a school setting where a small humanoid robot engaged Japanese elementary grade children over 18 days as a peer-like tutor to help children learn English words. The robot spoke the English words during playful interactions, such shaking hands, hugging, playing rock-paper-scissors, and playing a body-parts naming game. It used RFID nametags to identify individual children and included basic speech recognition and motion control. Children who remained engaged over 2 weeks learned the most English words, but children's engagement substantially waned over time, raising the challenge of sustaining long-term engagement.

A number of studies followed with several results regarding children's long-term engagement. A study by Tanaka et al. [2007] study showed that very young children (10–24 months) socialized with a robot that engaged them in social play (e.g., dancing, giggling in response to touch, sitting and standing, moving its hands). They noted that children appeared to bond with it, and the robot became part of the social ecology of the classroom. Touch-based interactions were among the most enduring, and also led to interesting teacher–child interactions, such as teachers showing the children how to treat the robot gently. This study suggested that children’s relationship with the robot helped maintain their engagement.

In other work, Salter et al. [2004] found that children grew bored of a robot that was designed for physical play even within the first few sessions if the robot’s
behavior was too repetitive. They changed the robot’s speech and behavior in later sessions and found that the increased variation improved interaction. Selecting different activities based on the child’s interests is another way to increase the variation in the interaction and can also improve engagement. For example, Coninx et al. [2016] found that switching between several different activities helped engage children in diabetes education over time. Further, as different children prefer different activities, adapting to switch activities to suit individuals’ preferences was also found helpful in maintaining engagement. Leite et al. [2015] highlights the importance of affective engagement, responding to users expressed emotions, and reinforcement-based adaption on long-term engagement in weekly sessions over 5 weeks.

Baxter et al. [2017] investigated the impact of personalizing a peer-like robot’s social behavior on children’s learning of novel or familiar subjects in a primary school classroom over a 2-week period. They found that social personalization (referring to the child by name, adapting speed of response) and adapting the speed of progression through educational material improved children’s acceptance of the robot as well as children’s learning of novel material. Over time, more sophisticated personalization algorithms based on reinforcement learning have been deployed in randomized controlled trials in schools over multiple months to improve children’s affective engagement over time [Gordon et al. 2016], and even personalizing to improve both engagement and learning simultaneously [Park et al. 2019]. Kory-Westlund and Breazeal [2019b] discovered a positive correlation between personalization and the quality of children’s reported relationship with a peer-like robot learning companion and children’s resulting vocabulary learning outcomes. Namely, children who reported a closer relationship with a social robot learned more, and children who had a close relationship with a personalized social robot learned the most.

### 19.5.1.2 Social Robots for Health and Wellness

The first long-term randomized control trial (RCT) study in the home was with Autom, a robot health coach that helped people to manage their weight [Kidd and Breazeal 2008]. However, the majority of papers that explore social robots for health and wellness has focused on aging. The rising number of older adults, and a growing shortage of clinical and non-clinical care providers, has motivated the development of social robots to address a wide range of issues and opportunities in the aging domain. Overall, people prefer to age-in-place in their own home for as long as they can, and there is a need and desire for affordable technological solutions that help older adults age with independence. A number of social robots
have been developed to provide support through social means to address physical decline, cognitive decline, health management, or psychosocial issues such as chronic loneliness or depression. The ability of a social robot to play the role of a motivating coach that can build rapport, track progress toward goals, and provide reminders has been explored in the context of physical rehabilitation, medication adherence, or serving as a health coach. Providing an educational function is relevant for healthcare, but it is also generally important to provide cognitive stimulation to help mitigate cognitive decline. To support people’s emotional well-being, social robots can entertain, help to socially connect people via telepresence or other means, as well as provide a sense of companionship.

These functions are relevant in assisted living contexts, too, where supporting the care staff is another important design dimension. Social robots for older adults have been designed with different physical embodiments and roles. For instance, zoomorphic robots have been studied over long-term encounters as pet therapy surrogates where affiliative touch is a positive emotion-eliciting interaction [Wada and Shibata 2007, Sung et al. 2015]. Other designs have an anthropomorphic embodiment that are capable of sharing gaze, making gestures, or using emotive expressions—sometimes with an accompanying touch screen to support a graphical user interface (GUI) integration in addition to spoken interaction. Designs can be more functional like a mobile kiosk such as SYMPARTNER [Gross et al. 2019]. Or they can merge the qualities of a helpful ally with those of a companion, such as Jibo [Ostrowski et al. 2019]. The ability to express emotion through multiple modalities (e.g., body movement, sound, and facial expression) contributes to people’s willingness to form an emotional bond with them as well as for the robot to provide emotional support. For instance, Jeong et al. [2020] reports on a 1-week study of the Jibo robot in undergraduate dormitories to serve as an emotional wellness coach to help students cope with stress and to promote emotional resilience. In general, careful consideration and alignment of how the physical appearance of the robot matches its intended function is important for user acceptance and long-term adoption.

The most extensive long-term studies with older adults have been with Paro, often with people experiencing cognitive decline or loneliness. These studies have verified Paro’s ability to provide pet-like companionship and emotional support similar to the benefits of animal-assisted therapy, for example, reducing stress and anxiety, reducing loneliness, and improving positive mood [Wada and Shibata 2007]. Mobile or legged social robots have also been explored in the context of being a component of a smart home designed for aging-in-place where acceptance by users over time has been of primary interest [Gross et al. 2012, Schroeter et al. 2013, Torta et al. 2014, Doering et al. 2016, Hebesberger et al. 2017, Gross et al. 2019]. These
robots often have a touch screen with a GUI to supplement a virtual user interface (VUI), or the GUI is the only mode of input. Given the exploratory nature of these long-term studies (lasting a couple of days, with multiple short-term encounters), the main research question focused on acceptance of the robot and what factors could lead to longer-term adoption and successful use.

Similar to what has been learned in long-term robot–child studies, it has been found that a change in the robot’s speech and behavior can help maintain user engagement and build a long-term relationship in aging, health, and wellness applications [Kidd and Breazeal 2008, Lee et al. 2012a]. Moreover, it will be important that social robots support more robust and flexible conversational abilities, they should be able to adapt and personalize to the users, and they need to support a variety of functions and tasks (both for the main users as well as caregivers).

As social robot platforms have matured, akin to the smart speaker market, recent studies have deployed commercial social robots with more robust VUIs and a multitude of skills—for example, information skills, entertainment skills, and social/persona driven skills. Ostrowski et al. [2019] reports on a 3-week study where the Jibo robot was deployed in community areas of an assisted living facility to support daily interactions among residents. They report that the robot served as a social catalyst to positively influence the social connectedness within the community of older adults as well as enhance senior citizens’ community engagement.

19.5.1.3 Living with Consumer Robots
There are a few long-term co-habitation studies of robots sold in the consumer market for the home. The first such studies explored how people adopted and used an autonomous floor vacuum cleaner (Roomba) over several weeks [Forlizzi and DiSalvo 2006] up to 6 months [Sung et al. 2009a]. Forlizzi and DiSalvo [2006] used an ethnographic approach to investigate how people interacted with the Roomba and how the presence of the robot changed their housekeeping practices. Although the Roomba is not a social robot, nonetheless, they found that half the participants developed a relationship with it. For instance, people named their Roomba, talked to it even though it had no voice interaction capability, made attributions about how pets related to the robot, demonstrated politeness to the robot (e.g., saying “excuse me” if they bumped into the robot), and would collaborate with the robot to clean as a team (e.g., either helping the robot by picking stuff off the floor or cleaning the room while the robot cleaned the floor). de Graaf et al. [2017] studied why participants stopped using a desktop robot (Kartoz) over a 6-month period. They report that a positive emotional experience with the product is important to capture users in the short-term (e.g., the first couple of weeks). However, relevant functionality is key to retain usage in the longer-term (e.g., the first couple of
weeks). Over the long-term (e.g., 6-months), users would replace the robot and use a different technology for the same function if they found the robot experience to be too annoying, too cumbersome or frustrating to use, or too boring and repetitive in comparison to the alternative.

More recent cloud-connected social robot platforms (such as Jibo) incorporate the technology advances of voice assistants (e.g., wake word, far-field speech, natural language understanding, regular content or skill updates, in-depth persona design) to address many of the shortcomings of these older consumer robot technologies. Singh [2018] presents comparative results from a 1-month in-home study with the Jibo social robot where they studied different engagement patterns between generations (children, younger adults, and older adults), as well as compared engagement and usage patterns to the Alexa smart speaker (see Section 19.6).

19.5.2 Long-term Interaction with Virtual Agents

Research in long-term interaction with virtual agents also began in the early 2000s. Bickmore [2003] first introduced the concept of relational agents: computational artifacts that build long-term, social-emotional relationships with users. For the purposes of this section, we surveyed papers on Google Scholar by searching for the keywords “embodied agents,” “conversational agent,” “virtual agent,” and “IVA” with the terms “long-term” or “longitudinal.” Based on our interaction threshold of five interactions or more, we found 24 papers published from 2003 to 2020. Bickmore and colleagues’ work has dominated research in investigating long-term interaction between users and virtual humans. Much of this work has been applied in the health and wellness domains for adults and older adults (see Figure 19.2 for examples). The approach has focused on the design and evaluation of virtual
humans that provide constructive and therapeutic support to users through verbal and nonverbal behaviors that build rapport, motivate, coach, and educate. Overall, use cases of virtual relational agents have primarily focused on three main areas of application: (1) health behavior change through motivation, (2) therapeutic/social companionship, and (3) health behavior change through education. We provide research highlights below, but this is not intended to be comprehensive.

19.5.2.1 Virtual Agents and Health Change

The first application relational agents investigated the use of a virtual human as a health coach to help motivate its user to sustain an exercise program, namely, to walk for at least 30 minutes a day for most days of the week [Bickmore et al. 2005]. A 1-month randomized controlled trial was carried out where users interacted with one of three interventions: a virtual human with relational behaviors, a virtual human without relational behaviors, and baseline with no agent. The agent in the relational condition used verbal rapport-building behaviors (social dialog, humor, etc.) and nonverbal immediacy behaviors (close conversational distance, gaze, facial orientation, etc.). All conditions included the standard behavioral interventions, self-monitoring, and educational content. They found that there was no significant difference in the amount of exercise that people performed between the virtual agent conditions, but people did exercise more compared to baseline. Furthermore, subjects in the relational condition liked the agent more, reported a closer relationship, and responded more favorably to continuing to work with the agent.

One drawback of Bickmore et al. [2005] was that the agent’s dialog was deemed too repetitive, reducing motivation among many participants. A follow-on study lasting several months explicitly examined the effect of dialog repetition on behavior change [Bickmore and Schulman 2009]. Users were divided into two groups where half interacted with an agent with a variable dialog structure for several months before switching to interact with an agent with a non-variable style for about the same amount of time. In the variable dialog condition, the agent could pick one of five different dialog structures (e.g., “Looks like you met your exercise goal of 5,000 steps. Great job!”), “Looks like you got your walking in and met your goal of 5,000 steps!”) during each interaction while in the non-variable condition the agent used the exact same dialog structure in every situation. The other group did the reverse order. Results showed that participants were significantly more likely to have and continue a conversation with the agent in the variable condition. However, participants walked a significantly greater number of steps in the non-variable condition.
A similar pattern was observed in another follow-on study to explore the effect of the virtual agent having a personal backstory on user’s engagement [Bickmore and Schulman 2009]. The agent in the first-person condition presented back stories in the interaction as its own. In the third-person condition, the agent presented back stories about a friend. Participants in the first-person condition reported significantly greater enjoyment and willingness to interact again. However, again subjects in the third-person condition walked significantly more steps. Yin and Bickmore [2018] evaluated the effects of cultural adaptation on behavior change in a 4-month study with Latino adults in an underserved population. The agent itself was culturally adapted in multiple ways: it resembled a Latina woman, participants could choose their preferred language between English and Spanish, and as part of social dialog, the agent demonstrated knowledge about Latino culture. Results showed that participants reported a high satisfaction with a significant increase in minutes of walking per week compared to the control arm.

In perhaps the longest RCT study with an SIA, Bickmore et al. [2013] performed a large-scale, 1-year study comparing the effectiveness of relational agents against computer-tablet based intervention with older adults. In the control condition, participants uploaded their pedometer readings to a computer. In the virtual exercise coach condition, participants uploaded pedometer data to a computer for the first 2 months, and then interacted with the virtual coach daily for the next 10 months in a kiosk in a clinic waiting room. They found that participants in the virtual agent condition walked significantly more than those in the control condition after the first couple of months, but this trend waned over time, and there was no significant difference in the two groups at the end of 12 months. A closer look revealed that participants with adequate health literacy in the virtual coach condition significantly benefited from the interaction and walked more than the control condition at the 2-month and 12-month interview period, while those with inadequate health literacy showed little to no improvement even after interacting with the agent at either interview point.

A number of studies followed where virtual agents were studied in the context of promoting health literacy and medical adherence [Gardiner et al. 2013, Jack et al. 2015, 2020, Kimani et al. 2016]. For instance, Gardiner et al. [2017] developed a virtual agent to promote mindfulness and lifestyle education among urban women. The agent provided mindfulness exercises, positive ways of managing stress, suggestions to increase physical activity, and motivation to eat healthy. In a 1-month RCT, participants in the experimental condition interacted with the virtual agent. In the control condition, participants were provided with written educational information and meditation audio files to listen to. Results showed that women in the experimental condition significantly reduced their alcohol consumption while increasing their intake of fruits by an average of two servings.
19.5.2.2 Virtual Agents and Wellness

Emotional wellness has been another application domain for long-term studies with virtual agents with affect-aware capabilities to explore their potential benefit in therapeutic/social companionship. For instance, research has shown that older adults with strong social connections have decreased health risks and mortality, raising the question of how SIAs might be able to help alleviate problems associated with social isolation. For instance, in a 1-week study with older adults, Ring et al. [2012] developed a relational agent to help mitigate loneliness. The agent used relational behaviors, motivational dialogue, short anecdotal stories in its interactions, and sensed the user’s mood to provide appropriate emotional feedback. The study explored whether the virtual agent should be passive (wait for the user to initiate interactions) or proactive where the agent could initiate interactions too. The results showed that while there weren't significant differences in the length of interactions between the two conditions, participants in the proactive group saw a greater decrease in loneliness. There was also positive correlation between comfort with the agent and time spent with the agent. Finally, the more lonely a participant was, the more they interacted with the agent, reducing their loneliness as a consequence.

In another long-term study, Ring [2017] explores the use of a virtual agent for depression counselling. Participants were divided into three conditions: control, standard, and affective. In the affective condition, the virtual agent responded to its user based on his/her emotional state (e.g., feelings of anger, shame, fear anxiety, etc.), and it detected emotional discrepancies to provide appropriate feedback for emotionally sensitive dialog. In the standard condition the agent did neither, and participants did not interact with any agent in the control condition but simply filled out weekly PHQ-8 and state anxiety questionnaires. While no differences were detected in PHQ-8 scores, the results indicated significant reduction in state anxiety scores in the affective condition compared to the standard and control conditions. No significant differences were detected between the affective and standard conditions in terms of likeability, willingness to continue, and interest with all measures tending in the positive direction for both conditions. In both conditions, participants found the agents to be caring and saw them more as a friend than a stranger. These results further motivate the use of virtual agents as social companions to alleviate anxiety and provide therapeutic benefits to users.

19.5.3 Long-term Interaction with Internet of Things Voice Assistants

Personified voice assistants, for example, digital assistants with a voice interface, embedded in smart devices such as smartphones, speakers, or displays are by far the most common SIA with millions of products sold. In contrast to social robots
and virtual humans, they often have no visual form (or a simple, abstract visual representation) and voice is the primary way the persona is conveyed. Siri was a spin-off project originally developed at SRI International, was first released as a voice app in 2010, was quickly acquired by Apple, Inc., and then released in the iPhone 4S in 2011. The first release of a voice assistant on a smart speaker was in 2014 with Alexa on the Amazon Echo.

Other companies have followed with their versions of voice assistants. For instance, the Google Assistant was unveiled in 2016 in the Google Home, and many other types of smart devices with voice assistants have since entered the Internet of Things (IoT) consumer market. These smart devices are “always listening” for their wake word and support far-field speech. These innovations have lowered the barriers to access, making it easy for all sorts of people, from young children to older adults, to quickly and easily launch skills (akin to mobile apps) through simple voice commands. Developer ecosystems have proven successful in populating these smart devices with literally thousands of digital skills and services—from playing music, setting timers, controlling other IoT devices, ordering pizza, getting news, Q&A, and much more.

Significant effort has gone into the persona design of these agents, and people enjoy having digital assistants tell jokes and engage in small talk or chitchat with them. Consumers have shown broad willingness and pleasure in asking these agents to offer their own “opinions” on a wide range of topics, to express their own likes or dislikes, and to express their associated “emotions.” It has been noted that users readily personify voice assistants, ascribe a gender to them, and refer to them in human-like terms such as “friend” and “someone to talk to” [Turk 2016, Pradhan et al. 2018]. They are cloud connected and support over-the-air (OTA) updates, so such devices are constantly updated with new content, skills, and personality quips.

Despite that millions of these devices are in people's homes, the long-term use of these devices has not been extensively studied or understood by the HCI research community. There are a limited but growing number of long-term study papers. For instance, Bentley et al. [2018] analyzed voice history logs of over 65,000 interactions with Google Home devices in 88 early-adopter homes over about 3 months of use. They identified particular patterns of activities by four distinct user groups based on the type of skills and the time of day they used the smart speaker. They also found that users settled pretty quickly into these patterns of use for which commands they tended to favor, and these did not change much over time past the first 3 weeks of use. They also identified that different demographics by age tended to use the Google Home differently. Younger adults were the most active users (age 18–44), favoring music, home automation, chitchat, Q&A, setting
timers/alarms, and getting the weather were pretty consistent over time. The device was not nearly so engaging over the long-term for older adults (age 45–64); however, music, chitchat, setting timers/alarms, and weather were among the most used features. The authors note that there are opportunities in the design of such devices to help introduce users to new skills in new domains, to support multimodal interactions (e.g., spoken interface with a screen), to anticipate user’s patterns of use to proactively offer information, and to support deeper agent-based interactions through richer conversation-based interaction.

Users may project a relationship onto voice assistants and form emotional attachments, as was seen for some users in a recent 1-month home study with the Alexa agent [Singh 2018]. Pradhan et al. [2019] studied how older adults perceive personified smart speakers as social agents or companions versus objects such as appliances. They deployed Amazon Echo Dot devices into the homes of adults over the age of 65 for a period of 3 weeks. They found that these personified devices were often treated as social agents, revealed through use of pronouns or polite behaviors (such as saying “please” and “thank you” to the agent). Interestingly, older adults would fluidly talk about the agent as both being human-like and object-like, depending on the specific encounter. For instance, greeting the user by name encouraged personification whereas interactions that conveyed a lack of “personal touch” encouraged objectifying the agent. Having the voice agent engage in small talk and greetings were more important to users who desired companionship (e.g., were more prone to feelings of loneliness).

Today’s commercial voice assistants generally do not attempt to build or maintain a social-emotional relationship, yet. Bentley et al. [2018] notes that the current style of interaction is far more transactional, where users task the agents using voice commands rather than engaging in collaborative conversation. However, research studies with virtual humans and social robots reveal users’ desire for relationship and companionship with these personified technologies, especially in areas where deeper engagement is needed—that is, to promote long-term quality of life and learning outcomes.

19.6 Similarities and Differences in Social Robots, Virtual Agents, Voice Assistants, and other SIAs

When we compare the long-term work using different relational SIAs—such as physical robots, virtual agents, voice agents, and more—we can examine a number of dimensions, including:

- The domain—for example, healthcare, education, therapy, entertainment.
- The population—for example, children, adults, the elderly.
— The agent’s embodiment—for example, physical or virtual, humanoid or non-humanoid.

— What was the impact of the SIA on human outcomes?

— What were the broader implications of the work, for example, for ethics or design?

Both social robots and virtual agents have been developed for a wide range of application domains and for use by a wide range of user populations. The embodiment of an SIA affects the kinds of behaviors it is capable of using, the tasks it can be used for, and the kinds of relationships people may form with it. In terms of long-term studies, a wider range of morphologies have been explored in social robots so far, including zoomorphic—for example, Paro (seal-like), Aibo (dog-like), and Pleo (dinosaur-like). There are also robots with functional embodiment (e.g., Roomba or Cero), as well as anthropomorphic characters that express according to principles of animation, for example, Tega (a squash-and-stretch robot) or Jibo (that can strike expressive postures through its line-of-action). Finally, there are humanoid or even android forms—for example, the humanoid Nao and the Geminoid, which emulates the human form with skin, teeth, and hair. The physicality of robots also affords touch-based interactions, which are particularly observed when robots take a pet-like or character-like embodiment. The wider variety of robot morphologies may lead to a wider range of possible human–agent relationships—pet-like companions, vacuums that people collaborate with to clean a room, learning companions for children, health coaches, and more. Virtual agents, on the other hand, are most often portrayed as virtual humans in long-term studies, though nothing precludes them from taking other forms. The majority of long-term studies with virtual humans has been in health and wellness domains where the agent often serves as a health coach and utilizes relational features to build rapport to sustain engagement (e.g. Bickmore and Picard [2005], Ring et al. [2012], Vardoulakis et al. [2012], Bickmore et al. [2018], Sidner et al. [2018]). There are also a growing number of long-term studies with smart devices with digital assistant personas—for example, Siri with smartphones, the Google Assistant with smart speakers, Amazon Alexa with smart displays, and so on.

We recently surveyed 79 peer-reviewed publications (from 2003 to 2020) comprising 87 unique comparative studies in HRI that compared virtual agents to co-present and telepresent robots and smart speakers. The vast majority of these are in the context of short-term encounters. We also performed a meta-analysis of 59 of these studies. We categorized the SIAs studied based on their embodiment (physical, virtual, or a mixed agent with some physical and some digital components) and their perceived physical presence (co-present or distant). Multiple studies also
included other kinds of agents, such as humans, tablets, or laptops without an SIA or a voice-only agent embedded in a static device. A total of 54 studies compared two agents (62.1%); 28 compared three (32.2%); and 6 compared four or more (6.9%). We also categorized whether the types of tasks performed with the agent included other kinds of agents, such as humans, tablets, or laptops without an SIA or a voice-only agent embedded in a static device. A total of 54 studies compared two agents (62.1%); 28 compared three (32.2%); and 6 compared four or more (6.9%). We also categorized whether the types of tasks performed with the agent included physical (e.g., a Towers of Hanoi or block-stacking physical puzzle), social (e.g., conversation, storytelling, judging emotions), or digital components (e.g., tasks shown on a screen, digital puzzles).

Overall, examining the results from both the survey and meta-analysis favoring each agent type, we see a trend that physically embodied co-present robots affect humans more deeply and strongly than virtual agents, telepresent robots, or smart speakers. However, this can be modulated somewhat by the type of tasks performed with the agent. Humans and physically present robots were favored most often, and most often during socially interactive tasks and physical tasks. During digital tasks that focus on information, the results generally did not favor one agent over another. In addition, humans and co-present physical robots led to stronger increases in important social metrics including attention, attraction/liking, empathy, persuasion, and trust. The results of our survey suggest that the robot’s co-presence and embodiment—but presence more so than embodiment—seem especially important for interpersonal, social tasks.

To date, there are only a handful of long-term comparative studies that examine different SIA embodiments. While not all measured emotional engagement or relationships—overall, the physical presence of social robots seems to support deeper emotional engagement and often stronger relationship scores.

For example, Kidd and Breazeal [2008] compared a robotic weight loss coach with a non-embodied computer coach and with self-report using a standard paper log. The goal was to determine which embodiment would be more effective at sustaining engagement. The exact same dialog model and screen interface ran on the robot weight-management coach as on the computer. It was hypothesized that all interventions would help people lose weight, but the real challenge is helping people to keep the weight off. Hence, long-term engagement was the main outcome studied. They performed a 6-week in-home study with 45 adults. The results showed that participants were significantly more likely to adhere to the weight loss program with the social robot. Participants also reported a stronger emotional bond with the robot as well as higher ratings for the robot in terms of working alliance, trust, credibility, and engagement.

In a health and wellness application, Sidner et al. [2018] developed a SIA system for home use comparing a social robot (Reeti) to a virtual human (a female avatar) with older adults. The study looked at health, wellness, and social engagement outcomes as measured through conversation, on-screen games, and filling out forms.
They tested the system with 26 older adults, each of whom interacted daily with either the robot or the virtual avatar for 30 days. Sidner et al. [2018] asked participants about their satisfaction with the agents, measured overall system usage and time with agent, and asked about a variety of other attributes of the agent (e.g., likeability, trustworthiness). They found that participants generally displayed reasonably positive attitudes toward both agents, and there was a trend toward participants finding the robot more trustworthy and wanting to have more conversations with the robot than with the virtual agent.

In the domain of education, Vogt et al. [2019] developed a humanoid language learning tutor to teach English words to Dutch children, using a tablet-based game. In a seven-session study, 194 children played the tablet game one-on-one with the robot or with the tablet alone. Language skills tests showed no differences in learning outcomes. This may have been because the tablet game played a crucial role in presenting the learning content; the robot’s presence may not have been as important. This study did not measure children’s relationship or social engagement with the robot or tablet.

Singh [2018] reports a comparative study on how families and individuals interacted with either a social robot (Jibo) or a smart speaker (Amazon Echo) for 1 month in the home. This study examined differences in how different generations interact with VUI agents over time: children (under 18 years old), adults (between 18–65 years old), and older adults (aged 65 and older). Both agents could perform a variety of tasks from entertainment (music, jokes, etc.), information (weather, news, Q&A, etc.) and social (sharing opinions, likes/dislikes, etc.). Overall, the Amazon Echo could perform many more skills than Jibo (Alexa had thousands of skills while Jibo had around 20). However, Jibo was a far more emotively expressive and companion-like agent. For instance, Jibo had a persistent lifelike presence, responded to being petted, could dance, express emotions through body posture, turn to look at people, proactively greet users in the morning through face recognition, and could inquire and remember about how family members slept to provide personalized, contextually relevant responses. Participants in the study could interact with the agent when they liked; they could use any or all of the agent’s functionality (classified as social, entertainment, or functional). Singh [2018] found that children and older adults tended to use all three categories of skills on Jibo. Children primarily only used the entertainment skills on the Amazon Echo. With Jibo, children’s reactions and preferences showed that they were drawn to the robot’s ability to be a social other, and they treated Jibo more like a companion. Interestingly, overall engagement was sustained better for children and older adults with Jibo where the companion-like social-relational capabilities seemed to be an important driver for this trend. Usage of the Amazon Echo by older adults dropped over time. In
contrast, younger adults tended to favor the utility of the Amazon Echo and practical skills tended to drive their usage over time. Overall, the social robot was often classified as a friend or member of the family, as opposed to being classified as an assistant. The robot was also seen as more open, agreeable, and extroverted than Alexa.

In sum, there are few long-term comparative studies, so we cannot draw any strong generalizations at this point. However, these prior works (both over short- and long-term studies) suggest that the physical co-presence and expressive behaviors of social robots enhances their ability to emotionally engage people, and this can be an important factor in sustaining engagement over time (more so than for virtual agents or smart devices). For tasks that focus on information, or where relational capabilities are less important, we do not see a significant difference in human behavior across embodied agents. However, we could anticipate that for tasks where establishing a long-term relationship is important for outcomes, social robots could lead to better results. Further studies are needed to verify whether or not this is the case.

19.7 Trends in Long-term SIA Research over the Past 20 Years
From this brief survey on long-term interaction with social robots, virtual humans, and voice assistants, we can identify a number of trends in SIA research. First, early studies in long-term interaction were exploratory in nature, focused on basic questions such as user adoption, identifying sustained patterns of use, and reasons that underlie engagement or abandonment. As these factors have become better understood, we see more effort on running long-term RCT studies to understand how to design SIAs to bring about intended desired outcomes for people. Currently, there are still few studies that last beyond a couple of months even though this may not be long enough to get past the novelty period. We discuss the novelty issue in more detail in Section 19.9.1. Those lasting 6 months or more are very rare. Second, we see the SIAs increasing in their AI and algorithmic sophistication over time. Earlier systems were fairly scripted in their verbal behavior or reactive in their physical behavior, and people would quickly lose interest, especially if there wasn’t sufficient practical functionality to entice usage. Recent papers are starting to investigate algorithmic innovations in long-term contexts such as data-driven personalization or context-aware adaptation. Third, the design of SIAs early on tended to either focus on information/decision support or emotional support—SIAs were either tutors/coaches or they were pet-like companions. More recently, SIAs are more often dovetailing cognitive, social, and emotional support—and are increasingly combining affective computing methods and reinforcement learning techniques with more advanced conversational AI. Finally, the technological platforms are
evolving to become far more capable and robust. While earlier systems in the 2000s were pre-coded with fixed behaviors, modern platforms are cloud connected with OTA updates and are equipped with SDKs to support developer eco-systems. It is now much easier to deploy social robots, virtual agents, or voice assistants in the field and capture much finer grained, continuous user data (under Institutional Review Board ethics protocols), and update algorithms or skills on-the-fly. This trend is making it easier to run larger and longer-term studies—deploying multiple SIAs over longer periods of time with more participants (although this work is by no means easy). This platform trend is also making it possible to develop more sophisticated AI methods to understand long-term usage across users (as well as with a specific user) to drive adaptation, personalization, and continuous expansion (e.g., persona backstory, natural language models, knowledge about its users). This mitigates the interaction from getting “stale” as well as providing a repertoire of functional skills. Both help to address key factors that have diminished long-term engagement in the past. This also opens the door to much more sophisticated relational AI research (see Section 19.9.2). It also raises critical issues around their ethical and responsible design (see Section 19.9.3).

19.8 Current Challenges
Developing SIAs for long-term interaction presents a wide range of research questions, design challenges, and shall require algorithmic and technical innovations. In this section, we focus on four key challenges, raised in the previous section, for the research community to address. For each, the challenge is to achieve these beyond a couple of months (with proof-points to date) to 6 months or more (which has been more elusive).

— Sustaining long-term engagement.
— Supporting flexible, engaging conversation.
— Adapting and personalizing to people effectively.
— Achieving long-term beneficial outcomes using relational properties.

19.8.1 Long-term Engagement
Maintaining engagement over long-term interactions of weeks, months, or years can be critical to the success of the SIAs, especially in domains such as education and healthcare. Numerous factors can help foster long-term engagement.

With respect to SIAs, four of the most important factors are (1) change over time, (2) shared experience, (3) backstory, and (4) design as a social agent. Change can be “scripted”—such as variation in how the agent speaks or acts, activities performed,
or backstory revealed over time—as well as “unscripted,” such as personalizing different aspects of the interaction in response to the user’s behavior. Change over time has been shown to increase engagement and help maintain and build relationships (e.g. Kidd and Breazeal [2008], Bickmore et al. [2010], Lee et al. [2012b], Gordon et al. [2016], Kory-Westlund and Breazeal [2019b]).

Shared experience can be considered part of change and personalization. It can contribute to the sense that the agent “knows you” and help build a relationship. For example, prior work on long-term child–robot interactions has found that children responded positively to the robot referencing shared experience—for example, using their name, talking about activities performed together, mentioning facts learned about the child such as their favorite color [Kory-Westlund 2019]. Other work has found that including a memory system that can track and reference prior interactions with the user can be beneficial for engagement and positive affect (e.g., Kasap and Magenat-Thalmann [2010], Leite et al. [2017]).

How a SIA is introduced, the stories told about it, and the story told by it all influence human perception of the SIA and their behavior with it (e.g. Stenzel et al. [2012], Klapper et al. [2014], Darling et al. [2015], Kory-Westlund et al. [2016]). Backstory can also be used to add interesting variation to dialogue to help maintain interest and engagement over time, for example, as was done with the robot receptionist [Gockley et al. 2005]. The agent’s story can also be used to help shape users’ expectations about the agent through sharing the agent’s history, capabilities, and limitations. The story can be used to establish the agent’s character, in the same way we learn about other people through conversation and disclosure. This story can be told by people who lead interactions with the agent, such as experimenters, as well as by the agent itself during conversation.

Finally, designing SIAs from the ground up for social interaction with humans will go a long way toward maintaining engagement over time. Social design includes all aspects of the robot’s social behavior and communicative abilities—including whether and how it speaks, how it moves, its nonverbal behavior, and its social contingency. Designing a SIA from the ground up with social interaction in mind means considering how to make the agent’s facial expressions, movement, gaze, dialogue, and other behaviors understandable to humans. It enables the SIA to be responsive, expressive, and social. All these social behaviors contribute to people’s engagement with the SIA as a social other, as well as their trust, relationship, and engagement with it.

The design of social robots as social agents may be more important for agents that interact with children than those that interact with adults. For example, in a recent study of in-home use of a social home robot and a voice-only home assistant, Singh [2018] found that children were more drawn to the entertainment and
social capabilities of the agents, while adults were more interested in the agents' functionality and usefulness.

19.8.2 From Voice Interfaces to Engaging Conversation

Most of the SIAs mentioned in this chapter utilize rule-based forms of multimodal dialog flow [also see Pieraccini, Chapter 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021] of volume 1 of this handbook [Lugrin et al. 2021]; Traum, Chapter 15 on “Socially Interactive Agent Dialogue” [Traum 2022] of this volume of this handbook]. In many of the long-term health coach SIAs mentioned, the user chooses their response from a set of pre-defined options, and the agent decides its response based on rules set by the designer of the agent. This provides for a rather rigid social experience with the agent. More modern NLU approaches use a flow editor tool to design dialog flows and may use machine learning methods to train models to recognize different intents from user utterances and automatically generate responses (see Chapters 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021], 6 on “Building and Designing Expressive Speech Synthesis” [Aylett et al. 2021], 7 on “Gesture Generation” [Saund and Marsella 2021], and 8 on “Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]).

19.8.2.1 Flexible Dialog

For truly social dialog, SIAs must be able to move away from brittle rule-based systems to more robust and flexible approaches that allow for a more free form of conversation. There is a significant body of research in the natural language understanding and generation community that focuses on statistical machine learning approaches to generate next utterance from a given context and history of conversations by relying on a large corpora of conversations. However, these approaches fail when there exists no corpora for the context of the conversation. This is often the case for SIA research deployments. In such cases, a mix of rule-based and machine learning approaches have been proposed. Researchers have also employed approaches where responses are crowd-sourced, for example, by having people perform collaborative tasks in multiplayer game scenarios, and then applying machine learning or reasoning methods to generate agent responses and plan networks based on human utterances and behavior [Orkin and Roy 2008, 2009, Breazeal et al. 2013].

For example, Kennedy et al. [2017b] developed an embodied agent for social chitchat that could self-author responses and grow its knowledge and ability to respond over time. It has been deployed on a robotic platform and as a virtual
agent on a phone. They integrate ideas from rule-based, machine learning, and crowd-sourcing approaches for automatic dialog generation. In their approach, every conversation between an agent and user translates into a dialog tree that becomes a part of the agent’s graph database. When an utterance is received, the agent finds the nearest node in the graph (using cosine distance between average word2vec embeddings of words in the sentence) and randomly chooses a child node (i.e., the response to the utterance). When the agent fails to find a nearest neighbor (constrained by a threshold), it ends the conversation and attempts to grow its graph. Apart from the natural growth of the graph from having different conversations, the agent also crowd-sources responses so that it doesn’t fail the next time it is in the same state (i.e., at the same node). The model was evaluated in a 12-day study where users were encouraged to chat with the system multiple times a day. Results show that with time the number of conversations increased significantly with the number of failures significantly lower on the last day compared to the first day. There was also a trend of increased conversation length with time. These results are encouraging, and much more research in conversational AI with SIAs is needed for them to support robust and flexible social conversation that satisfies users expectations and desires.

19.8.2.2 Nonverbal Cues

Use of appropriate nonverbal cues such as gesture, gaze, displays of affect, behavior mimicry, and turn-taking can drastically improve how engaging, understandable, interactive, and believable conversation with SIAs can be. For example, use of nonverbal mirroring and behavioral mimicry increased a virtual agent’s likability and persuasiveness [Bailenson et al. 2005], and use of nonverbal mirroring and affective support decreased frustration and increased flow [Burleson and Picard 2007]. With robots, use of appropriate social cues, social contingency, nonverbal immediacy, vocal entrainment, and expressivity have led to increased learning and trust in the robot as an informant (e.g., Breazeal et al. [2016], Kennedy et al. [2017a], Kory-Westlund et al. [2017a, 2017b], Lubold [2017], Lubold et al. [2018]).

Nonverbal cues are important in establishing joint attention, trust, and rapport in both human–human relationships (e.g., Tickle-Degnen and Rosenthal [1990], Dijksterhuis and Bargh [2001], Lakin et al. [2003], Rotenberg et al. [2003], Dijksterhuis [2005], Harris [2007, 2012], Chartrand and van Baaren [2009], Semin and Cacioppo [2008], Wiltermuth and Heath [2009], Valdesolo and DeSteno [2011]) (also see Chapter 8 on “Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]) and in human–agent relationships (e.g., Bell et al. [2003], Breazeal [2002], Breazeal et al. [2016], Gordon et al. [2016], Levitan et al. [2016], Suzuki and Katagiri [2007]).
Many SIAs currently use nonverbal cues that are, at least in part, scripted—for example, rule-based systems that direct the agent to look at the user when speaking or look down at a shared work surface during pauses, display certain facial expressions upon detecting particular user facial expressions, use beat gestures when certain lines of dialogue are played back. More recent approaches use machine learning to automatically generate nonverbal cues (see Chapter 7 on “Gesture Generation” [Saund and Marsella 2021] of volume 1 of this handbook [Lugrin et al. 2021]) and even to learn personalized policies for nonverbal cue generation. An increasing number of comparative studies examine different nonverbal behavior generation systems and different methods of directing gaze or gesturing in an effort to determine more effective and natural ways for SIAs to communicate. Like with conversation and dialogue, to achieve truly social nonverbal behavior, we need to move from brittle rule-based systems to approaches that allow for more robust, free-form interaction.

One promising direction has been taken by Justine Cassell and colleagues. They have performed several long-term human–human studies in which they have collected data on how humans coordinate behavior during conversation and use nonverbal cues to, for example, establish rapport and a positive relationship [Cassell et al. 2007a, Sinha and Cassell 2015a, 2015b]. They have then used this data to build models for SIAs in human–agent interaction [Zhao et al. 2014, 2016], thus enabling the SIA to be more reactive and interactive in human-like ways.

### 19.8.3 Long-term Adaptation and Personalization

One important aspect of several long-term studies so far is personalization. For instance, tailoring educational content to individuals can lead to greater engagement and improved learning outcomes. This has been seen in HRI with children [Leite et al. 2012b, Kory and Breazeal 2014, Gordon et al. 2016, Palestra et al. 2016, Scassellati et al. 2018a, Park et al. 2019] as well as in other learning contexts, for example, with virtual agents or with older children and adults [Thrun et al. 1999, D’Mello et al. 2012, Kasap and Magnenat-Thalmann 2012, Leyzberg et al. 2014, Gordon and Breazeal 2015, Ramachandran and Scassellati 2015] (also see Chapter 18 on “Adaptive Artificial Personalities” [Janowski et al. 2022] of this volume of this handbook). So far, personalization has been studied far more often in longitudinal studies than in one-session studies. This is likely because nearly all personalization studies so far have focused on providing personalized educational content or feedback, using the results of the previous sessions to plan out the content or feedback types for the next session. For example, in Leyzberg et al. [2014], two different models of personalization were used to determine which lessons individuals received about how to solve logic puzzles over the course of four sessions.
One model tallied positive and negative demonstrations of a relevant skill; the other model used Bayesian updates to model the probability of mastering a relevant skill. They found that receiving personalized lessons significantly improved participants’ performance in the puzzle-solving task.

In the AutoTutor intelligent tutoring system, the system monitored students’ affective and cognitive states and selected actions to increase learning and help students regulate negative emotional states [D’Mello et al. 2012]. It modeled human tutor dialogue styles and used semantic matching algorithms and conversation rules to pick next dialogue moves in the curriculum script. It detected learning-centered emotions, including engagement, boredom, confusion, and frustration, using facial feature tracking, body posture measurements, and contextual cues. It provided feedback via the virtual tutor’s affective facial expressions and verbal responses. They found that the supportive tutor increased students’ deep learning, but primarily for low-domain knowledge students and only the first session—that is, after there was sufficient context to know the student had problems and actually needed support.

Leite et al. [2009, 2012b, 2014] studied how enabling a robot to express empathy and support during a chess-playing activity might increase children’s engagement over time. The robot detected children’s affect and made assessments about the child’s emotional state using facial expressions and the chess game’s state. It used this information to select appropriate supportive behaviors, such as providing advice or guidance, reinforcing the child’s sense of competence, or showing expressions of caring and empathy. It also stored information about prior interactions with the child and used reinforcement learning to learn what support strategies worked best with each child. In addition, in the earlier work [Leite et al. 2009], a human instructor chose level-appropriate chess exercises for each child. This work showed that personalizing the robot’s supportive behaviors to individual children increased children’s engagement and their ratings of the robot’s social presence and helpfulness.

Multiple studies were preliminary in that they presented personalization strategies but did not test them in full experimental studies or did not report all results as yet. For example, Serholt and Barendregt [2016] used information about children’s affective states to determine the pedagogical strategy. Although that paper did not report learning results, they found that children expressed significant social engagement and the robot’s personalization appeared to increase engagement. In a preliminary case study, Palestra et al. [2016] scaled up the difficulty of several social skills games played by three children with autism (e.g., about eye contact, joint attention, and body mimicry) and stopped leveling up when children
were unable to complete a task. Two of the children appeared to benefit from the leveling.

More recently, Park et al. [2019] conducted an eight-session study where 44 children aged 4–7 interacted one-on-one with a fully autonomous robot. The robot told stories and children were asked to retell the stories, thereby practicing language skills and learning new vocabulary. This work found that personalizing the robot’s story curriculum improved children’s engagement in the interaction that also led to higher vocabulary learning. Children were given language assessments prior to the study, which were used to select the first stories children heard and to select curricula for children in the non-personalized condition. In the personalized condition, children’s story retells, task behavior (e.g., answering dialogic questions during the robot’s narration), and affective arousal were used as input for a Q-learning algorithm, which selected personalized storybooks for each child at an appropriate syntactic and lexical level while maximizing for engagement and learning. In this study, it was observed that children who reported having closer relationship with the robot also achieved higher learning gains and vice versa, and this trend was more significant when the robot interaction was personalized to the child [Kory-Westlund et al. 2018]. This work showed how closer relationships between SIAs and their users can achieve higher long-term beneficial outcomes and further motivates why relational properties need to be considered in the interaction design of SIAs. All of the work so far on long-term interactions and personalization provides evidence for several takeaways. First, personalization of curricula, support, and feedback can improve students’ learning, engagement, and positive emotions. Agents that provide support and feedback may be seen as having greater social presence and as being more helpful. The relationships students developed with the agents appeared to influence their engagement and interest in further interaction. Including change and variation in the agents’ behavior and the learning content over time can also increase engagement and social interaction.

19.8.4 Achieving Long-term Beneficial Outcomes from Relational Properties

One difficulty in using relational properties in SIAs is determining which properties to use in a particular SIA, and how to design and use said properties effectively to achieve beneficial outcomes. Although research so far suggests that relational properties can indeed lead to, for example, increased user engagement, learning, or adherence to health-related programs, it is unclear which relational properties may be most helpful in promoting particular desired outcomes—or which may not contribute positively at all. While many studies found social contingency and social interaction being positively associated with increased trust and
learning (e.g., Breazeal et al. [2016], Kennedy et al. [2017a], Lubold et al. [2018], Kory-Westlund and Breazeal [2019b]), Kennedy et al. [2015] observed that an excessive amount of social behavior by a robot may detract from children’s learning during a math-learning activity.

In another example, Kory-Westlund [2019] found that using a variety of relational properties in an educational social robot leads to increased engagement and learning compared to a robot that did not use any relational properties. However, the benefits appeared to be moderated by children’s gender and overall affiliation with the robot—that is, children who formed a stronger relationship with the robot tended to engage more and learn more, regardless of whether the robot they played with was using relational properties.

SIs may need to be designed to use different behaviors—for example, dialogue, emotional reactions, ways of expressing information—in order to connect with different users and meet their needs in ways that work best for them. More research is needed to disentangle how different relational properties contribute to the achievement of different outcomes. For example, acknowledging shared experiences and showing mutual change/personalization may be more effective at creating a sense that the SIA “knows” you than responsiveness and use of appropriate emotion—or it may be that responsiveness and building rapport may contribute just as much, for different people, or in different situations.

Another challenge in using relational properties in SIs is measuring them and tracking the changes over time. So far, relationship measures in SIA research were heavily reliant on participants’ self-reported surveys or experimenter-conducted interviews. The Godspeed Questionnaire Series (GQS) is one of the most frequently used questionnaires in HRI [Bartneck et al. 2009, Weiss and Bartneck 2015]. The GQS consists of five scales that are relevant to evaluating the perception of the interaction with the robot: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Some studies have borrowed measures from social psychology to assess response to SIs, such as the Working Alliance Inventory [Horvath 1989] or the Dyadic Interaction System [Ickes et al. 1990, Broadbent et al. 2007]. However, these popular scales are focused on evaluating SIs as useful tools, and they lack metrics in measuring and tracking the change in the relationship and bonding between users and SIs. Few works have designed new metrics to specifically measure human–SIA relationships, such as measuring young children’s relational perception toward a peer-like robot learning companion [Kory-Westlund et al. 2018, Kory-Westlund 2019]. Kory-Westlund et al. [2018] proposed an interview protocol to measure children’s perception of robot as social-relational other and developed questionnaires to measure perceived closeness to the robot (inclusion of other in self), a narrative description task comparing
existing relationships with others, pets, and toys to relationship with the robot, and a self-disclosure task that measures if children would share something they are not good at to a robot that discloses its incompetence. However, most of these interviews and surveys were conducted by the experimenter pre or post to the study, and there still is a huge lack of automatic behavioral measures for relational properties. Zhao et al. [2014, 2016] showed some potential by building computational models of humans building, maintaining, and destroying rapport through the use of conversational strategies with verbal and nonverbal behaviors. Being able to measure relational cues autonomously during interactions will open up for much advanced SIA personalization and adaption for long-term beneficial outcomes.

19.9 Future Directions

In this section, we offer three future directions that will be important for SIA research to address in the coming years. The first is understanding and measuring when SIs are beyond the novelty period. This is a critical question for long-term SIA research: indeed, when is the SIA truly engaged in long-term interaction? Presently, this is poorly understood, standard metrics don’t exist, and long-term studies don’t quantify it. The second concerns advancing AI methods for SIs to realize more socially and emotionally intelligent relational agents with more sophisticated, flexible, and robust behavior. This also includes developing more tools to measure and assess the quality of relationship and its impact on human engagement and behavior. The third topic concerns the responsible and ethical design of relational SIA technologies. This is particularly important as many applications and intended benefits are designed to support vulnerable populations. And given the popularity of voice assistants, with millions of units already sold, this creates a glide path to relational agents in the near future. There is still much to be understood in terms of how our evolving relationships with SIs will shape our behavior, attitudes, policies, communities, and more.

19.9.1 Understanding Novelty

Currently, there is no consensus on what amount of exposure is needed to say that novelty has worn off, nor consensus on exactly what novelty is or how to measure it. Sung et al. [2009a] studied the adoption and sustained use of the Roomba in the home and found that stable usage patterns developed after the 2-month mark. de Graaf et al. [2017] studied the long-term use and reasons for abandonment of a simple desktop robot-information device over 6 months in the home. They found that participants had different reasons for ceasing use, and they caution against defining the novelty period based on a fixed amount of time. Rather, they recognize that the novelty period will be different for different devices and different use cases.
According to the psychology of behavior change literature, it takes about 2 months on average to establish a new behavior—however, the amount of time ranged from 3 weeks to a year depending on the new type of behavior being established [Lally et al. 2010].

It is also worth questioning whether novelty ever does completely “wear off.” In human–human relationships, there seems to always be the potential for some novelty in some level. While interactions may reach some kind of steady state with less novelty, there is potential, for example, for one’s spouse of 30 years to still cause surprises (e.g., “keeping things fresh”). Furthermore, published works in human–SIA interaction rarely measure when repeated encounters move past a novelty period. Given this is the case, how do we know that research about long-term effects is reporting accurate results?

For instance, is novelty different than unfamiliarity, and if so, how? What are we contrasting novelty with—familiarization? Habituation? Boredom? How does novelty relate to engagement? For example, we could define the novelty effect in SIAs as engagement that is due to newness rather than due to intrinsic qualities of a thing (e.g., a virtual agent, a robot, a talking speaker) being engaging or fun. Then the question is, when does boredom or engagement overtake engagement-from-novelty? However, novelty is not intrinsically associated with either positive or negative valence and could lead an individual into a curious/interested state or a state of threat/risk [Gillebaart 2012]. There are likely many individual differences in preference for seeking novelty, for example, children may prefer novel toys and pictures over familiar ones, which may promote development and acquisition of new concepts [Gillebaart 2012]. Boredom-prone people may be more focused on novel experiences and may find them more interesting. Given this, defining novelty in terms of engagement may not make sense.

One useful framework for examining novelty may be Novelty Categorization Theory (NCT). NCT suggests that appraisal of events as novel relates to categorization, in that an event is novel if it does not fit in any existing categories one has [Förster 2009, Förster et al. 2010]. Novel events are processed in a more global processing style that uses broader, more inclusive mental categories in order to assimilate the novel information and integrate it into existing mental categories or knowledge structures. Assimilated knowledge becomes more familiar and likeable. We could use this framework to evaluate people’s categorizations of robots. For example, there are multiple tasks that measure global versus local processing on which people have performed differently when the task is framed as novel versus as familiar, such as the Gestalt Completion Task discussed in Gillebaart [2012]. Perhaps one could administer these kinds of tasks either during or following a SIA interaction to learn whether people are using a more global or local processing...
style, and thus, whether they are perceiving the activity with the robot as more or less novel.

Clearly this is a very important topic for long-term interaction research with SIAs. Advancing how we define and assess novelty (or when an interaction is past the novelty phase) is important work that is relatively unexplored in long-term human–SIA work. In the future, any paper claiming long-term effects should empirically determine that the findings are not due to novelty.

19.9.2 Relational AI for SIAs

A great deal of work remains to be done to advance relational AI to improve the socio-emotional intelligence and response of relational SIAs over longitudinal encounters to help people collaboratively achieve long-term goals. There is active research in developing and studying key real-time behaviors such as gaze, reciprocity, entrainment, dialog patterns, affective response, and the like that contributes to building rapport, trust, and working alliance. Algorithmic methods need to be advanced to make these behaviors more flexible, robust, adaptable, and socially appropriate in terms of style and timing. Relational behaviors, for instance, can be divided up in many different ways: by timescale (e.g., behaviors developing in the present on shorter timescales in matters of seconds or minutes vs. behaviors that develop over longer times, such as days, weeks, or years), or by modality (e.g., verbal vs. nonverbal cues, linguistic vs. non-linguistic). Investigating how these contribute to building an engaging and appropriate relationship over time—beyond days and weeks to months and even years—is needed. Adaptation to context also matters, whether that is about the task, the larger social context such as the setting, or specifics about the individual.

Much work remains to be done on computational methods for personalization to an individual’s needs and differences, as well as adaptation to changing contexts. What can and should a relational SIA learn and remember about you (or forget)? Perhaps the user should have control over this explicitly. The memory capability of SIAs raises both technical and ethical issues. What kinds of memory are needed—for example, memory of facts, events, emotions, and personal preferences or other details? At what point in time should the relational SIA probe a person again to gain more information or reassess what it knows about the user? If people grow and change, the relationship will need to as well. Part and parcel of learning and adaptation is data capture, privacy, security, and data ownership issues.

Also, the automatic perception of relational properties in the interaction and tracking changes over time is a remaining challenge. In human-to-human interactions, use of verbal and nonverbal expressions are direct observable cues that reflect the relationship between the involved parties [Canary and Stafford 1994].
When relationship changes over time, so do the use and style of language, prosodic cues, and facial and body expressions. While much prior work relied on interviews and self-reported surveys for SIA relationship measures, being able to detect and track the behavioral relational properties in real time during interactions through SIA will impact the way agents can personalize and adapt its behavior policies for better user engagement and greater beneficial interaction outcomes in the long term.

These kinds of question show that we need better ways of measuring and assessing the state and quality of the human–SIA relationship. Such measures need to also be designed to be appropriate for the human counterpart (e.g., considering age, task, setting, personality, and other contextual factors). At what point do relationships move from novelty to familiarity and habituation, and how does novelty continue to play a role in relationship continuation (e.g., to keep the relationship interesting)?

19.9.3 Ethical Issues and Design Practices

The promise of relational SIAs that can help and support people in humanistic, high impact ways is a commonly held goal by many researchers in the community. Because of the distinct human-engagement of relational SIAs, they have the potential to engage diverse people in innovative ways across many domains: education, therapy, healthcare, and so on. This also comes with ethical concerns about their appropriate and responsible use as well as potential for (unintended) misuse. Relational SIAs such as social robots, embodied agents, and personified smart devices raise many different ethical concerns—most of which are also encountered in other technologies and domains—all at once. Many of the ethical concerns are most contentious with vulnerable populations, such as children and older adults, who potentially also have the most to benefit from relational SIAs. All of these concerns are pressing given that social and relational technology is swiftly entering the market. We highlight a few of these ethical concerns below as well as some ethical design practices to consider, especially when working with vulnerable stakeholders.

19.9.3.1 Social Bonds and Authenticity

One concern that has been expressed about relational technology is that it will replace the social bonds people have (or would have had) with other people [Turkle 2007, 2017]. One part of this concern pertains to deception—that is, whether relational SIAs are deceptive in their display of relationship, emotions, and empathy, causing people to think, act, and believe that they have emotional and relational capabilities that they do not “really” have [Picard and Klein 2002, Turkle 2007,
Questions about deception and authenticity are, at the heart, about the effects of deception on people—that is, deception is a problem because it causes harm to people. One possible harmful effect relates to human attachment to and reliance on relational technology. Will we come to depend on it too much, when we should not, to our social detriment (e.g., Turkle [2007])?

Coeckelbergh [2012] argues that what robot ethicists really mean when arguing about emotional deception is either (1) that the robots intend to deceive, (2) that the emotions robots have are not real, or (3) that the robots pretend to be a kind of entity they are not. In the first case, he argues that it is not the robot that intends to deceive but the robot designer, and that designers have a long tradition across many disciplines (literature, video games, movies, etc.) of creating believable characters. No one is fooled that these characters are “real,” though, or if they are, this is generally considered an acceptable kind of deception based on widespread prior art. In the case of relational SIAs, the question is whether people are fooled—and then, whether this is necessarily a bad thing. Research has reported numerous benefits from the relational properties of SIAs, from boosting children’s learning outcomes, improving engagement in health protocols, serving as a catalyst to promote human–human connection in assisted living facilities, and more. People have deep conversations with chatbots and virtual therapists [Bobicz and Richard 2003, Bickmore et al. 2005, Pontier and Siddiqui 2008]; often, people consider these agents less judgmental than humans [Bickmore et al. 2005, Gratch et al. 2007, Lucas et al. 2014, Utami et al. 2017].

When researchers do probe whether these users actually believe that relational SIAs have human-equivalent emotions, we see nuance in their answers, even from young children [Kory-Westlund et al. 2018, Kory-Westlund and Breazeal 2019a]. Thus far, that data suggests that people do not see SIAs as a relationship human-equivalent but more of a “sort-of” comparison. Young children, for instance, treat robots as social others, apply social judgments to robots, and respond to their social cues in ways similar to how they respond to people. Nonetheless, when explicitly assessed, children seem to place robots in a different “in-between” ontological category than either living or non-living things [Severson and Carlson 2010, Kahn et al. 2011, Gaudiello et al. 2015]. They have shown a moral objection to the object-like treatment of robots, such as putting a robot away in a closet, because of the perception they have formed about the robot as a social other [Kahn et al. 2012]; however, they may also say that like other objects a person made the robot, people can own robots, and that robots can break [Kory and Breazeal 2014, Kory-Westlund et al. 2016]. In several studies in the early 2000s, children categorized the robot dog Aibo as not a dog and not a robot but as a “robotic dog”—a dog with robotic attributes [Kahn et al. 2002, Bartlett et al. 2004, Melson et al. 2009, Weiss et al. 2009].
What does it mean to have “real” or “authentic” emotions in the first place? Sherry Turkle [2007], for example, has argued that social robots are inauthentic: they may provoke emotional attachment, trust, caring, and empathy that is not deserved because the relationship and the feelings are not reciprocal. Must a relationship be reciprocal in a human or equal way? Reciprocity in equal measure is not a requirement even of human relationships. People are capable of having many different kinds of relationships, simultaneously: with peers, our children, our parents, our pets, and so on. Human–SIA relationships may simply be one more different kind of relationship that we are still figuring out. Given that SIAs are becoming increasingly mainstream (e.g., Alexa, Siri, the Google Assistant), an important area of ongoing research is to turn these into empirical questions to understand what effects these relationships with relational AI actually have. These are far from simple topics where data captured about human behavior with relational SIAs often reveals greater nuance and complexity than expected. Speculation or theorizing about the goods and bads of authenticity, attachment, and deception are no longer enough.

### 19.9.3.2 Persuasion and Social Manipulation

Another ethical concern pertains to social manipulation and persuasion. Technologies often mediate and implicitly shape human interaction with and perception of the world, by encouraging or inviting some forms of actions while discouraging or inhibiting others [Verbeek 2006]. As we have discussed, some research has focused on creating SIAs for behavior change in health contexts—for example, to help someone with particular weight loss goals to stay on track or engage isolated older adults, among others [Kidd and Breazeal 2008, Bickmore et al. 2018, Sidner et al. 2018]. Social robot learning companions that exhibit curiosity [Gordon et al. 2015], creativity [Ali et al. 2019], or a growth mindset [Park et al. 2017b] have been shown to promote the same behaviors and attitudes in children. This kind of change is usually considered acceptable: it is a “positive” change, with the goal of helping people achieve what they want to achieve. When used for “good,” then persuasion, in a SIA or in a human, is often seen as a positive attribute—it gets us to the end we want. If used for “bad,” it is another story altogether. For example, robots that are used to provide the elderly with shopping assistance may be seen as beneficial [Iwamura et al. 2011], but those that target potential customers may raise some eyebrows [Kanda et al. 2008]. When SIAs enter people’s homes from corporations who want to nudge human behaviors to increase profits certainly raises concern, in addition to concerns around data privacy, security, and use of personal
data. The IEEE guidelines for Ethically Aligned Design address some of these issues.\textsuperscript{1}

People are socially manipulative and persuasive all the time with each other—this is part of how social interaction works. However, while it may be acceptable for the car salesman to use conversation and rapport tactics to up-sell expensive features for a new car, an SIA that does the same thing could be considered alarming. The question here is whether being socially manipulative or being persuasive is acceptable for technology, and if so, to what degree? Verbeek [2006] argued that persuasion is not an intrinsic property of any technology but comes from both the designer and the user. He argued that we should assess potentially persuasive technologies on three fronts: (1) whether the intended persuasions are morally justifiable, for example, that they do not cause harm, and promote beneficence or justice; (2) that the methods of persuasion used are morally acceptable, for example, that they respect human autonomy; and (3) that the outcomes or consequences of persuasion are morally justifiable. The biggest challenge, here, however, is that people are often not going to agree on what is considered morally acceptable or morally justifiable.

However, we can look to other domains for inspiration on how to handle these ethical quandaries. Marketing and advertising are two domains that frequently raise similar questions about social manipulation using human-made artifacts and face similar challenges regarding lack of consensus about what ethical behavior is [Drumwright and Murphy 2009]. Some marketing agencies have adopted codes of ethics promoting transparency, honesty of relationships, opinions, and identity—that is, promoting the idea that they should make sure consumers know when they are being advertised to. Relational SIAs could follow this example of promoting transparency and honesty, for example, using backstory or dialogue to explain to users what it is capable of and what its goals are for others’ behavior. How the SIA talks about itself can continue to remind and reinforce this transparency and appropriate relationship, such as not answering certain questions or doing certain tasks on the grounds of those being as “only appropriate for humans”—an approach adopted by the design team of Jibo, a social robot for the home. Informing and reminding users through the interaction design of relational SIAs can be helpful, but there are additional questions we can raise about how much users can really trust the designers of the SIA technology. Who is held responsible for the behaviors of users, how can we be sure that a technology will not have undesired

\textsuperscript{1} https://standards.ieee.org/industry-connections/ec/autonomous-systems.html, retrieved February 7, 2019.
or unintended effects, and whether being persuasive is in itself an ethical thing for an SIA to do? The development of design guidelines, best practices, and policies are all important areas of ongoing work as SIAs move into the human environment for longer and longer periods of time.

### 19.9.3.3 Privacy and Security

SIAs for long-term interaction will need to collect data about users—but what data, how much, and how will it be stored and protected? For example, we raised the question earlier of SIAs needing memory of facts, events, emotions, personal preferences, or other details in order to perform their tasks, learn and adapt, build relationships, and maintain relationships over time. For a SIA, memory is data. The capability of SIAs to monitor and surveil beyond the capacities of human sensing (e.g., through the use of infrared or ultrasonic sensors, or at ranges or distances unavailable to humans on their own) is concerning. Ryan Calo [2010] describes three areas of privacy that we should be concerned about: direct surveillance—that is, SIAs that magnify the human capacity to observe; increased access—for example, new access to historically protected spaces, like inside homes; and social meaning—for example, people may act differently as a result of feeling observed and evaluated.

These issues are not unique to SIAs—they arise with many current technologies such as laptops and devices in the IoT [Arnold 2010, Goldman 2015]. Beyond issues of privacy, we also need to be aware of security in how data is collected, transmitted, and stored: for example, data breaches are increasingly common. Finding satisfactory solutions for issues in data capture, security, privacy, and data ownership will likely require joint action from governments and regulatory bodies regarding, for example, what surveillance is acceptable in different circumstances (e.g., by SIAs in public spaces), accountability for anyone dealing with protected data (e.g., legal consequences for negligence in protecting user data, similar to HIPAA-protected data), and imposing standards regarding security, encryption, data forensics, and so forth. We may also need designers of SIAs to adopt an ethical code similar to codes that professionals in other fields follow that emphasizes privacy, accuracy, intellectual property, and access [Calo 2010, Riek and Howard 2014].

### 19.9.3.4 Ethical Design Practices

These ethical design questions are difficult to answer. There is compelling opportunity to include philosophers and ethicists more directly in the design of future relational SIAs. We also need designers of SIA technologies to be aware of the ethical and moral issues involved in the things they are creating and to attempt
whenever possible to create technology that supports and affirms people in becoming who they want to be—that supports human flourishing. Toward this goal, we offer a few ethical design principles of relational SIAs:

— Design responsibly. When designing new technology, involve philosophers and ethicists who have specific training in relevant ethical and moral frameworks and applications. Also include domain experts in the application area and user demographic to be engaged. Design with empathy and human flourishing in mind—companies are often criticized for designing for “addiction” or only to maximize profit. Rather than creating technology that serves as a “crutch” that people may become over-reliant upon, consider how to design technology that empowers and respects human agency and dignity.

— Be informed by data as well as theory. An increasing number of research studies are exploring questions highly relevant to the ethical design of relational AI, such as questions about engagement, trust, and attachment. We need to use the data from both human–human studies and human–agent studies to learn how people actually form relationships, develop trust, and interact with relational agents, and use these data to inform future design.

— Co-design and involve all stakeholders. There are emerging disciplines in the area of ethical design where stakeholders have an important role and voice at the design table. Work in the area of participatory design and design justice are becoming increasingly relevant and important as SIAs move from research labs into human environments for longitudinal time frames.

— Be transparent and honest. Inform users about what a technology can do and what it will do. Use the technology’s packaging, introduction, framing, and backstory to share information and set user expectations appropriately about the technology, its capabilities, and its limitations. Verify that users actually understood the technology’s capabilities and limitations.

— Implement security and privacy by design as well as safety by design. Collect only data that are needed for the agent to fulfill its tasks, only data that can be sufficiently protected, and only data that are acceptable to users. Be transparent about what set of data are collected, how data are stored and transmitted, and how data are used.

19.10 Conclusion

Relational SIAs have great potential to support people of all ages in areas that can profoundly contribute to quality of life and opportunity—from education and
lifelong learning, to health and wellness, and more. These are all long-term endeavors for people, and each can benefit from having a strong relationships with a supportive ally. Relational AIs have the potential to offer personalized, high-touch support that is far more scalable, accessible, and affordable than hiring human professionals. However, this also raises critical ethical and societal questions as to the responsible and appropriate use of relational AIs—to ensure that this technology empowers people to achieve important personal goals and supports our human networks and stakeholders in the process. Our goal should be to use relational SIAs to help all people flourish, to augment and support human relationships, and to enable people to be happier, healthier, more educated, and more able to lead the lives they want to live. Much work remains to be done, and many questions and issues need to be understood and addressed. Long-term SIA research continues to make exciting progress with the potential for positive, transformative impact for quality of life for many.

Acknowledgments
We want to thank the members of the Personal Robots Group at the MIT Media Lab who helped to survey papers on long-term interaction with social robots. Part of this research was supported by an MIT Media Lab Learning Innovation Fellowship and the National Science Foundation (NSF) under Grants IIS-1122886, IIS-1122845, IIS-1123085, IIS-1523118, CCF-1138986, and a Graduate Research Fellowship under Grant No. 1122374.

References


Chapter 19  Long-Term Interaction with Relational SIAs


References


Chapter 19  Long-Term Interaction with Relational SIAs


References


Chapter 19  Long-Term Interaction with Relational SIAs


T. Kanda, D. F. Glas, M. Shioimi, H. Ishiguro, and N. Hagita. 2008. Who will be the customer?: A social robot that anticipates people’s behavior from their trajectories. In
Chapter 19  Long-Term Interaction with Relational SIAs


Chapter 19  **Long-Term Interaction with Relational SIAs**


References 253


A. Pradhan, L. Findlater, and A. Lazar. 2019. “Phantom friend” or “just a box with information”: Personification and ontological categorization of smart speaker-based


Motivation

Developing a Socially Interactive Agent (SIA) with the goal of simulating complex human behavior in all of its intricacies and nuances is a daunting undertaking. It requires not only in-depth knowledge of individual research areas, including specialized fields within computer science, artificial intelligence, social sciences, art production, game development, and psychology, but also how these fields interconnect, both theoretically and practically. These interdisciplinary requirements go beyond the capabilities of individuals or even teams.

Fortunately, there are many tools and platforms available that researchers and developers can take advantage of, allowing us to:

— Leverage previous work, either to expand upon or to provide context for their own work.
— Contrast and compare approaches and implementations.
— Avoid starting from scratch and reinventing the wheel, reducing time and costs.
— Explore the various problem spaces within SIA collaboratively.
— Define standards, which enhance interoperability.

Research and development is becoming ever more complex. Platforms and tools allow both researchers and developers to collectively advance our understanding of and capability in the exploration and creation of SIAs in an ever more efficient
and effective manner. In this chapter, we provide an overview of commonly used software solutions to the different aspects of behavior and interaction described earlier in this handbook.

### 20.2 Overview

We start by discussing the history and trends of using tools and platforms, including interoperability, hardware, and distribution. Section 20.4 provides an overview of common platforms while Sections 20.5 and 20.6 delve deeper into individual tools for creating characters and character interactions, respectively. Our main focus is on Intelligent Virtual Agents (IVAs), but we briefly discuss how many of the discussed areas apply to Social Robotics (SRs) as well in Section 20.7. We end with a discussion on current challenges, future directions, and a summary.

### 20.3 History and Trends

While research and development in artificial intelligence (AI) is typically acknowledged to have started in the 1950s [Crevier 1993], initial enthusiasm gave way to an “AI Winter” during the 1970s and 1980s [Hendler 2008]. In the early 1990s, interest in more holistic approaches grew, with a focus on embodiment and being part of the real world, often utilizing robotics [Brooks 1991]. These techniques would be combined with real-time computer graphics and a focus on simulating human-to-human interactions to give birth to the field of embodied conversational agents [Cassell et al. 2000]. Within the context of platforms and tools we see three main trends, which we will explore below:

1. **Interoperability**: from relatively isolated research efforts toward broader collaboration and the development of standards.

2. **Hardware platforms**: from supercomputers to desktops to multiplatform solutions, including mobile, web, Augmented Reality (AR) and Virtual Reality (VR).

3. **Distribution**: from dedicated research websites and propriety software to standard open source sharing platforms and commercial web services.

text-to-speech generation. These isolated efforts made collaborative research, integration, and sharing challenging, in particular because approaches would not align [Gratch et al. 2002]. As a result, standards were proposed and iterated upon. Examples include Knowledge Query and Manipulation Language (KQML) [Finin et al. 1994], Speech Synthesis Markup Language (SSML) [Taylor and Isard 1997], Virtual Human Markup Language (VHML) [Marriott 2001],1 is a wonderfully preserved artifact of the early aughts and the authors encourage any interested reader to take a trip down memory lane. Affective Presentation Markup Language (APML) and Discourse Plan Markup Language (DPML) [De Carolis et al. 2002], MPEG-4 facial animation [Pelachaud 2002], Avatar Markup Language (AML) [Kshirsagar et al. 2002], Multimodal Utterance Representation Markup Language (MURML) [Kranstedt et al. 2002], Character Markup Language (CML) [Arafa and Mamdani 2003], Artificial Intelligence Markup Language (AIML) [Wallace 2002], Web3D Consortium’s HAnim [Web3D 2006], and Behavior Markup Language (BML) [Kopp et al. 2006]. Out of these, SSML and BML are the main standards still in use today. SSML allows a character’s utterance to be marked up to indicate how its speech should be generated. BML describes the nonverbal behavior for a character and is based on a large academic collaboration of integrating previous standards. It is part of the SAIBA framework [Kopp et al. 2006], which also includes the Function Markup Language (FML) [Heylen et al. 2008]. Most of these are not true standards in the traditional sense of the word; they have not been ratified by official bodies but instead are common formats often used by researchers in the field of SIA. For more details on multimodal interaction architectures, see Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022] of this volume of this handbook.

For hardware platforms, computing power was still relatively scarce around the turn of the millennium, which led to either basic applications or the need for large, distributed computing servers [Rickel et al. 2001] with specialized and proprietary functionality and content [Rickel et al. 2002]. Increasingly powerful desktop systems resulted in ever more powerful applications being able to run on personal computers [Swartout et al. 2006]. The web has seen early adoption of SIAs [André et al. 1998, Evers and Nijholt 2000, Noma et al. 2000] and supporting tools [Bickmore et al. 2009]. Smartphones and tablets have seen a range of SIAs [Bickmore et al. 2010, Doumanis 2013] as well as tools to support its development [Klaassen et al. 2012, Feng et al. 2015]. Finally, a sizable effort is currently focused on AR and VR [Holz et al. 2011, Hartholt et al. 2019a]. The move toward ever more personalized, pervasive, and immersive computing devices has resulted not only in

---

1. The website https://www.vhml.org
the democratization of computing but also in a proliferation of tools that support the development of SIAs, in particular the use of game engines (e.g., Unity, Unreal Engine) (see Section 20.4.2).

Early SIA *distribution* methods often relied on providing software binaries or source code through university or personal websites, followed by centralized services that include version control, bug tracking, and documentation, predominately SourceForge in the 2000s [Howison and Crowston 2004] and GitHub in the 2010s and beyond [Kalliamvakou et al. 2014]. While the open source philosophy initially was mainly supported by researchers and a small number of companies, open sourcing software and data is becoming increasingly commonplace, with traditionally closed companies opening up portions of their IP in order to leverage the advantages of community-based development, including Microsoft,² Apple,³ Facebook,⁴ Epic Games,⁵ and Unity.⁶ At the same time, the proliferation of personal assistants and related “AI” capabilities has resulted in turning previously challenging areas into commodity technologies (e.g., speech recognition, text-to-speech). These are often available in the form of online services, offering researchers and developers a range of interconnected capabilities they can leverage in designing and developing SIAs.

These three trends have led to the democratization of many capabilities, which in turn leads to more technical solutions, increased competition, and improved accessibility.

### 20.4 Agent Platforms

We define a platform as having a suite of capabilities, which are integrated in a principled manner, that together cover several required features of an SIA, and that should be extendable. Based on these criteria, we define three categories of platforms, which we will explore below:

1. **Cognitive architectures**: principled approaches to simulate aspects of the human mind.
2. **Commercial platforms**: privately developed game engines to create animated characters and their behaviors.

---

(3) Academic platforms: integrated systems built by research organizations that cover many SIA-specific capabilities.

20.4.1 Cognitive Architectures

Cognitive architectures typically only cover the mind of an intelligent agent but many have incorporated perception of and acting in the world as explicit notions. This makes them suitable candidates to integrate them with SIAs. We will cover several of the more commonly used architectures from a practical point of view. For a more in-depth discussion of cognitive architectures, see Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022] of this volume of this handbook.

Soar is one of the early cognitive architectures, designed and developed by Allen Newell, John Laird, and Rosenbloom in 1983 [Laird 2012]. Soar is a general cognitive architecture for developing systems that exhibit intelligent behavior using symbolic reasoning. It is goal oriented and uses operator rules within a problem space to select its next action which affects the overall state. It includes procedural, semantic, and episodic memory that work together with short-term memory and learning mechanisms. Soar started off using general-purpose mechanisms, but later versions incorporated dedicated modules to better support specialized functions, including emotions and visual sensing. It has been used to develop intelligent conversational agents [Rickel et al. 2001, Swartout et al. 2006] and robotics [Laird et al. 2012]. Soar can use several external programming languages, including C++, Java, Python, and TCL, through the Soar Markup Language (SML). It includes a suite of supporting development tools, including visual and command line tools. It is available under a modified BSD 2-clause license. See https://soar.eecs.umich.edu for more details.

ACT-R was originally developed by John Anderson [Anderson et al. 1997] who was heavily influenced by Allen Newell and Soar. ACT-R uses symbolic reasoning, declarative and procedural memory, and perception and motor modules to interface with the world. Production rules fire when matching certain states which in turn affect the overall state of the system. ACT-R very explicitly bases its cognitive assumptions on results derived from psychology experiments and allows researchers to collect quantitative measures that can be directly compared with similar measures obtained from human participants. It has been used most notably in an intelligent match tutor [Ritter et al. 2007]. It is developed in Lisp and can interface with Python and other languages using JSON. ACT-R is licensed under LGPL v2.1. For more details, see http://act-r.psy.cmu.edu.

OpenCog was initially released in 2008 by Ben Goertzel and David Hart [Hart and Goertzel 2008]. It aims to achieve general intelligence through tightly integrated
cognitive algorithms, a strategy dubbed cognitive synergy. OpenCog includes native support for natural language processing, reasoning and inference, embodiment, and psychological states. It includes a symbolic knowledge representation that multiple cognitive processes can work on, called AtomSpace, which is a graph-based knowledge representation database with a query and reasoning engine. Applications have been developed both for IVAs and robotics. OpenCog uses C++, Python, and a custom language in support of AtomSpace called Atomese. It is available under the AGPL license (see https://opencog.org).

**TinyCog** was released in 2015 by Frank Bergmann and Brian Fenton [Bergmann and Fenton, 2015], following the tradition of Soar and ACT-R. It specifically aims to provide a minimalist implementation of a cognitive architecture, which makes it a good starting place for newcomers even though it is no longer in active development. TinyCog uses what they call Scene Based Reasoning, in which scenes represent real world environments on which plans can be executed. The main focus is on robotics with implementations for perception, actions, and learning. TinyCog is written in ProLog, licensed under GPL v3, and can be found at http://tinycog.sourceforge.net.

**Sigma** ($\Sigma$) is one of the more recent efforts and is headed up by Paul Rosenbloom and Volkan Ustun [Rosenbloom et al., 2016]. Its design and development is driven by four goals: grand unification, generic cognition, functional elegance, and sufficient efficiency. It aims to achieve this by combining traditional cognitive architecture concepts with the use of factor graphs [Kschischang et al., 2001] as the main underlying mechanism. Sigma primarily targets IVAs and has developed several sample projects that include perception, reasoning, learning, emotions, and natural language processing. Sigma is written in Lisp and available under the BSD 2-clause license at https://cogarch.ict.usc.edu.

### 20.4.2 Commercial Platforms

There are many individual commercial tools available for use within SIA, which we will cover in Sections 20.5 and 20.6, yet not many offerings are available that combine these into platforms specifically in support of SIA development. However, game engines do provide a range of integrated functionality, in particular for IVAs. Modern game engines do not only cover rendering but also animation, sound, networking, and so on. They typically cover multiple hardware platforms (e.g., mobile, web, AR, VR) and offer flexible development environments. While they are mainly focused on game development, they are general purpose tools that are very useful in creating IVAs.

The two most popular game engines are **Unity** and **Unreal Engine**. Unity started off as a game engine for beginners and smaller developers, while Unreal Engine
targeted large, professional teams. As a result, Unity is easy to pick up, allows for rapid iteration, has a large community and asset store, and excellent multiplatform support. Unreal Engine on the other hand has a longer history and shines in creating higher fidelity graphics—at the cost of an increased learning curve—and is completely open source. Unreal Engine and Unity are in fierce competition with each other, which ultimately benefits researchers and developers. Both engines are now free for academic use and small developers. Unity continues to catch up in terms of visual fidelity, while Unreal Engine provides more and more functionality to streamline development. Both typically require at least some level of programming to create IVAs. We will discuss each game engine in more detail below. Afterwards, we point toward several alternatives.

20.4.2.1 Unreal Engine

Unreal Engine is developed by Epic Games, who started licensing it in 1996. Epic Games, unlike Unity, develop their own games in addition to the engine. The most well-known of these are Unreal, Gears of War, and Fortnite. Unreal Engine can be obtained at https://www.unrealengine.com and is free for non-commercial use. Access requires an account, including for use of the source code.

Unreal Engine is a powerful game engine used for many high-fidelity video games on PC and consoles. It offers all of the main required functionality out of the box, including networking, GUIs, animation, physics, and audio. Unreal Engine no longer has a scripting language but instead uses either C++ or a visual scripting language called Blueprints Visual Scripting. This approach is one indication of Unreal Engine traditionally focusing on large, professional teams, where programmers would do the core development and create tools and templates for designers.

While Unreal Engine supports most common platforms, including, mobile, web, AR and VR, the implementation is typically less robust and user-friendly than Unity. The Unreal Engine Editor offers many graphical interfaces for most areas, though, and usability is improving. There are many tutorials available, including a dedicated section for developers transitioning from Unity.

Unreal Engine particularly shines in graphical fidelity (see Figure 20.1). It offers powerful tools and shaders in order to fully customize the look and feel of the environment, characters, and objects, which has made it a successful tool for non-gaming application, including architecture, previz production, and advertisement. Unreal Engine now offers easy to create, high-fidelity characters through its MetaHuman tool, although using these as conversational agents is as of yet non-trivial. Epic Games’ massive success with Fortnite has resulted in solid multiplatform and multiplayer capabilities, as well as enough income to grow the engine. Unreal Engine is released several times a year, with version 5 released in 2022.
20.4.2.2 Unity

Unity can be obtained at https://unity.com through a subscription model. There is a free version for students and individuals, with professionals and larger teams paying a monthly fee. All versions are technically equivalent, with the main differences lying in online features, team communication features, technical support, and available assets. Unity has open sourced portions of their product at https://github.com/Unity-Technologies, but the source code of the main engine, written in C/C++, can only be obtained through a paid license.

The Unity Editor is a development environment that allows visually manipulating game objects that can have C# scripts associated with them. This enables a range of developers with different levels of technical expertise, including designers, artists, and programmers, to be productive. Unity supports graphical user interfaces (GUIs), audio, networking, pathfinding, and character animation out of the box for both 2D and 3D projects. Its animation system, originally called Mecanim, is quite powerful, offering graphical state machines and blending parameters. It offers solid tutorials and has a wide community, which—while varying in skill level—offers lots of help as well as scripts and assets through the Unity Asset Store. Unity itself offers many online services, for example, cloud build solutions for all platforms it supports, which includes Windows, Mac, Linux, iOS, Android, WebGL, and major game consoles. Unity's robust multiplatform support has resulted in 69% to 91% of VR and AR applications being developed with Unity [Marvin 2018].
Looking toward the future, Unity has implemented machine learning tools as a separate package, called ML-Agents. This allows for reinforcement learning (based on TensorFlow), imitation learning, and other methods using a Python API, with Unity visualizing agents and their environments. Unity keeps advancing their rendering pipelines in order to improve graphical fidelity and to provide developers with more control. See Figure 20.2 for one of their demo characters. Unity is also working on enhancing performance by moving from an object-oriented to a data-oriented design of the core engine, an approach it has called Data-Oriented Technology Stack (DOTS). DOTS uses a range of techniques, one of which is the Entity-Component-System (ECS) architectural pattern, which decouples aspects of real-time simulation (e.g., graphics, physics, AI) in order to process entities more efficiently and to support safe multithreading.

Starting in 2020, Unity releases major new versions two times a year (e.g., 2020.1, 2020.2) and supports a final yearly release for up to 2 years (e.g., 2020.3). A unity roadmap can be found at https://unity3d.com/unity/roadmap.

20.4.2.3 Honorable Mentions

There are numerous other game engines available, of which we will briefly discuss a selection.

Id Tech (https://github.com/id-Software) is a series of game engines by Id Software that started after one of the first popular 3D games, Wolfenstein, with Doom (1993) and Quake (1996). Older versions up till Id Tech 4 are released under the GPL license.

CryEngine (https://www.cryengine.com) is a high-fidelity game engine that was originally released in 2002. It has lost some popularity in recent years, partly due to a lack of funding for continued development. It is still used for professional game development and aims to make a comeback. Part of this strategy is to release the engine open source (https://github.com/CRYTEK/CRYENGINE), freely available for non-commercial use.

Panda3D (https://www.panda3d.org) was originally created by Disney and made open source in 2002. It is written in C++, uses Python, supports multiple platforms, and is currently available under the BSD license.

The Source Engine (https://developer.valvesoftware.com/wiki/SDK_Installation) was originally released in 2004. It is created by Valve, co-creator of the HTC Vive VR headset. It is not used much beyond Valve games itself (e.g., Half Life, Counter Strike, Team Fortress, DOTA 2) and only offers an SDK rather than the full game engine, but recent VR interest may change this.

Godot (https://godotengine.org) is relatively recent, with an original release date of 2014, and aims for smaller projects. It offers a user-friendly development environment, supports C++ and C#, is multiplatform, and is open source under the MIT license.

Lumberyard (https://aws.amazon.com/lumberyard) is developed by Amazon and released in 2016. It uses CryEngine as a foundation, maintaining its high-fidelity approach while aiming more broadly at developers to create virtual worlds rather than just games. It integrates with both Twitch—a popular game streaming service—and Amazon’s AWS cloud services. Its source code is available at https://github.com/aws/lumberyard.

Amazon Sumerian (https://aws.amazon.com/sumerian) is an online web authoring tool for creating web, AR, and VR experiences that includes interactive characters, originally released in 2017. It integrates with many of Amazon’s own services in order to cover a range of SIA related capabilities, including speech recognition, natural language processing, nonverbal behavior generation, and text-to-speech. The development is web-only, using a visual editor and JavaScript, and as such not very customizable or expandable. It can, however, be combined with 3rd party software [Monteiro and Pfeiffer 2020]. Amazon Sumerian requires a monthly subscription, based on the level of services used.
20.4.3 **Academic Platforms**

Few academic organizations possess the expertise and funding to create platforms that incorporate all aspects of SIA. As a result, platforms typically slowly emerge over time, based on previous work and collaborations, with varying levels of continued maintenance and support. We discuss here two of the main SIA platforms, *Greta* and the *Virtual Human Toolkit*, chosen because of their lengthy history, broad coverage of SIA areas, extendability, and ongoing use and support. Afterwards, we point toward several alternatives.

20.4.3.1 **Greta**

One of the earliest and most fleshed-out SIA platforms is *Greta*, originally released around 2005 [Poggi et al. 2005]. It grew out of earlier work that focused on the importance of gaze in coordinated verbal and nonverbal communication [Poggi et al. 2000], the design of a reflexive agent capable of showing and hiding emotions [De Carolis et al. 2001], and the creation of a detailed 3D face capable of showing nuanced facial expressions as well as detailed facial and skin deformation, including wrinkles [Pasquariello and Pelachaud 2002]. It is most notable for being able to modulate verbal and nonverbal behavior based on personality and other factors as well as including automated listening behaviors and backchanneling [Bevacqua et al. 2010].

The architecture of Greta is modular and follows the publish and subscribe whiteboard approach [Niewiadomski et al. 2009]. It uses either the Psyclone messaging system [Thórísson et al. 2005] or ActiveMQ [Snyder et al. 2011]. It is compliant with the SAIBA framework [Vilhjálmsson et al. 2007], developed in Java, and extendable with custom modules. Its main platform target is Windows, with versions available for mobile and web as well. See Figure 20.3 for an overview of the architecture.

Greta can interface with a range of external audio-visual sensing and speech recognition systems, including Watson [Morency et al. 2005], PureData,\(^9\) and SSI [Wagner et al. 2013]. These signals are processed by the Intent Planner, which outputs a communicative intent in FML-APML [Heylen et al. 2008a]. The Behavior Planner takes this as input and creates a behavior schedule in BML [Kopp et al. 2006]. Finally, the Behavior Realizer creates the actual movement of the agent, which can be done in the FAP-BAP Player using the MPEG-4 standard [Ostermann 2002], a robot (e.g., Aibo) [Niewiadomski et al. 2009], or a game engine, including Ogre3D

---

or Unity. Text-to-speech is provided by either MaryTTS\textsuperscript{10} or CereVoice [Aylett and Pidcock 2007].

Many research efforts have used Greta, including the SEMAINE project, which aimed to develop a Sensitive Artificial Listener [Schröder et al. 2011a]. The HUMAINE project focused on the emotional aspects of human to agent interactions [Petta et al. 2011], which led to the creation of the HUMAINE Database that contains clips of the use of emotion in everyday interactions [Douglas-Cowie et al. 2011]. The TARDIS project created a framework for developing agents who could offer social coaching within the context of job interviewing [Anderson et al. 2013]. Ask Alice is an example of the ARIA project (Artificial Retrieval of Information Agent) in which a user can interact with a virtual Alice from Wonderland [Valstar et al. 2016].

Greta is available at https://github.com/isir/greta under a mix of LGPL v3 and GPL v3 licensing. It provides documentation and tutorials as well as an overview of associated projects.

\textsuperscript{10} http://mary.dfki.de.
20.4.3.2 Virtual Human Toolkit

The Virtual Human Toolkit (VHToolkit) is a convergence of approaches and technologies researched and developed at the University of Southern California Institute for Creative Technologies (ICT), released in 2009 [Hartholt et al. 2013]. The MRE [Rickel et al. 2001] and SASO projects [Swartout et al. 2006] provided the overall architecture and nonverbal behavior technology, combined with natural language processing and rendering technologies from the SGT Star project [Artstein et al. 2008], which were integrated into a common platform as part of the Gunslinger project [Hartholt et al. 2009].

The VHToolkit follows the SAIBA framework [Vilhjálmsson et al. 2007] and has a modular architecture. Modules mainly communicate with each other through message passing using a custom protocol called VHMsg, developed on top of ActiveMQ [Snyder et al. 2011], see Figure 20.4, where regular arrows indicate messages and bolded arrows direct connections. Modules can be written in a range of languages, the most common of which are C#, Java, and C++. This allows relatively easy incorporation of new modules as long as they adhere to the VHMsg protocol. The VHToolkit mainly supports Windows, although a multiplatform version is in internal development [Hartholt et al. 2020].

The default speech recognition solution is PocketSphinx [Huggins-Daines et al. 2006], with options for Google ASR and native Windows 10. Audio-visual sensing is provided by MultiSense, which is built on top of SSI [Wagner et al. 2013]. MultiSense combines sensing producers and consumers to provide sensing and behavioral

---

Figure 20.4 The Virtual Human Toolkit architecture, adapted from its website.
information to the rest of the system using Perception Markup Language (PML) messages [Scherer et al. 2012]. Natural language processing is provided by the NPCEditor, a statistical text classifier that matches novel user input to the best character response output [Leuski and Traum 2011]. It can take custom Groovy scripts to provide dialogue management functions. The NPCEditor sends FML to the NonVerbal Behavior Generator (NVBG), which generates a BML schedule based on syntactic and semantic rules [Lee and Marsella 2006]. This BML schedule is sent to SmartBody, a procedural character animation and simulation platform and one of the most powerful BML realizers available [Shapiro 2011]. Text-to-speech defaults to Festival [Black et al. 1998], with options for CereVoice [Aylett and Pidcock 2007] and MS SAPI [Shi and Maier 1996]. Rendering is provided by Unity.

The VHToolkit has formed the core of many virtual human prototypes, both for research as well as deployment. ELITE and INOTS combine virtual humans with intelligent tutoring to offer a system that allows young officers in the US Army and Navy to learn and practice leadership and counseling skills [Hays et al. 2012]. The Museum Guides were a lifelized exhibition at the Boston Museum of Science, providing kids with information about the museum and STEM topics [Swartout et al. 2010]. PAL3 is a meta-tutor that will accompany a student or professional throughout their career, providing advice and recommendations along the way [Swartout et al. 2016]. New Dimensions in Testimony allows museum visitors to interact with a lifelized capture of a Holocaust survivor [Traum et al. 2015]. VITA allows young adults with autism to practice job interviews [Burke et al. 2018]. It has been expanded to also serve veterans transitioning back into society as well as juveniles [Hartholt et al. 2019c]. An AR prototype was developed for the Magic Leap AR headset [Hartholt et al. 2019b]. The Battle Buddy is a mobile passive sensing agent designed to collect multimodal data with passive sensors native to popular wearables (e.g., Apple Watch, Fitbit, and Garmin) as well as through user self-report. It delivers personalized and adaptive multimedia content via smartphone application specifically tailored to the user in the interdependent domains of physical, cognitive, and emotional health [Mozgai et al. 2020].

The VHToolkit is free for academic use and can be obtained at https://vhtoolkit.ict.usc.edu. This website includes documentation, tutorials, and a forum.

### 20.4.3.3 Honorable Mentions

Relational Agents are web-based 2D IVAs often used within the healthcare domain [Bickmore et al. 2009]. The framework offers a task planner and dialogue manager with associated ontology [Bickmore et al. 2011], a web BML realizer, and text-to-speech integrations. It uses Java and Flash and is available at https://relationalagents.com/demo/index.html.
WASABI aims to combine physical emotion dynamics with cognitive appraisal in order to simulate infant-like primary emotions as well as cognitively elaborated secondary emotions [Becker-Asano and Wachsmuth 2010]. The source code is freely available under the LGPL v3 license through https://www.becker-asano.de/index.php/research/wasabi.

The Virtual People Factory is a web authoring and runtime platform, often used to create virtual patients [Rossen and Lok 2012]. The website can be accessed at http://virtualpeoplefactory.com.

Visual SceneMaker is an authoring tool that allows non-experts to create interactive presentations [Gebhard et al. 2012]. It has been used in many projects, including TARDIS [Anderson et al. 2013] and is freely available at https://github.com/SceneMaker/VisualSceneMaker.

ADAPT focuses on full body animation, navigation, and object interaction by combining a range of different approaches and technologies, including SmartBody, into a single framework [Shoulson et al. 2013]. It can be obtained at https://github.com/ashoulson/ADAPT.

The Articulated Social Agents Platform (Asap) provides a collection of software modules for both IVAs and SRs [Kopp et al. 2014]. Asap is SAIBA compliant and includes Flipper for dialogue management [Ter Maat and Heylen 2011]. Its main language is Java and is available under the LGPL v3 license at https://github.com/ArticulatedSocialAgentsPlatform/Asap/wiki.

Agents United combines Greta and Asap into a single platform that mainly uses message passing through ActiveMQ, with a small number of custom modules using web services [Beinema et al. 2021]. An application can contain multiple agents, for both desktop and mobile, and has native support for longitudinal studies. It mainly uses Java and C#, and can be found at https://github.com/AgentsUnited, under the LGPL v3 license.

The Generalized Intelligent Framework for Tutoring (GIFT) is mainly focused on providing intelligent tutoring capabilities both on desktop and the web [Sottilare et al. 2017]. It provides some SIA capabilities through integration with a subset of the VHToolkit. It is open source and accessible at https://www.gifttutoring.org.

M-PATH focuses on empathetic conversations, using SmartBody [Shapiro 2011] in combination with a custom dialogue manager [Yalçın and DiPaola 2019]. It is hosted at https://github.com/onyalcin/M-PATH.

The Standard Patient Studio allows doctors and medical students to create and practice with their own standard patients [Talbot and Rizzo 2019]. It is an online authoring tool that starts with an interactive, healthy patient as a baseline to deviate from. It includes student feedback capabilities and is accessible at https://www.standardpatient.org.
Tools to Create Appearance and nonverbal Behavior

For a digital character to be rendered, it first needs to be modeled, which includes creating the mesh (i.e., the shape of the character) as well as the textures (i.e., the paint on that shape). The mesh needs to be rigged to a skeleton, which contains all the joints that can be animated (e.g., legs, fingers, jaw). The animated skeleton drives the deformation of the mesh based on skinning information. This joint-driven animation approach can be used for both the body and the face. In addition, the mesh can be deformed directly by creating a series of blend shapes—also called morph targets—which are detailed poses of a portion of the mesh (e.g., the mouth or cheeks) that can be blended together to create the desired effect (e.g., a smile, bulging biceps). When all these aspects are integrated, this source art can be exported to a game engine, typically in the FBX format.\(^\text{11}\)

In this section, we will discuss individual tools that cover any aspect of the appearance and nonverbal behavior of an IVA, including modeling and animating a character as well as generating and realizing nonverbal behavior.

20.5.1 Modeling

Modeling a character includes both creating the mesh (i.e., shape) and textures (i.e., paint) of the character. Traditional modeling software allows artists to create the character mesh as well as the UV-layouts that indicate how a 2D texture should be mapped to the 3D mesh. Maya\(^\text{12}\) and 3DMax\(^\text{13}\) by Autodesk are often used by professionals, while Blender is a free and open source alternative,\(^\text{14}\) and Houdini offers a free version.\(^\text{15}\) These programs typically contain the overall character rig, with all the necessary elements to be exported to the game engine. Textures can be created with general purpose tools like Adobe Photoshop,\(^\text{16}\) the open source alternative Gimp,\(^\text{17}\) or with specialized software, including Mari\(^\text{18}\) and Substance Painter.\(^\text{19}\) Specialized sculpting software, including Mudbox\(^\text{20}\) and ZBrush,\(^\text{21}\) allows artists

---

to sculpt and paint high-detail models that can be exported into Maya, 3DMax, or Blender. The mesh typically gets downsized to contain less detail, while many of the original details get maintained in the textures. This technique allows for the simulation of detail while maintaining performance. Upcoming advances in game engine technology may fully automate this process, retaining all details without the need for low-fidelity meshes with high-fidelity texture maps.

Scanning real people in order to digitize their appearance is becoming ever more feasible, either through commodity hardware [Shapiro et al. 2014, Achenbach et al. 2017, Chibane et al. 2020] or through specialized hardware like the USC ICT LightStage [Debevec et al. 2000]. This process consists of creating a series of photographs of the subject, either full body or just the face, to then stitch together using photogrammetry into a 3D mesh with textures. While commodity hardware typically results in a character that has the original lighting conditions baked in, high-end solutions like the LightStage can capture images under a myriad of lighting conditions, allowing them to re-light the resulting 3D character in any novel environment. More recent efforts focus on generating models from a single image, for example, Tex2Shape [Alldieck et al. 2019], available together with related tools at https://virtualhumans.mpi-inf.mpg.de/software.html.

Character art assets can also be created or obtained directly from 3rd party sources, including Autodesk Character Generator, iClone, MakeHuman, Mixamo, Renderpeople, TurboSquid, Unity Asset Store, Unreal Marketplace, and Unreal MetaHumans.

Ultimately, any of these assets will need to be run in real-time in a game engine, combining skeleton, mesh, textures, and blend shapes. In addition, modern characters require materials and shaders in order to provide realistic reflections of light on different surfaces (e.g., cloth, skin, eyes) as well as physics on cloth and hair. Tools like Substance Designer and PhysX can support these.

20.5.2 Animation

Animating IVAs brings with it a set of particular challenges compared to offline animation for movies or even real-time animation for video games. IVA characters need to be able to respond dynamically and in real-time to the user, which requires lip-synching to match the generated character speech, conversational gestures in support of the words and overall meaning, as well as for body language, facial expressions, and gaze to match the underlying communicative intent, personality, and cognitive processes. For use with nonverbal behavior generation and realization (see below), they will need to be annotated with metadata that indicates the timing of the phases of the animation so that it can be synchronized with the character’s speech.

Animation techniques can be divided into traditional keyframe animation, motion capture (mocap), and procedural animation.

With keyframe animation, an artist uses an animation rig to create a series of poses—the keyframes—that an animation system then blends together to create the final performance. This can be labor intensive but allows the artist complete control over the final result. This is typically done in tools like Maya, 3DMax, or Blender.

Mocap maps an actor’s movements onto a digital character. High-end mocap studios use dedicated sets with specialized cameras that look at markers on a suit that the actor wears. This typically gets processed with tools like MotionBuilder before it’s used in the rest of the animation pipeline. Markerless suits can be used outside of expensive studios, for example, Rokoko. More commodity hardware like webcams or 3D depth cameras can be used for a lower-cost (and typically lower quality) solution, for example, f-clone and iPi Soft.

Finally, animations can be procedurally generated, either using math functions or example-based controllers, where existing animation data forms the basis for further manipulation or blending to get the desired performance.

Regardless of the process, animations can be purchased from third-party sources, typically from the sources mentioned in the Modeling section above.

While the above approaches can broadly apply to both the body and the face, the latter typically requires special attention. Facial animation is complex, layering and blending lip-synching with facial expressions and gaze. Most systems base the

facial expressions on the Facial Action Coding System (FACS) framework [Ekman 1997]. Lip-sync tools like FaceFX\textsuperscript{37} or dedicated Unity or Unreal Engine plugins can generate a viseme schedule (i.e., individual base units of mouth shapes) based on the character’s phoneme schedule (i.e., individual base units of speech sounds), either by analyzing an existing audio file or by obtaining a phoneme schedule in real-time from a text-to-speech provider. Full facial performance capture uses cameras to track an actor’s face and translate the performance to a digital character, for example, using Blender\textsuperscript{38}, FaceWare\textsuperscript{39}, and Unreal Engine.\textsuperscript{40}

### 20.5.3 nonverbal Behavior Generation and Realization

Many traditional IVAs are developed using the SAIBA framework [Kopp et al. 2006], where an agent “mind” generates a communicative intent in the form of FML, which gets translated into BML by a generator, which in turn gets realized in an animation system and renderer. BML describes at a high level what a character should do (e.g., speak words, gaze at object, gesture) and how to synchronize all behaviors. For instance, it can describe that the emphasis point of a conversational gesture coincides with the pronunciation of a specific word. The BEAT system was instrumental in laying the groundwork for this approach [Cassell et al. 2004]. For more details on multimodal interaction architectures, see Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022] of this volume of this handbook.

There are several generators available that take FML as input and produce BML as output. The Greta platform (see Section 20.4.3.1) provides the Behavior Planner. It takes the high-level intent of the agent and generates through rules a series of associated nonverbal behavior intents, which can be based on role, personality, and context. This was first described in De Carolis et al. [2002] as the APML, one of the efforts on which FML is modeled. The VHToolkit includes the NVBG [Lee and Marsella 2006], a rule-based system that analyzes a character’s surface text semantically through keyword scanning and syntactically with the Charniak parser [McClosky et al. 2006]. A series of rules fire to generate head gaze, nods, and shakes, and to select conversational gestures and facial expressions. The resulting schedule gets pruned based on rule priorities when competing overlapping behaviors are triggered. Defaults can be overwritten per character or personality.

\textsuperscript{37} https://facefx.com.

\textsuperscript{38} https://blender.community/c/today/9sdbbc.

\textsuperscript{39} https://www.facewaretech.com.

\textsuperscript{40} https://docs.unrealengine.com/en-US/Engine/Animation/FacialRecordingiPhone/index.html.
BML realizers take the high-level behavior schedule and are responsible to execute this in real-time, synchronizing all requested behaviors. Greta includes the Behavior Realizer [Le et al. 2012]. It is written in Java and includes the ability to synchronize speech, gestures, gaze, and facial animations. It integrates with dynamic listening behaviors and allows users to tweak and create their own gestures with a custom tool, which can be modulated for intensity. SmartBody [Shapiro 2011] ships as part of the VHToolkit but is also a stand-alone character simulation platform available at https://smartbody.ict.usc.edu under the LGPL license. It is written in C++ and includes its own renderer and debugging tools, offers multi-platform support, and includes speech, locomotion, gazing, object manipulation, and physical simulation. AsapRealizer 2.0 is part of Asap, with a particular focus on incremental behavior plan construction, graceful interruption, and adaptation of ongoing behavior. More information can be found in Van Welbergen et al. [2014], which also includes a detailed comparison to other BML realizers. AsapRealizer is developed in Java and is available at https://github.com/ArticulatedSocialAgentsPlatform/AsapRealizer under the LGPL v3 license. LiteBody is part of the overall Relational Agents approach [Bickmore et al. 2009]. LiteBody is specifically developed for the web; however, it requires Flash, which is no longer supported. It is developed in Java and available through https://relationalagents.com/demo/litebody.

20.6 Tools to Model Interactions

In this section, we discuss tools that address human–SIA interaction. In particular: speech recognition, which turns user speech into text; audio-visual sensing, which perceives and analyzes the user’s face, body, and voice; natural language processing, which understands the user’s verbal input, generates the character’s verbal output, and manages the overall dialogue; and expressive speech, which generates character speech based on its communicative intent.

20.6.1 Speech Recognition

Speech recognition turns user speech into text the rest of the system can use. Systems process user audio either locally on the device or on a server in the cloud. Local systems are typically more secure, don’t require an Internet connection, and are more flexible, allowing, for instance, the definition of custom acoustic or language models. Acoustic models describe the acoustic qualities of a target audience (e.g., you, you in a specific recording environment, accents, children vs. adults). Language models describe the linguistic qualities of a target domain
(e.g., specialized vocabulary, common phrases). Cloud-based services have access
to more computing power and therefore can have a higher accuracy while typi-
cally costing money and requiring an Internet connection, which can add latency.
Either approach can provide sequential results (i.e., audio is processed once the
user stops talking) or continuous results (i.e., audio is continuously processed,
and text strings are intermittently sent). Most solutions provide a confidence score
for the recognition and may provide additional information beyond the text (e.g.,
prosody, emotion, filler word removal).

One of the original local solutions is the \textit{CMUSphinx} suite \cite{Lamere2003}.
It contains several tools for both Java and C. It supports Windows, MacOS, Linux,
and Android, and it allows you to create custom acoustic and language models. It
is available at \url{https://cmusphinx.github.io}. While development on the core Sphinx
suite has slowed down, it provides links to related efforts. \textit{Kaldi} is a research-
focused, local speech recognition toolkit for Windows, MacOS, Unix/Linux, and
Android \cite{Povey2011}. It is written in C++ and has an active development com-

Cloud solutions are typically provided by large technology companies, driven by
voice and personal assistant applications. This has resulted in big improvements in
speed and accuracy, which can be leveraged for SIA. The main ones include Google
Speech-to-Text\textsuperscript{41}, Amazon Transcribe\textsuperscript{42}, Microsoft Azure Speech to Text\textsuperscript{43}, and IBM
Speech to Text\textsuperscript{44}. These services are able to leverage large quantities of data and
computing power in order to provide solid accuracy and a single solution for mul-
tiple hardware platforms, at the cost of customization, possible data collection,
and service fees.

Most hardware and Operating System (OS) platforms also offer APIs for native
speech recognition solutions, for example, Microsoft Windows\textsuperscript{45}, Apple MacOS
and iOS\textsuperscript{46}, and Android Speech\textsuperscript{47}.

\textsuperscript{41} \url{https://cloud.google.com/speech-to-text}.
\textsuperscript{42} \url{https://aws.amazon.com/transcribe}.
\textsuperscript{43} \url{https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text}.
\textsuperscript{44} \url{https://www.ibm.com/topics/speech-recognition}.
\textsuperscript{45} \url{https://docs.microsoft.com/en-us/windows/apps/speech}.
\textsuperscript{46} \url{https://developer.apple.com/documentation/speech}.
\textsuperscript{47} \url{https://developer.android.com/reference/android/speech/package-summary}. 
20.6.2 Audio-visual Sensing

Audio-visual sensing uses cameras and microphones to perceive a user’s face, body, or voice in order to recognize facial features, gestures, voice acoustics, and so on. This can be used for a variety of purposes, for example, to recognize people, an increasingly controversial use. Within the context of SIA, audio-visual sensing is typically used to detect the affect of the user, in particular in support of real-time conversational goals, including rapport building and improved understanding through nonverbal behavior. As with speech recognition, tools are either local or cloud-based.

Local solutions include OpenCV (Open Source Computer Vision Library), an open source computer vision and machine learning software library that aims to provide a common infrastructure for a range of computer vision-related applications, including detecting faces and classifying human actions. It is written in C++ and supports Windows, Linux, Mac OS, and Android. It is freely available at https://github.com/opencv/opencv under the BSD license.

Social Signal Interpretation (SSI) is a framework for real-time recognition of social signals, including tools to record, analyze, and recognize human behavior in real-time, such as gestures, mimics, head nods, and emotional speech [Wagner et al.]. SSI allows the integration of multiple data producers and consumers for both audio and video. It is written in C++ and available under the GPL v3 and LGPL v3 licenses at https://hcm-lab.de/projects/ssi. Closely related to SSI is the NOnVer-bal behaviour Analysis tool (NovA) [Baur et al.], which supports the analysis and interpretation of social signals conveyed by gestures, facial expressions, and others as a basis for computer-enhanced social coaching. See https://github.com/hcmlab/nova for more details.

MultiSense [Stratou and Morency] is part of the VHToolkit (see Section 20.4.3.2). MultiSense combines multiple existing and custom data producers and consumers into a single framework, based on SSI (see below). It primarily focuses on the face using commodity web cams but can analyze the full body with the original Microsoft Kinect and has the possibility to add custom voice analytics modules. Results are communicated through message passing using the PML. See Scherer et al. [2012] for more details.

OpenSMILE, despite its name, is focused mainly on acoustic analysis of voice and music. SMILE stands for Speech and Music Interpretation by Large-space Extraction and can be used in both real-time as well as offline feature extraction on large datasets. Within the context of SIA, it is most useful for voice activation detection and speech emotion recognition. OpenSMILE works on Windows, Linux, and Mac OS. Its source code is available under a custom license at https://www.audeering.com/opensmile.
OpenFace 2.0 performs real-time facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation using a webcam [Baltrusaitis et al. 2018]. The underlying models can be re-trained and the source code is freely available for research purposes at https://github.com/TadasBaltrusaitis/OpenFace.

OpenPose offers 2D pose detection, including body, foot, hand, and facial landmarks [Cao et al. 2018]. It can detect these in real-time for multiple people in an image or video stream. OpenPose comes with a C++ and Python API. The source code is available for non-commercial use at https://github.com/CMU-Perceptual-Computing-Lab/openpose.

Microsoft Platform for Situated Intelligence (psi) is an open source framework for multimodal intelligent systems [Bohus et al. 2017]. It consists of a runtime for real-time data collection and manipulation, a suite of tools for analytics, visualization, and training, and a collection of components that can be combined to create applications. It uses C# with interfaces to other languages (e.g., Python, JavaScript). It integrates with Azure Cognitive Services as well as the Azure Kinect DK (https://azure.microsoft.com/en-us/services/kinect-dk). psi is available at https://github.com/microsoft/psi under the MIT license.

OpenSense follows MultiSense and provides a framework in which producers and consumers can be combined in a flexible way [Stefanov et al. 2020]. It builds on psi (see below) and is written mainly in C# and C++. It is available at https://github.com/intelligent-human-perception-laboratory.

The big US tech companies offer web services for audio-visual sensing as well. The main focus is visual, including face detection, facial landmark detection, emotion detection, and object recognition. These services can be accessed directly through REST calls or using dedicated SDKs for the most common languages. Services include Amazon Rekognition, Google Vision, and Microsoft Azure Computer Vision.

20.6.3 Natural Language Processing

Natural language processing (NLP) can be divided into three areas:

— Natural language understanding: comprehend what the user is saying.
— Natural language generation: generate what the agent should say.
— Dialogue management: manage the conversation between two or more entities.

Tools can cover a mix of these three areas. One important element is who has the initiative in the conversation: the user (e.g., a question-answering system, personal assistant), the agent (e.g., virtual interviewer), or both. The latter, mixed-initiative systems can handle more complex conversations and are typically more difficult to develop. For a more in-depth discussion of natural language approaches, see Chapter 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021] of volume 1 of this handbook [Lugrin et al. 2021]. Here, we will primarily discuss common tools that can be used in real-time systems. Given the focus on digital and voice assistants, this is a very active area of research and development, performed at both academia and within industry. This results in many available tools ranging from individual libraries for natural language research to fully developed solutions and everything in between, often made open source, regardless of the origin.

**Olympus**, from Carnegie Melon University (CMU), is one of the earliest available NLP tools. It is a suite of tools that cover speech recognition, natural language understanding and generation, and dialogue management [Bohus et al. 2007]. **RavenClaw** is the dialogue manager and can handle mixed-initiative conversations that match external input to an internal agenda linked to a task model [Bohus and Rudnicky 2009]. The tools in Olympus are available under the BSD license at http://wiki.speech.cs.cmu.edu/olympus.

The **NPCEditor** is part of the VHToolkit (see Section 20.4.3.2). It is a statistical text classifier that matches novel user input to the best pre-authored character response output using an information retrieval approach [Leuski and Traum 2011]. Authors provide the NPCEditor with examples of how user input should be matched to character output. Novel user input is analyzed against close known inputs and their linked outputs, resulting in a set of possible answers above a certain threshold. This set is processed by its dialogue manager, a Groovy script that can be customized. For example, it can avoid repeating the same utterance or prompt the user if no suitable response can be retrieved.

**OpenDial** is a toolkit with a focus on dialogue management, with the possibility to add natural language understanding and generation, text-to-speech, and multimodal processing [Lison and Kennington 2016]. It is written in Java and provides a hybrid approach that combines human readable rules with a Bayesian network that contains the dialogue state. It is available through http://www.opendial-toolkit.net under the MIT license and seems to no longer be in active development.

**PyDial**, the Cambridge University Python Multi-domain Statistical Dialogue System Toolkit, is a more research-focused toolkit [Ultes et al. 2017]. It offers natural
language understanding and generation as well as dialogue management, provided by both rule-based and model-based approaches. PyDial uses Python and is available at http://www.camdial.org/pydial under the Apache 2.0 license.

*Rasa* is a commercial company that offers an open source solution for conversational AI, including dialogue management and natural understanding [Bocklisch et al. 2017]. More advanced services, including annotations, multiple deployed versions, and support require a subscription. Rasa is written in Python and the open source portion is available under the Apache 2.0 license through https://rasa.com.

*DeepPavlov* is an open source library for creating natural language solutions, combining machine learning and deep learning models with traditional rule-based approaches. These form the basis for individual skills (e.g., question-answering, goal-oriented dialogue) that can be combined into a single agent, which can be integrated with existing systems of services. DeepPavlov uses TensorFlow and Keras, supports Windows and Linux, and mainly uses Python. It is available at https://deeppavlov.ai under the Apache 2.0 license.

*ChatScript* is a rule-based scripting language that forms the foundation for many custom natural language systems (see https://github.com/ChatScript/ChatScript). It does this through describing patterns of user input, combined with an ontology and built-in memory of conversations. ChatScript works on Windows, Linux, MacOS, iOS, and Android and has a server version. It is available under the MIT license.

The main big US tech companies offer a suite of natural language processing services, including text understanding, semantic analysis, and conversational interactions, for example, Google Dialogflow, Amazon Lex, and Microsoft LUIS. These are typically focused on personal assistant type interactions that drive Amazon Echo, Google Assistant, and Microsoft Cortana. This means conversations are authored around user intents (e.g., play music, book a vacation), with parameters to be filled in for the specifics of each request. Online authoring tools are aimed at domain experts rather than natural language researchers. The ability to connect to a service from any device offers flexibility, at the cost of requiring an online connection, providing user data, and paying for used data and compute cycles. These and other large companies increasingly open source parts

---

51. https://keras.io.
of their technology stack, including Pluto from Uber,\textsuperscript{55} ParlAI from Facebook,\textsuperscript{56} and Google BERT\textsuperscript{57} and ALBERT.\textsuperscript{58} They join forces with more traditional academic approaches, including Stanford's suite of software.\textsuperscript{59}

### 20.6.4 Expressive Speech

Expressive speech, or text-to-speech generation, turns character text into speech (see Chapter 6 on "Building and Designing Expressive Speech Synthesis" [Aylett et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]). Most tools allow at a minimum the definition of a particular voice and optionally allow annotations of the speech with SSML [Taylor and Isard 1997] to, for instance, indicate emphasis, prosody, or emotion. In order to utilize these tools in real-time, in addition to the resulting audio file, they need to provide a phoneme schedule in order to drive lip-synching (see Section 20.5.2).

One of the earliest available tools is Festival [Taylor et al. 1998]. This is an offline available tool that offers various male and female voices, using a range of approaches. While some of these approaches are outdated, it provides an out-of-the-box solution that is relatively easy to integrate using C++. It supports English and Spanish and is available at http://www.cstr.ed.ac.uk/projects/festival.

MaryTTS supports roughly ten languages with a host of options for metadata, including phoneme and intonation schedules using a custom XML schema [Schröder et al. 2011b]. It is written in Java and can be used locally or set up as a service. MaryTTS is available at http://mary.dfki.de under the LGPL v3 license.

There are several commercial options available. CereVoice\textsuperscript{60} was one of the first viable companies in this space and offers both local and online solutions, with either a proprietary SDK or web services [Aylett and Pidcock 2007]. Just as with other personal and voice assistant-related technologies, the big tech companies offer online services, including from Amazon,\textsuperscript{61} Google,\textsuperscript{62} and Microsoft.\textsuperscript{63} Many of these offer the functionality to clone a voice as well, where a relatively small amount of voice data from a specific person can be used to generate novel speech by that

---

\textsuperscript{55} https://github.com/uber-research/plato-research-dialogue-system.
\textsuperscript{56} https://github.com/facebookresearch/ParlAI.
\textsuperscript{57} https://github.com/google-research/bert.
\textsuperscript{58} https://github.com/google-research/ALBERT.
\textsuperscript{59} https://nlp.stanford.edu/software.
\textsuperscript{60} https://www.cereproc.com.
\textsuperscript{61} https://aws.amazon.com/polly.
\textsuperscript{62} https://cloud.google.com/text-to-speech.
\textsuperscript{63} https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech.
same person [Arik et al. 2018]. This allows pre-recorded speech to be matched with generated speech. Finally, much work is currently put into making generated voices sound more natural, including introducing disfluencies [Oord et al. 2016].

20.7 Similarities and Differences in IVAs and SRs

Both IVAs and SRs aim to perceive a human, process that input, and integrate it with its internal state in order to respond appropriately both verbally and nonverbally. However, while IVAs inhabit a virtual world, SRs are fully embedded within the real world, which brings with it many additional challenges and constraints, including perception (e.g., mapping the real, dynamic world to an internal representation), physics (e.g., using mechanical elements to realize conversational gestures), navigation (combining perception and physics), and energy (balancing the power of an energy source with volume, weight, and mobility). One school of thought views this physical embodiment and being part of the real world as a fundamental necessity of creating SIAs. Per Brooks [1991], human evolution took a long time to create low-level systems to interact with the world, on top of which intelligence evolved relatively quickly. In addition, the complexity of a system itself may be determined by the complexity of the environment it operates in. Finally, by abstracting the real world, researchers and developers necessarily do the heavy intellectual lifting in these abstractions rather than in the systems themselves and may create simplifications and dependencies that cannot be overcome when these systems are deployed in the real world.

Many of the platforms and tools discussed here can be and have been applied to robotics. For instance, Greta has been integrated with an Aibo robot [Niewiadomski et al. 2009] and a Nao [Le and Pelachaud 2011] robot. The VHToolkit has also been used with a Nao robot [Artstein et al. 2016] as well as with mobile robots that use a traditional monitor screen to display a virtual head on [Si and McDaniel 2016, Pang et al. 2018], while SmartBody is used in the Sophia robot by Hanson.57

Mixed approaches like these are more commonplace, aiming to combine advantages from IVAs with SRs. Furhat takes a novel approach in back projecting a virtual

64. For example, the complex route an ant takes to get home is largely a reflection of the obstacles and entities it encounters in its environment rather than of its quite simple set of behaviors [Simon 1969].
face on a physical mold in order to address the limitations of a 2D screen, in particular in regard to gaze behavior [Al Moubayed et al. 2012]. It allows for multiparty interactions in a 3D physical space while leveraging smooth facial expressions.

Furhat is driven by IrisTK, which is a modular, Java-based system with the specific aim of supporting SR research and development [Skantze and Al Moubayed 2012]. IrisTK is available at http://www.iristk.net under the GPL v3 license.

Another approach is to leverage the virtual space to simulate robots before physically creating them in order to speed up research and development [Coevoet et al. 2017]. For instance, work with the VHToolkit has shown that new algorithms can be rapidly iterated upon in a simulation, where physical and time constraints are reduced, before implementing promising candidates in the real world. This includes experimentation with a larger number of robots than may be feasible in the real world, more advanced sensors than are currently available, or multiple labs in different locations [Hönig et al. 2015].

Similarly, shared frameworks allow for the exploration of IVA and SR communication [Rahman 2019]. ScoutBot is developed using the VHToolkit and is integrated with the Robot Operating System (ROS) [Lukin et al. 2018]. ROS is an open collection of tools, libraries, and conventions that support a wide range of robot research and development [Quigley et al. 2009]. Dedicated migration platforms allow for an agent to migrate between different embodiments (e.g., from an IVA to an SR) [Kriegel et al. 2011, Hassani and Lee 2014].

While IVA research typically focuses on simulating human appearance and behavior, this may not be desirable for SRs [Gratch et al. 2015]. Abstracting away from realistic humans to animals or stylized humanoid robots allows for focusing on the social and interactive elements without the challenge of re-creating lifelike, physical human representation. Examples include Aibo and Nao mentioned above, the Jibo robot,69 and the Nabaztag robot,70 the last of which has been supported by Asap Realizer [Reidsma and van Welbergen 2013]. Recently, Embodied revealed their robot, Moxie, a lifelike robot-companion for children to provide support in the development of social and emotional skills, as described in their recent study [Hurst et al. 2020].

Regardless of the type of robot, most benefit from the tools discussed in this chapter, in particular audio-visual sensing, speech recognition, natural language processing, and expressive speech. More humanoid-like characters also benefit from nonverbal behavior generation [Mead et al. 2010, Mlakar et al. 2013].
Matsuyama 2015). With ever more advanced capabilities and tools, increased fidelity in graphics, and improved hardware, we’re bound to see the cross-pollination between IVAs and SRs that has started at the beginning of the field continue to increase in the coming years.

Current Challenges

While much progress has been made in the past 20 to 30 years in regard to platforms and tools, as well as their effectiveness and dissemination, many challenges remain.

In individual specializations, these are often directly tied to the challenges of the field itself. Modeling the human mind, for example, or accurately inferring a person’s mental state based on audio-visual input are by no means solved problems. As a result, theories, approaches, and technologies have not yet matured to a point where they can be elegantly captured in user-friendly solutions for further use in research or development.

However, much progress has been made in individual fields driven by the rise of machine learning techniques. This approach does bring its own set of challenges. Effective machine learning requires huge amounts of data, which is difficult for smaller teams to acquire. Even for larger teams, properly curating datasets is not only labor-intensive but it is rife with challenges regarding biases, which is only exacerbated by the inherent black-box nature of most machine learning techniques. Effective tools will require validated datasets combined with explainable AI features in order to create systems that can be trusted to give the intended results.

As a whole, the combined trends of increased specialization and expanded democratization of many SIA relevant technologies, partly driven by the push of digital personal assistants, has led to an explosion of available tools. This increases the potential of being able to leverage existing capabilities while also increasing its complexity. Furthermore, relatively little progress has been made to standardize interfaces in such a way that individual requirements can be met by interchangeable solutions. This is due to many factors, including diverse sets of requirements (e.g., voice-based call center interactions vs. embodied story-driven agents), divergent incentives (e.g., research vs. commercialization), the availability of multiple hardware platforms and form factors (e.g., web, mobile, desktop, AR, VR), and the fact that formalizing principled representations on how exactly the human mind and body operate is just, well, hard. However, the lack of formal standards does not prevent software as a whole to become ever more modular, allowing distributed systems to be developed more organically from available microservices. This allows researchers and developers to pick and choose from an ever-expanding suite of relevant services at the cost of interfacing with them individually.
This points to one of the most enduring remaining challenges: providing an integrated solution that provides the tools to create and validate both appearance and behaviors in all their nuanced interplays. This requires a deep level of understanding of not only the individual research fields but also how they all interconnect; a level we have not yet reached, neither within the social sciences nor the “hard” sciences.

Even while we continue to gain in our level of understanding, creating solid tools that are easy to use is a challenge in and of itself. Developing these tools require (1) a deep understanding of all underlying research fields that the tool aims to capture, (2) decisions regarding the tradeoff between power and complexity as well as the possible abstraction levels for each, and (3) a solid understanding of the end users, their skill level, their likely knowledge of the domain, and how they can and want to leverage the tool in a user-friendly manner. As with the research that underlies all aspects of SIA, this requires an interdisciplinary team of researchers and developers, with the additional challenge of translating gained knowledge, capabilities, and procedures to a user-friendly package that the end user can take advantage of.

20.9 Future Directions

As our understanding of a particular SIA field advances and the maturity of related technologies increases, the tools that support these grow increasingly more powerful and user-friendly. The current interest in SIA in general and the associated commercial applications in particular will lead to a continuous democratization of tools that support SIA exploration and creation. As before, this starts with relatively isolated aspects of human behavior (e.g., speech, hearing) and continues with more complex behavior (e.g., multiparty dialogue, longitudinal relationships). As these tools become more advanced and user-friendly, it lowers the barrier of entry to the inherently integrated nature of SIA research and development.

In terms of platforms and tools, the authors aim to focus on the following aspects:

1. **Microservices architecture.** As per [Hartholt et al. 2020], the aim is to provide a modern, modular architecture that leverages cloud services in order to offer both researchers and developers a powerful and flexible framework to collaboratively explore SIAs. The modularity allows for multiple implementations for a given service, allowing for tradeoffs between power, flexibility, and performance.

2. **Multiplatform support.** The aim is to enhance the VHToolkit to not only support desktop applications but also web, mobile, AR and VR in order to better
explore the strengths of each platform [Hartholt et al. 2019a]. This leverages the microservices architecture and provides per-platform capabilities and best practices.

3. **Audio-visual sensing.** The more real-time information can be gathered from the user, the better SIAs are positioned to converse with end users in a manner that increases rapport and avoids frustration. This requires more research to go from external feature extraction (e.g., smile, frown) to internal inference (e.g., happy, confused) for all members of the human population. The aim is to integrate a range of cloud services and local solutions (e.g., Stefanov et al. [2020]) to provide a testbed for integrated exploration.

4. **Character generation.** As exploring our own humanity is a key pillar of SIA research, we collectively should strive for IVAs to match our diversity in order to (1) represent populations from all over the world, (2) support social research related to race, ethnicity, gender, sexuality, age, and so on, and (3) avoid unnatural repetition. It is therefore vital that we lower the required effort to create diverse, high-quality IVAs. We aim to pursue this by leveraging recent progress in generating high-fidelity characters [Li et al. 2020].

5. **Integrated authoring tools.** Humans are complex beings with advanced capabilities that have no clear boundaries; it is the system as a whole that leads to complex behaviors and overall intelligence. Similarly, SIAs require tightly integrated capabilities that collectively realize the goals for which they have been created. This requires authoring tools that take the integrated nature of SIAs into account and provide ways to create and validate agents that encompasses the whole rather than solely focus on individual areas.

### 20.10 Summary

One of the defining aspects of humanity is the use of tools to increase our productivity and enhance our understanding of ourselves and the world we live in. As our tools have become more sophisticated, so has our ability to create and understand. There is no better field than SIA to exemplify this duality of creation and understanding, from exploring ourselves mentally, physiologically, and socially to developing theories and implementations of virtual counterparts. It is tools that underlie this intertwined process and allow us to progress.

The sophistication of SIAs have advanced considerably in recent years, and with it, more and more tools have been made available to support the research and development of both IVAs and SRs. As a result, the barriers to enter this important field have been lowered to the point where it is easier than ever for individuals
and small groups to advance our understanding of what makes us human and to leverage that knowledge in creating applications that benefit humanity.

Big challenges remain, given the complexity of the subject matter. Progress in individual areas will advance our understanding and lead to ever more powerful tools. These will have to come together in order to provide a holistic approach to researching and developing SIAs. This requires an interdisciplinary approach, where researchers, developers, artists, and usability experts work together to understand, create, and refine the tools that enable not only the creation of powerful new systems but the understanding of the human body and mind itself.

References


References


Chapter 20 Platforms and Tools for SIA Research and Development


References


PART V

AREAS OF APPLICATION

This handbook consists of 28 chapters, covering the major topics in Socially Interactive Agents (SIAs). Each chapter provides a survey that summarizes the theoretical background, approaches for implementation, history/overview of the topic, alongside with current challenges and future directions. All chapters discuss similarities and differences between Intelligent Virtual Agents (IVAs) and Social Robots (SRs) and highlight important work of both fields in the topic.

The chapters are clustered into five parts, representing broad themes in SIA research. Volume 1 [Lugrin et al. 2021] contains Parts I–III. Part I “Establishing SIA Research” helps the reader understand how research in this area is conducted and discusses the impact on individuals and society. Part II “Appearance and Behaviour” deals with the most immediately noticeable external features of SIAs across multiple modalities. Part III “Social Cognition—Models and Phenomena” investigates internal processes known from human cognition that are driving forces in human–human interaction and demonstrates how they are addressed in SIA systems.

This Volume 2 contains Parts IV “Modeling Interactivity” and V “Areas of Application”.

Part V gives an overview of the major domains in which SIAs are employed, directing the reader to systems and research findings and highlighting the benefits of SIAs to individuals and society. It is important to note that this part is not about summarizing existing applications, but rather gives detailed surveys of some of the main concrete areas of applications that require different underlying theories, methods, and implementations. For example, it will make a huge difference
for a SIA researcher whether (s)he is working in the domain of child education, aging support, autism, or games, as the target user group, types of interactions, and concrete goals of systems in these different domains are very diverse.

The chapters in Part V build upon the concepts presented in Parts I–III (Volume 1 of this handbook) by applying the outlined research methods, and considering individual and ethical implications (Part I), and using knowledge on designing appearance, modeling multimodal behaviour including verbal, para-verbal and nonverbal behavior (Part II), as well as relying on underlying cognitive phenomena such as theory of mind, empathy, rapport, or culture (Part III). It also relies on Part IV (this volume) by using interactivity with the user and other agents in applied settings. Thus, some concepts, topics, and concrete applications that appeared in earlier chapters re-emerge in the chapters in this part but from different points of view.
21.1 Background and Motivation

A pedagogical agent (PA) can be defined as a virtual character or physical robot that seeks to promote learning, enhance motivation, and provide support to engage in an educational activity. PAs can be non-interactive or interactive, and typically seek to emulate naturalistic communication with learners through speech, gesture, emotions, and action. In this chapter, we outline the design considerations, summarize the state of empirical research, and suggest several important directions for future research for the PA field. While our emphasis is on research conducted with virtual agents (i.e., characters on a display), our position is that many of the fundamental design aspects of PAs are applicable whether the PA is physical or virtual. Similarly, from a learning sciences point of view, a known effective pedagogical strategy should be effective regardless of the medium. That said, we certainly recognize the fundamental practical differences between intelligent virtual agents (IVA) and social robots (SR), and the potential for one to have advantages over another in given contexts. Later in the chapter, we identify some recent research to explore these important differences. Our hope is that with growing interest in educational robots, the now extensive body of research on virtual PAs can be leveraged and applied so that mistakes are not replicated and that, together, the two fields can advance with shared goals and common methodologies.

There are many reasons to think that PAs can be beneficial for learning. Given that much learning occurs with other people, such as in classrooms [Kumpulainen and Wray 2003] and museums [Leinhardt et al. 2003], it is intuitive to consider learning as a social experience involving questions, dialogue, smiling, pointing, and much more. Introducing PAs may therefore activate a more natural, social frame for learners as they gain new skills or knowledge. From a more technological perspective, in an early review of the field Johnson et al. [2000] explained that “agents can demonstrate complex tasks, employ locomotion and gesture to focus
students’ attention on the most salient aspect of the task at hand, and convey emotional responses to the tutorial situation” (p. 47). In other words, PAs greatly expand available communicative bandwidth beyond a text-only interaction, and they introduce the potential to virtually inhabit a learning space with a learner. This suggests a significantly wider range of potential collaborative and shared learning interactions as well as the compelling possibility of modeling learning that occurs in the real world to a higher degree of fidelity.

However, the decision to include a PA in a computer-supported learning environment is a non-trivial one that involves both theoretical and technical considerations. Including a PA in a virtual environment could potentially increase the learner’s cognitive load and possibly hinder learning [Clark and Choi 2007, Sweller et al. 2011]. Also, from a software engineering point of view, implementing a sufficiently robust, interactive, and appealing agent can be cost and time prohibitive. To achieve a meaningful level of interactivity with a PA, significant effort is often required to achieve speech (or text) input/output, animations of nonverbal behaviors, appropriate and recognizable gestures and movement, and more. Each of these aspects of PA design can require associated expertise on behalf of the designer, who must understand both the theoretical and technical implications of different design choices. Given the complexity of PAs, it is important to understand the history of the technology and potential benefits they can provide learners.

### 21.2 History of Pedagogical Agents

From the earliest examples of mechanical teaching machines [Pressey 1926] to the highly interactive, immersive, social, and digital environments of today, the development of educational technologies has brought profound changes in the ways people learn with technology. While some of these changes represent the application of more general technologies to educational problems (e.g., data mining, virtual reality), others have emerged from novel properties of learning itself. For example, one of the best forms of education known to humankind is to learn from an expert human tutor in a face-to-face interaction [Bloom 1984]. Using what we know about effective human learning to influence the design of educational technologies captures a large portion of modern educational technology research.

Thus, it is important to understand the strategies and techniques of effective teachers and tutors so that those same productive interactions can occur in interactions with educational technology. Because an effective human tutor can focus completely on a single learner, they are able to more precisely fine-tune help that is given to that learner’s specific emerging needs. They can constantly assess that learner’s progress, offer personalized guidance and encouragement when that learner needs it, and optimize the time available for practice. When freed from
the demand of simultaneously meeting the needs of all members in a group of learners, such as in a classroom, these kinds of specialized interactions become possible. Thus, over time, computer-based learning environments have sought to emulate many of the same tactics and provide similar interactive learning experiences that human tutors are able to provide [Merrill et al. 1992]. In particular, tutoring is fundamentally a social experience and so the trajectory of research on computer-based tutors has been to increase their social capabilities with the hope that their effectiveness would simultaneously increase. In other words, a clear trend in the evolution of research on computer-based tutoring environments has been one of making them more like humans—that is, more communicative, social, and emotionally aware.

The earliest intelligent tutoring systems (ITSs) tended to play the role of “homework helper” by focusing exclusively on problem-solving support [Anderson et al. 1995, Shute and Psotka 1996]. Limited by the technology at the time, these early ITSs communicated by displaying text on a screen and tended to focus exclusively on cognitive aspects of learning. For example, a typical ITS would flag an action as wrong by coloring it red, or after a student requested a hint, say “You should try to combine like terms.” By providing just-in-time help and attending immediately to the needs of the student, these and similar later systems achieved impressive learning gains of about 1 sigma, roughly a letter grade [Kulik and Fletcher 2016].

Researchers were quick to realize that much more interactivity was possible, allowing these systems to pursue more human-like interactions. Inspired by the rich conversational tactics used by human tutors [Chi et al. 2001], a second generation of systems sought to leverage Natural Language Processing (NLP) techniques to allow learners to interact more naturally using their own words to explain and ask questions. Some systems emphasized analysis of student-generated explanations to tutor questions and rely exclusively on natural language interaction, whereas others use dialogue as a tactic during problem-solving to support remediation of misconceptions and to better assess learner beliefs [Graesser et al. 2001].

While PA research has existed for more than 20 years, implementing a PA within an ITS can be technically challenging. However, recent advances in animation and sound has enabled modern ITSs to be embodied—that is, to have a virtual body [Johnson and Lester 2016]. PAs typically help students learn by providing feedback that is explanatory as well as corrective [Moreno 2005] and personalized [Kim and Baylor 2016]. By using nonverbal communication in the set of pedagogical strategies (e.g., nodding, eyebrow raising) the PA can utilize, the bandwidth of communication between the PA and the learner is greatly expanded. For example, when explaining concepts to the student, agents can leverage relevant supporting
gestures known to facilitate thinking and learning [Goldin-Meadow 2003, Davis 2018]. In digital environments such as virtual worlds, an embodied agent can even inhabit the same learning space, provide orientation actions (such as pointing), and demonstrate activities in the environment for the learner, enabling learning through observation.

Body movement, gestures, and facial expressions also expand the communicative bandwidth for PAs to convey emotional messages, such as empathy and excitement. This enables implementation of more socially adept agents, which has been shown to benefit student learning in many cases [Atkinson et al. 2005], increase self-efficacy [Lane et al. 2013] and, more broadly, enable systems to better motivate and engage learners [Krämer and Bente 2010, Schroeder and Adesope 2014, Lane 2016]. A growing body of evidence supports claims that when designed properly, PAs enhance students’ motivational and learning outcomes [Schroeder et al. 2013, Kim and Baylor 2016]. Many of these effects are believed to be due to overt attempts to convey empathy and build rapport with human partners and engage in social conversation, which distinguish PA-based learning environments from many (non-PA enhanced) computer-based learning environments.

### 21.3 Designing and Implementing Pedagogical Agents

Incorporating a PA into a computer-based learning environment entails a substantial amount of additional design and implementation work. Designing effective PAs requires an in-depth understanding of the context they are being implemented in, such as the learners who will be interacting with the PA, the information being taught, and the interaction between the learners’ prior knowledge and the learning content. As such, not only is a strong understanding of learning theory necessary to know when and how to implement a PA, but a designer or design team must also have the technical skills to implement an effective PA. More specifically, in addition to deciding on the external properties such as the appearance of the agent as well as its role in the learning environment, defining internal models and properties (such as expert knowledge, learner model, and pedagogical polices) are also essential [Dehn and van Mulken 2000]. We provide an overview of these topics below but refer readers to Heidig and Clarebout [2011] for more detailed frameworks.

#### 21.3.1 External Properties—Roles and Appearance

A consistent argument given for the use of PAs is that they are believed to create a more social learning experience for a learner. While much of the success behind creating a social experience depends on the underlying technologies (e.g., NLP, animation quality), the basic design of the agent can have a profound impact on the experience. Two of the most fundamental design choices that must be made
Table 21.1 Examples of important elements of the design space of pedagogical agents

<table>
<thead>
<tr>
<th>Role</th>
<th>Appearance</th>
<th>Language/speech</th>
<th>Backstory/personality</th>
</tr>
</thead>
<tbody>
<tr>
<td>expert (teacher, tutor, coach)</td>
<td>fidelity (cartoon, photoreal)</td>
<td>voice type (machine-generated or recorded human speech)</td>
<td>place of origin</td>
</tr>
<tr>
<td>learning companion (peer)</td>
<td>shape, size, species, race/ethnicity (humanoid, animal, “living” object)</td>
<td>voice qualities (pitch, prosody, rate, disfluencies)</td>
<td>likes/dislikes</td>
</tr>
<tr>
<td>simulated role player (e.g., virtual patient for doctor training)</td>
<td>gender, age, size, &amp; shape</td>
<td>word choice (formal, informal, colloquial/culturally aligned)</td>
<td>sense of humor</td>
</tr>
<tr>
<td></td>
<td>hair, eye, skin color; clothing &amp; accessories</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nonverbal behaviors (gestures, body language)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

are (1) the role that PA will play in the learning experience and (2) what the agent will look like, how it will speak, and the agent’s “personality”. Table 21.1 provides examples of some of these important design choices.

A systematic review revealed that a large percentage of PAs reported in the literature play the role of teacher or coach [Schroeder and Gotch 2015]. In this case, the PA takes on the traditional role of guiding the learner through learning tasks such as solving problems or gaining conceptual understanding of some topic. PAs in this category are often responsible for structuring a learning experience by selecting problems to solve or topics to discuss, providing hints and feedback, and generally managing the learner’s experience. The next most popular role for PAs falls under the pedagogical approach of reciprocal teaching [Palincsar and Brown 1984] by positioning the student as a simulated learner with the human learner acting as teacher. Such teachable agents implement the learning-by-teaching paradigm and have been explored in a variety of contexts [Matsuda et al. 2013, Biswas et al. 2016]. Sometimes a teachable agent can take the role of a peer collaborator as well, meaning that the human learner may both learn from and teach during simulated collaborative interactions, see also Chapter 22 on “Socially Interactive Agents as Peers” [Cassell 2022] of this volume of this handbook.
The appearance of an agent is a similarly important design choice that can have consequences on learning and motivational outcomes [Veletsianos 2010]. Basic design choices like age, gender, or ethnicity may impact how a PA is perceived and in what implicit assumptions are made by a learner. Further, the level of realism in the appearance, or whether the agent is humanoid or not, also likely influences the nature of the interactions a learner may, or choose not to, have with an agent. For example, research has shown that the age, gender, or ethnicity of an agent may challenge or reinforce stereotypes that a learner may have about the role the agent is playing [Baylor and Kim 2004], and research has also shown that learners can have biases against PAs based on the PAs skin tone [Zipp and Craig 2019]. Yet, research does not always provide clear answers as to how agents should be designed in regard to their appearance. There may be benefits to offering learners a choice of agent, specifically for learners from minority and underrepresented groups [Baylor 2005]. Relatedly, evidence has been found for the similarity attraction hypothesis which says that people are naturally drawn to others that look and behave in ways that are similar to them. In a study of over 500 middle school-aged learners, there were motivational benefits for gender matching between a learner and a PA [Ozogul et al. 2013]. However, another study found that learners who chose to learn with PAs that were of the same ethnicity as themselves performed worse on learning tests than those who chose to learn with PAs of a different ethnicity than themselves [Moreno and Flowerday 2006]. As such, it is important to note that such findings are challenging to generalize due to the vast individual differences between learners, including important factors like age, interests, and background.

As noted, we have only shared a few examples of a vast body of complex empirical research on the design of PAs for learning. Heidig and Clarebout [2011] provide a far more in-depth framework for capturing the design aspects of implementing PAs.

### 21.3.2 Internal Models and Properties—Policies, Actions, and Behaviors

While the external properties of a PA can be important, there is substantial evidence that the pedagogical activities of intelligent systems can and do impact learning. The best ITSs are known to improve learning by effect sizes of 1.0–1.2 standard deviations [VanLehn 2011, Kulik and Fletcher 2016]. Given that most of the ITSs in these studies do not include an embodied agent, it is likely that the learning gains are largely due to the pedagogical interventions provided by the system.

Dehn and van Mulken [2000] refer to these as internal properties of a system given that they involve pedagogical activities (e.g., assessing learner actions, deciding what feedback to give, delivering help) and do not require use of an embodied
agent. The pedagogical decision space of intelligent tutors includes critical decisions points such as how to select problems, how often to give feedback and hints, what level of detail to include in hints, how to respond to emotional changes, and more [Anderson et al. 1995, VanLehn 2006]. In this section, we focus primarily on how PAs enhance and extend this design space and explore some possible features of PAs that may improve upon a disembodied tutoring system (Figure 21.1).

Inhabiting the Same Space as the Learner  In the physical world, a teacher or coach exists in the space and can interact with the same objects as the learner. For example, if a learner is learning how to paint, an instructor can step in to demonstrate how to hold a brush in their hand or perform a stroke for that learner on the canvas. PAs that exist inside virtual environments alongside a learner's avatar also have this ability. A PA might lead the learner to a specific location in a virtual world by having the learner follow them (navigational support). Naturalistic supports are possible as well by gaze changes (where the agent looks) or pointing in the virtual environment. A PA can also demonstrate a skill enabling the learner to play more the role of apprentice or provide additional scaffolding during problem-solving for a learner by doing a step for them or undoing a mistake. One of the first embodied PAs, STEVE, exhibited these actions for repair of HVAC (Heating, Ventilation, and Air Condition) systems on ships [Rickel and Johnson 1997].

PAs can also play meaningful roles in narrative-based learning environments by portraying part of a larger story and collaborating with the learner to solve problems. A good example of this approach can be found in Crystal Island [Rowe et al. 2011], a multiagent educational game for middle school biology. In the game, the learner is trapped on an island with everyone getting sick from a mysterious illness.
The PAs play various roles, including scientists and medical personnel. As one the last remaining conscious inhabitants of the island, the learner must conduct experiments and work with the remaining agents to learn about their progress and unravel the mystery.

**Emoting and Social Signals** While a certain level of emotion is certainly possible in text, such messages can be more impactful when combined with corresponding physical gestures and facial expressions. One such example can be found in Coach Mike, a virtual agent who supports learners at a museum exhibit for computer programming [Lane et al. 2011, 2013]. Coach Mike used synthesized speech along with a number of gestures and animations to build rapport with museum visitors and convey emotional messages such as empathy and excitement (Figure 21.2). Gestures are also known to play an important role in learning, especially when they align conceptually with the content. For example, PAs that used gestures during a lesson on mathematical equivalence were more effective than agents that did not [Cook et al. 2017]. Importantly, this preliminary evidence suggests the fundamental role of gesture in promoting learning and is impossible without the use of a PA (for computer-based learning).

**Conversational Interactions** One of the most obvious benefits of using a PA is the potential for more naturalistic communication and to achieve learning gains that approach human tutoring [Bloom 1984]. Basic research on human communication tells us that humans instinctively naturally bring a social frame to interactions with machines [Reeves and Nass 1996], which is promising for work on agents given
that the aim is also to emulate human communication, both verbal and physical. As mentioned in Section 21.2, intelligent tutors based on conversational interaction were around long before agents were possible [Graesser et al. 2001]. What is possible with a PA is to go beyond the textual components of a conversation to incorporate additional conversational behaviors such as nodding, headshaking, eyebrow raising, and cognitive gestures (e.g., hand to chin, head scratching). These behaviors all contribute to establishing a sense of social presence, which has been positively linked to learning [Frechette and Moreno 2010]. While not immediately applicable to achieving any particular learning goal in a conversation, such interactions do contribute to at least two related aims. First, improved conversational skills can help an agent build rapport, an important element of any conversation, but especially teaching and learning [Gratch et al. 2007, Estepp and Roberts 2015]. Second, building trust between a human and agent is widely regarded as important [Rheu et al. 2021] and related to conversational skills and relationship building. Early studies have shown, however, that trust primarily influences perceptions of an agent and, so far, has minimal impact on learning outcomes [Schroeder et al. 2021].

21.4 Empirical Status of Pedagogical Agents

Researchers have explored the use of PAs in a wide variety of learning environments, for a diverse population of learners, and in varied contexts for more than 20 years. Meta-analyses of PA research have consistently showed that PAs can have small, positive effects on learning when compared to conditions that do not have PAs present [Schroeder et al. 2013, Castro-Alonso et al. 2021]. In contrast, systematic reviews of the literature have not found consistent, statistically significant benefits of PAs for facilitating learners’ motivation [Heidig and Clarebout 2011, Schroeder and Adesope 2014], although Schroeder and Adesope [2014] did find more positive effects than negative. Finally, in regard to cognitive load, Schroeder and Adesope’s systematic review found few significant effects, and non-significant effects were also mixed with some studies finding PAs increased cognitive load while others found PAs decreased cognitive load.

It is important to note that researchers have argued that examining the overall effectiveness of PAs across these varied contexts is too broad a question that overlooks the nuances of context and PA design [Heidig and Clarebout 2011, Schroeder and Craig 2021]. This is exemplified by the findings of Schroeder et al.’s [2013] meta-analysis, which found strong, positive effects when K–12 students learned from PAs. While the number of studies included in the analysis was relatively small, the effect stands in contrast to the small positive effect found for postsecondary learners.
In order to help understand the design and contextual factors that may influence the impacts of PAs, many researchers have moved beyond comparing PAs to non-PA conditions and instead have begun investigating specific design aspects of PAs, the impacts of PAs on various types of outcomes other than learning, or exploring the use of PAs in novel contexts.

21.4.1 Research-based PA Design—Exploring Agent Voice

Designing an effective PA requires attention to a wide range of details, as illustrated in this chapter and also by Heidig and Clarebout’s frameworks. Discussing all of these components in detail is outside the scope of this chapter, but we feel it is important to understand the depth of the research that takes place within each major design component. As such, in this section we explore one aspect of PA design that has been investigated in current research and is also broadly applicable: in the way the agent communicates with the learner. This aspect of PA design is multifaceted, spanning various design components such as the role of the PA, its voice and language, and its backstory. For the purposes of our example in this section, we explore only the voice and language used by a PA in detail.

A critical question when designing PAs is how they will communicate with the learner, and therefore a designer has a few critical choices to make. First, one must choose a type of voice for the PA (i.e., machine-generated or recorded human speech), which will have ramifications with regard to the qualities of the voice itself, and they also must consider different types of speech patterns.

Choosing what type of voice to use for a PA used to be a relatively simple process. Research around the voice principle [Mayer 2014] had shown that in multimedia learning situations recorded human speech consistently lead to higher learning outcomes than machine-generated, or text-to-speech (TTS), voices [Mayer et al. 2003, Atkinson et al. 2005]. However, there is a growing body of research investigating the type of voice one should use to aid learning in multimedia learning situations [Craig and Schroeder 2019] and, more specifically, the type of voice that a PA should use to communicate with the learner [Craig and Schroeder 2017, Chiou et al. 2020], which is showing that modern TTS voices can facilitate learning as well as recorded human voices. As such, PA designers should consider their software and its technical requirements as part of their decision-making process when picking a type of voice for the PA to use. Specifically, if a system needs to be able to reply with a wide variety of responses, or can create its own responses rather than replying with pre-made phrases, this may be more easily implemented with a TTS engine rather than a recorded human voice. In contrast, if one is creating a video that contains a PA but the learner does not interact or converse with the PA, they simply listen to a recording, a recorded human voice may be preferable in some
circumstances. To conclude our discussion of voice type, it is important to note that research cited in this section was often dealing with PAs that did not naturally and adaptively converse with the learner but rather conveyed either pre-scripted speech or were simple recordings of PAs delivering a narrative. It is not clear if TTS or recorded human voices would be superior for facilitating learning in systems that generate unique responses in the course of a conversation with a learner.

While the choice of TTS compared to recorded human speech may seem relatively straightforward due to software requirements, one must also consider the qualities of the voice itself. For example, the prosody of the voice may be important. Prosody refers to the stress, intonation, and rhythm of speech that can provide important information to a listener [American Psychological Association 2014, Shintel et al. 2014, Davis et al. 2019]. This is potentially an important factor in PA design, especially if the learner is communicating in their non-native language [Davis et al. 2019]. Few studies have examined voice prosody in relation to PAs specifically. However, Davis et al. [2019] investigated the use of PAs with strong or weak voice prosody for teaching non-native speakers. Their results revealed no statistically significant differences on learning outcomes, but they concluded that more research is needed to better understand what features of human voices are important for facilitating learning. A PA designer must then decide how important prosody is in their context. This potentially has implications for if they can use a TTS voice or would be better served by a recorded human voice, depending on the software they have available. Please see Chapter 6 on “Building and Designing Expressive Speech Synthesis” [Aylett et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021] on expressive speech synthesis for further details.

### 21.4.2 Communication Styles

A PA designer must also decide how the PA will use language when communicating with the learner. In the research literature, this is often characterized by using a conversational style compared to a more formal style. In the broader literature outside of PAs specifically, researchers have investigated this question in some depth. A meta-analysis found that using a conversational style aided both retention and transfer outcomes, with a noteworthy effect on learning transfer outcomes, \( d = .54 \) [Ginns et al. 2013]. Importantly, the transfer outcomes were not significantly moderated by either the specific language (e.g., English, German) nor the learners’ grade level, and Ginns et al. found that the conversational style provided strong benefits for effective cognitive processing \( (d = .62) \). It is important to note, however, that Ginns et al. found that the length of the intervention significantly moderated the transfer outcomes, and few studies were longer than 35 minutes in length.
As such, it seems that, especially for relatively short interventions, conversational style language may benefit learning.

### 21.4.3 Summary of Empirical Status

Veletsianos and Russell [2014] observed that while PAs have often been associated with providing learning benefits to learners, it is not necessarily their presence alone that typically facilitates learning but rather it is often the pedagogical strategies that the PA facilitates that may benefit learning compared to non-PA conditions. In addition, Craig and Schroeder [2018] suggested that PAs may benefit learning most when the social aspects of their use can be leveraged, such as when they can act as conversational partners that convey emotions and model and facilitate effective learning strategies. As such, it is prudent that designers consider what educational benefits a PA could provide in their specific system within their context and design each aspect of the PA as appropriate.

There is a vast amount of research around many of the design choices one needs to make when designing a PA. It is also important to remember that there can be interactions between different PA design components. For example, the role the agent plays in the environment may also influence how an agent should communicate, either conversationally or more formally, and thus a designer should consider all aspects of their PA design together rather than in isolation.

### 21.5 Relationship with IVA and SR Research

Research on IVA and SR has many shared goals and similar challenges, with the key (obvious) difference being the use of a physical or virtual agent/robot. As mentioned in Chapter 1 on “Introduction to Socially Interactive Agents” [Lugrin 2021] of Volume 1 of this handbook [Lugrin et al. 2021], comparisons between these two modalities are revealing strengths and weaknesses of both. The challenges, costs, and limitations of a robot may present challenges to implementation; however, the benefits of interacting with the physical world of a learner has great potential for supporting learning in deeper ways. For example, learning very often occurs in a social context (classrooms, museums, outdoors) and so bringing a PA from a computer screen into these physical contexts has obvious benefits, such as in-context guidance and coordinating group work on a collaborative, physical task (e.g., planting a garden). Emerging research on using SR for education is beginning to explore these benefits [Mubin et al. 2013, Belpaeme et al. 2018].

Research in a wide range of areas has clear applications for both IVAs and SRs. For example, the challenge of increasing naturalistic interactions with speech and language understanding, natural language generation, dialogue modeling, animation of nonverbal behaviors, and more have value broadly for IVA and SR, and
specifically for PAs. In one study of college students, robots were viewed as credible instructors by students [Edwards et al. 2016], due in part to believable and natural interactions that resulted from remote human control (e.g., “Wizard of Oz”). Researchers have also shown that robots that emote are more effective in terms of supporting learners, due most likely to the feelings of empathy and concern that learners experience [Saerbeck et al. 2010, Obaid et al. 2018]. Clearly, underlying algorithms that detect, express, and enable naturalistic interactions can contribute to both virtual and physical PAs.

21.6 Current Challenges

PA researchers face a variety of challenges in their work. Notably, PAs can be time and resource-intensive to develop. One positive outcome of this challenge is that researchers have developed systems that vary greatly, from individual agents delivering a narrative without any back and forth dialogue with the user [Schroeder 2017], through ITSs with one or more PAs [Graesser et al. 2017], teachable agents [Silvervarg et al. 2021], and even virtual worlds [Rowe et al. 2011]. However, the challenge of creating PAs persists. To date, there are few user-friendly software programs with a graphic-user interface that can help a researcher create a PA, and likely fewer yet that allow for the development of interactive PAs. While some platforms do exist to facilitate the interactive PA creation process [Hartholt et al. 2013, Lane et al. 2015], the general process of authoring a new intelligent agent, embodied or not, still represents one of the most persistent and complex challenges for the field [Murray et al. 2003, Sottilare et al. 2017, Dermeval et al. 2018], in part due to its interdisciplinary nature.

Like many fields, the PA literature is also limited by the lack of psychometrically strong measurement tools within the field. These limitations were noted by researchers more than 15 years ago [Clark and Choi 2005]; however, relatively few validated measures developed specifically for use with PAs exist in the literature. In fact, much of the literature still relies on the Agent Persona Instrument [Ryu and Baylor 2005] or Schroeder et al.’s [2018] revision of it. While this instrument has been useful for examining different aspects of the PA persona, researchers have found that it does not strongly predict or influence learning outcomes [Schroeder 2017, Schroeder et al. 2018]. The field would benefit from new instruments being developed that can help push the field forward. For example, while a PA may only facilitate a small learning benefit in some circumstances, which may not rationalize the cost and effort required to create it, there may be other benefits (e.g., social or emotional outcomes) from the agent that are viewed as more important. However, without adequate measurement techniques it is challenging to approach those areas of work.
Finally, the field has also found challenges in creating PAs that express emotions well and understanding if and how these emotions impact learners. However, recent work is making advancements in this area. For example, Lawson et al. [2021b] compared a human instructor and a PA and found that learners could differentiate positive and negative emotions in both cases. In a different study, Lawson et al. [2021a] found that learners were able to differentiate a PA’s emotions and generally perceived positive PAs better than negative PAs. Meanwhile, other lines of work have examined how to best create believable PA emotions. Anasingaraju et al. [2020] found that the PA’s body was the most important aspect of creating the believable expression of emotions. Together with basic research in affective computing to model and recognize emotions [Calvo et al. 2015], these represent significant ongoing challenges for the field.

21.7 Future Directions

While many ongoing research areas related to PAs will continue to see much attention in coming years, some areas remain elusive. In this section we provide our opinions on what the field should emphasize in the coming years.

21.7.1 PA Support for Lifelong Learning

Much of what we know about learning from PAs and specific aspects of PA design comes from studies that are of relatively short duration. For example, many studies investigating specific design aspects of PAs contain interventions less than 10 minutes in duration [Davis 2018, Chiou et al. 2020], whereas a longer duration study may be a few hours [Li and Graesser 2021]. This being the case, a long-standing question in PA research is how learners will learn from [Gulz 2004, Choi and Clark 2006] and develop relationships with PAs over longer durations [Veletsianos and Russell 2014]. For example, one question posed by researchers is if the benefits of PAs are simply due to novelty effects that may wear off over time [Choi and Clark 2006]. Importantly, however, Veletsianos and Russell [2014] raised a number of questions about the ethics related to designing agents that make emotional connections between the learner and the agent. In terms of technical demands of supporting lifelong learning, a PA would require a long-term model of a learner that spans different domains and contexts. This challenge is well-known [Kay 2008] and different approaches have been pursued that seek to inform pedagogical decision making, such as meeting learner preferences and modeling memory decay [Kay and Kummerfeld 2012, Pavlik et al. 2021]. This area surely needs more research to address it, especially if longer-term PA interventions are pursued by researchers.
21.7.2 PA Support for Medical and Health Literacy

Research on building conversational agents for health literacy and supporting medical information needs has grown substantially in recent years [Montenegro et al. 2019, Morrow et al. 2021]. Given the cost of human–human support, and the challenges associated with printed materials, this line of research seeks to provide intuitive and natural ways for acquiring information about diagnoses, treatments, and self-care. In overlapping with support for lifelong learning, Bickmore et al. have implemented numerous agents that provide such support [Bickmore and Picard 2005]. This work has provided a framework for agent–human relationship building as well as methods for behavior change [Bickmore et al. 2013]. Agents have particularly been successful with aging adults and supporting complex medication schedules and self-care techniques [Azevedo et al. 2018]. While such systems have implicitly recognized patients as learners, such as by answering questions and providing information, there remains a long list of techniques from PA research that focus on supporting learning that have the potential to transform agents in the area of health literacy. Please consult Chapter 24 on “Health-Related Applications of Socially Interactive Agents” [Bickmore 2022] of this volume of this handbook for further details.

21.7.3 New and Emerging Areas for PA Research

While researchers have long examined the effects of PAs on learning, researchers have also began looking at factors other than learning that PAs may influence, such as learners’ trust in the agent. For example, Chiou et al. [2020] examined how different types of PA voices influenced learners trust in the PA. Outcomes such as trust and credibility, as well as other outcomes such as social presence, may be particularly important as researchers are moving PAs beyond science and mathematics teaching contexts to more sensitive topics such as health (Section 21.7.2) and counseling [King et al. 2020].

Another emerging line of research is how various learner perceptions of the PA can influence learning outcomes [Schroeder et al. 2017, 2018, 2021]. Presumably, the goal of these lines of work is work toward identifying the specific outcomes that most influence learning, then researchers can isolate specific design features to help facilitate these outcomes and improve learning from PAs.

21.8 Summary

In this chapter, we have summarized research on PAs by describing the “external” design considerations (age, gender, ethnicity, form, etc.), the “internal” properties that define how such agents can promote learning, and documented specific interaction techniques for which PAs are unique or can improve upon previous
non-embodied ITSs. We summarized current research on PAs and provided an overview of findings as they relate to the design of agents, pedagogical strategies, and roles. While empirical evidence is modest to date, it is nonetheless positive for many contexts and learners (young and older). In the future, we suggest that PA research should focus on providing longer term support, across domains, integrate learning technologies into health literacy agents, and continue to focus on specific ways that PAs connect with learners either via enhanced trust, social presence, and quality of learner experiences.

References


22

Socially Interactive Agents as Peers

Justine Cassell

22.1 Introduction

Embodied conversational agents (ECAs) were originally implemented as experts in their respective domains—bankers, real estate agents, or health care workers, for example. In this sense, those early ECAs, that provided information to their users, were similar to early dialogue systems, such as those developed to give information about trains [Ferguson and Allen 1998]. Even early chat systems, less concerned with accomplishing a task, still represented expertise in a domain, such as Rogersian psychology [Weizenbaum 1966]. More recently, smart speakers, such as Siri, Alexa, and Google Home, still give access to information but represent themselves as assistants rather than experts. It is an interesting twist (and beyond the scope of this paper to discuss); however, the relationship still embodies a power differential between the user and the system.

Intelligent tutoring systems and pedagogical agents (see Chapter 21 on “Pedagogical Agents” [Lane and Schroeder 2022] of this volume of this handbook) have followed much the same vein as early ECAs, focusing in vast majority on agents that act as expert tutors. These tutors pose questions, assess student knowledge, and selectively deliver tutorials for skills that students have not yet mastered. They have a representation of a subject area, of best practices in how to teach, and a model of the student that is updated as knowledge is acquired. Historically, many of these systems grew out of research into theories of how people learn, particularly by Herb Simon, John Anderson, and their respective students (e.g., Anderson et al. [1985]). Tutoring has been shown to lead to significant learning gains when carried out by human tutors and has also transferred well to AI-based intelligent tutoring systems that have been used with impressive results in thousands of K–12 and university classrooms worldwide [Koedinger and Corbett 2005]. Increasingly, these systems have amplified their performance by including the ability
to recognize student emotional states. However, even those intelligent tutoring systems that include emotional awareness focus in large part on the emotions of the students, with the adult tutor monitoring and responding but not engaging at the level of an equal peer (e.g., Zakharov et al. [2008]).

While tutoring agents are derived from theories of how people learn, ECAs have been based on theories and descriptions of how people behave with one another [Cassell 2000]. As the paradigm of ECAs developed at the end of the last century, it was natural, then, that eventually researchers begin to include children’s behavior with one another as the model for their systems. This was particularly the case for researchers who had studied children’s interaction with other children through computational systems [Cassell 2002] or children’s interaction with non-embodied computational systems that listened and engaged without demonstrating expertise or assuming the stance of a teacher—what were called Story Listening Systems [Cassell 2004].

Children’s interaction with peers is enormously influential in their cognitive, social, and emotional development [Ladd 2005], and story listening systems evoked and supported children’s stories as ways of scaffolding cognitive, social, and emotional development. These story listening systems eventually became embodied in child-like ECAs called virtual peers [Cassell et al. 2000].

In what follows we focus on those virtual peers—the kinds of ECAs and robots where the computer takes on the role of a peer, often communicating with age-appropriate language, and even looking like a child of the same age as the young person interacting with it. For the most part, the application of contemporary virtual peers is learning, and so in this chapter we narrow the focus to learning (as opposed to interactions to simply pass the time without other goals, for example) among students up through university, but we broaden the discussion beyond the classic school curriculum to informal learning outside the classroom. We also broaden the focus beyond what are sometimes called “core literacies”—reading, writing, and arithmetic—to include the learning of socio-emotional skills such as curiosity and establishing social bonds. In this way, we focus both on ECAs without social competencies and also specifically Socially Interactive Agents (SIAs) as peers.

While many topics could be covered in a chapter of this sort, we organize the remainder of the chapter around the nature of natural peer interaction in young people and its role in development and learning. We begin by outlining some core ways that communicative behaviors among peers play a role in children’s development and in learning. We then turn to the implementation of virtual peers and how they can engage with children along the same key dimensions. Finally, we turn to ethical considerations and a roadmap for future work.
22.2 Peer Interaction

Children in virtually all societies spend significant periods of time with their peers. The context may be school or home, overseen by adults or free of adult intervention, while playing or working in dyads or in larger groups. In all of these contexts, peer interaction plays several important roles in children’s development. For example, while adults may provide notions of culturally appropriate narrative structure for storytelling, it is often with peers that children learn how to construct stories for an audience who cannot see what they are thinking [Vygotsky 1978, Christie and Stone 1999]. Likewise, while adults remain for the most part polite and encouraging when teaching children, children amongst themselves engage in argument and disagreement that in fact provides a unique resource for problem-solving and learning [Azmitia and Montgomery 1993, Pellegrini et al. 1998].

There has been increased focus on peer interaction in the developmental literature. This may well be because much of the developmental literature before a certain period considered that children spend their childhoods becoming adults, reproducing the best versions that they can of adult norms, and acquiring adult ways of interacting with peers. More recently, however, there has been a realization that children are members of a culture that is different from that of the adult world [Cook-Gumperz and Kyratzis 2001, Ladd 2005]. As Kyratzis [2004] notes,

Children are not merely unformed adults... they reformulate social categories (e.g., friendship, gender) appropriated from the adult culture in ways that are sensitive to context... and reflective of children’s personalities and momentary goals and agendas in the culture of peers (p. 626)

In this context, Zadunaisky and Blum-Kulka [2010] have identified peer interaction as a “double opportunity space,” on the one hand offering opportunities for children to construct their own kinds of childhood cultures, with important rules and roles for the group, and on the other hand providing opportunities for individual development in cognitive, linguistic, and social spheres. While action certainly plays an important role here (such as stomping away after losing a game), language and associated nonverbal (gesture, facial expression, posture shifts) and paraverbal (prosody and loudness) behavior is the primary way that both individual and social development takes place in the context of peers. For our purposes in this chapter, certain of these communicative behaviors with their peers are particularly important:

1. Assuming and switching roles and participation frameworks in pretend play and collaborative learning.
(2) Taking positions in debate and disagreement using both language and embodied behaviors.

(3) Constructing variants of speaking styles to demonstrate affiliation or allegiance to a particular cultural group and particular identity.

(4) Different ways of communicating.

(5) The role of social interaction in development and learning.

### 22.2.1 Roles and Relationships in Peer Interaction

Many of our daily activities rely on designated roles, and relationships among those roles, that set expectations, rights, and responsibilities for how we participate in an activity with a partner [Allwood 2000]. Asymmetric role-relationships like lecturer–audience, teacher–student, or mentor–mentee, often come with fairly rigid expectations about who should say what when, and how the other should respond. For example, in large lecture classes, it is rare that students challenge the choice of homework assigned. Some of these roles—such as CEO–employee—are based on status or power relationships of the participants and remain relatively stable across activities; however, others are based on expertise about the current topic or are locally allocated in other ways. For example, a student in my small conversational agents class may that same evening be my yoga instructor.

Peer-to-peer interaction, on the other hand, reflects a socially symmetric relationship that gives both participants the same rights and expectations to speak, even while roles may shift seamlessly over the course of an activity. Two children building a block tower in a kindergarten classroom may start on equal footing until one steps back and says, “you’re doing it wrong.” In this symmetric relationship, however, the first child has the right to disagree and continue building the tower as before. In the workplace, two colleagues may begin revising a document together until one notices that the other is deleting text and suggests the other begin working on the references. In this symmetric relationship, the colleague who is deleting text may defend her choice and continue to delete. The social equality of the relationship has benefits for the roles that each takes. Neuman and Roskos [1992] observed that children engaged in instructional conversation with a peer negotiate and coach each other’s literacy activities. Unlike the exchanges in adult–child conversation, children instructing one another often reverse roles and attribute the role of the more capable peer according to the purpose of the play at hand. Based on studies such as these, taking roles has become an important part of the formal curriculum in many subjects. One striking example is in “pair programming” where students take turns being the driver and navigator in the writing of code [Campe et al. 2020].
22.2.2 Debate and Disagreement
In the absence of hierarchy, debate and disagreement can flourish, and these play an important role in learning and development. For adults, differences in life experience and perspective can lead to what has been termed “productive disagreement.” This concept has been popularized recently as a way of overcoming unproductive political conversations [Benson 2019]. However, the notion of productive disagreement in children goes back at least as far as Piaget. He described the benefits of cognitive conflict among peers as leading children to revise their views, unlike parent–child interaction where children are more likely to take as given the opinion of the adult ([Piaget 1959], cited in Tudge and Winterhoff [1993]).

There are in fact many positive functions of conflict in children’s peer interactions. Some of these can be traced to the need to deploy sophisticated linguistic strategies such as justification of one’s opinions [Eisenberg and Garvey 1981]. Others derive from children’s ability to maintain an argument as opposed to resolving it quickly [Genishi and DiPaolo 1982, Maynard 1985]. Recent studies have focused on the complex moment-to-moment interactional processes by which kinds of conflictual talk, such as gossiping about one another, teasing, and debates over who has won and who has played fairly in childhood games, are used among peers [Goodwin 1990]. Piaget traces disequilibrium in beliefs provoked by these kinds of conflict, and the subsequent change in beliefs that leads to learning [Piaget 1947:1950]. A more contemporary interpretation of this approach, referred to as socio-cognitive conflict, demonstrates the ways that cognitive conflict, specifically embedded in social situations, can lead to better learning than when individuals learn alone [Mugny and Doise 1978]. Sinha et al. [2017] observed how children in a small group, working on building a Rube Goldberg machine (a contraption where the goal is to make the mechanism as complex as possible), challenged and disagreed with each other’s ideas. These challenges and disagreements were in fact the most important predictor of an increase in curiosity in those who were challenged, and also in the group as a whole.

22.2.3 Varying Speaking Styles
In addition to seamlessly shifting roles, children in interaction with one another also seamlessly shift their ways of speaking, adopting and adapting language that they hear spoken by various adults around them, as well as constructing their own variants of speaking styles to display allegiance to a broader youth culture. These speaking styles play an important role as children experiment with who they want to be, and how they want to be perceived. They also allow children growing up in situations where different dialects or languages are in contact to mark their affiliation to dominant and minority ethnic, racial, and gender identities [Rampton 1995].
A 2021 article in the New York Times newspaper described in detail the struggles of a young Black woman as she moved back and forth between her poor family in New York City, and the elite boarding school she attended in Hershey Pennsylvania. A large part of the girl’s struggle, as she described it, was around the decision of whether to adopt “speaking like a white person” full-time, and whether doing so would be a betrayal of her family [Elliott 2021]. Young people (and adults) who move between marginalized and mainstream communities often report this kind of code-switching as a way to maintain their affiliation with their home community, while also making their way in a world where the standard dialect is associated with increased earning power and other kinds of success [Kallmeyer and Keim 2003]. However, the movement back and forth is not without the kind of stress that is reported in the New York Times article, and Ogbu [2008] has coined the term “oppositional culture” to describe the ways that school systems may inadvertently set up a situation where the student feels the need to define her identity contra the expectations of the school, and for that reason to refuse the dialect that the school insists on.

In addition to code-switching of this kind, children may engage in pretend play where they assume the voices of the characters they are enacting. In this context, even quite young children are capable of “playing teacher” or “playing mother,” both roles they have participated in only as observer and interlocutor (Goodwin [1993], cited in Kyratzis [2004], Cekaite and Aronsson [2005]).

### 22.2.4 Difference in Peer Social Interaction

Thus far we have described peer interaction and peer learning in neurotypical populations.

However, non-neurotypical individuals, such as those diagnosed with autism spectrum disorders (ASD), tend to exhibit social-emotional skills that differ from their neurotypical peers, which impact their peer interactions in fundamental ways, such as difficulties with integrated verbal and nonverbal communication and with interpersonal relationship development, and insistence on behavioral and environmental sameness [American Psychiatric Association 2013]. ASD is called a spectrum disorder because each individual with ASD may demonstrate the above traits to a greater or lesser extent. Nevertheless, the research and clinical communities have identified what they refer to as “high-functioning autism” or Asperger’s, where individuals tend to exhibit different socio-emotional skills from their typically developing neurotypical peers, but to a lesser extent. For this reason, individuals from this population are likely to be mainstreamed into schools where they will interact with neurotypical peers. In contemporary classrooms, where group learning is the norm in many countries, these individuals may have
difficulties in benefiting from the learning context if they cannot engage in productive interaction with neurotypical peers, and their neurotypical peers likewise have difficulties engaging with them.

In reflecting on the role that virtual peers can play in the learning of socio-emotional skills as well as disciplinary topics, it is therefore important to understand the nature of social interaction and friendship among individuals with ASD, and in their interaction with neurotypical peers [Bauminger 2002, Bauminger et al. 2003]. Diagnostic criteria for autism define it in terms of “restrictive interests and behaviors” and “deficits in social interaction and communication” [DSM IV 1994]. The mechanism underlying these behaviors is often described as an “impaired theory of mind” [Senju 2012] whereby the individuals in question may have difficulty “de-centering” or imagining a perspective on the world different than their own. In the lay literature, this is often described as difficulty in imagining the thoughts and feelings of others, although this does not align perfectly with the technical definition of the cognitive capacity to infer other’s mental states. It is true that individuals with ASD may find it difficult to understand the behavior, and intentions behind that behavior, of others in social interaction. On the other hand, the same is true in the other direction, where neurotypical individuals find it difficult to understand the intention of individuals with ASD in social interaction [Humphrey and Symes 2011]. This has been referred to as the “double empathy problem” [Milton 2012]—a breakdown of empathy and mutual understanding between people with differing ways of experiencing the world, or the “cross-neurological theory of mind” [Beardon 2017].

Perspectives such as these are important not just for improving our understanding of individuals with ASD but also for our understanding of social interaction in general and peer social interaction in particular. These theories point out the kinds of obstacles posed by different experiences (including language or dialect spoken, socio-economic status, as well as other life experiences) in creating a social bond that can productively support learning and other aspects of development.

22.2.5 Social Interaction during Task Behavior

It would be inappropriate for a teacher to share his problems with his spouse during his high school class on linear algebra. The same teacher, however, while grading papers with a colleague, may well share those marital difficulties. Similarly, when friends engage in a task, regardless of its nature, they often refer to past shared experience, disclose their feelings about what they are doing, and, depending on the cultural context, they may engage in mutual teasing. They also engage in more mutual eye gaze and greater alignment in their speech rate. These behaviors serve to build and maintain rapport between the participants [Cassell et al. 2007].
Cappella [1990] goes so far as to argue that the “construct of rapport is arguably one of the central, if not the central, construct necessary to understanding successful helping relationships.” In fact, such behaviors between friends, apparently unrelated to task, are associated with higher learning gains in a peer tutoring task, while the same behaviors between strangers are negatively correlated with learning [Finkelstein et al. 2012, Wang et al. 2012].

The reciprocal relationship between peers is in part what allows this kind of off-task behavior, whether the peers are children or adults. Sociality and the interweaving of social and task talk is perhaps the most representative behavior of peer interaction, and it plays an essential role in cognitive development and learning [Hartup 1996]. Interleaving the two is a primary way in which peers bond and create solidarity and rapport with one another, manifest their alliances, and demonstrate that their relationship is special and not subject to the same politeness rules of the wider culture [Kyratzis 2004, Zhao et al. 2014]. As noted above, the off-task talk that achieves these goals may on its surface appear quite negative, such as joking at the expense of the other or teasing [Corsaro 1997, Kyratzis 2004]. As a clue to the affiliative function of these apparently disruptive behaviors, researchers have observed that children justify their actions during arguments but do not try to resolve their disagreement, as the conflict remains an important part of healthy peer interaction [Genishi and DiPaolo 1982, Maynard 1985].

22.2.6 Peer-based Learning

In peer-based learning, students learn both from and with one another, in dyads or small groups. They learn by explaining their ideas to others and by participating in activities where they can learn from their peers. Not only do the learners develop skills related to the material being discussed, they also develop invaluable interpersonal skills as they work with others, give and receive feedback, and evaluate their own performance [Blum-Kulka and Dvir-Gvirsman 2010, Sin et al. 2019]. Considerable convincing evidence has accumulated demonstrating that if students are asked to discuss their answers with other students, their understanding of the material increases more than if they did an active learning component on their own [Bonwell and Eison 1991, Johnson et al. 1991]. Similarly, structured group work can promote problem solving at a higher level than possible with individual effort alone [Millis and Rhem 2010]. Additionally, there is evidence that peer learning may help reduce attrition rates and increase engagement [Crouch et al. 2007, Porter et al. 2013]. Peer-based learning is particularly effective for underperforming students [Robinson et al. 2005] and those in low-resource environments [Jacobson et al. 2001].
22.3 Research on Virtual Peers and SIAs as Peers

These studies demonstrate that learning does not happen in a cultural or social void. When students learn together, a number of cognitive advantages accrue, linked in large part to the relative social equality of their relationships [Webb 1989]. It is not just the tutee who learns, however. The “tutor effect” refers to the fact that explaining a subject to somebody else can lead to learning [Sharpley et al. 1983]. Self-explanation, in the absence of a learning partner, has been linked to learning in many studies. However, the tutee’s challenges and questions also may encourage deeper reflection about the topic on the part of the tutor [Webb 1989]. Nevertheless, simply tutoring another is not sufficient as the experience of being both tutor and tutee may also encourage students to view learning as socially desirable [Rohrbeck et al. 2003] and to better understand how to learn. In fact, reciprocal tutoring [Palincsar and Brown 1984], where the students take turns as tutor and tutee, has been shown to be an important tool in classrooms and forms the basis for many of the SIAs we will discuss below.

Learning how to read and write is particularly facilitated by peer talk [Teale and Sulzby 1986, Fuchs and Fuchs 2005], as is the learning of math and science [Newcomb and Brady 1982]. While these subjects require cognitive skills, they also rely on self-confidence and self-efficacy and perhaps for that reason, peers play a particularly important role [Rohrbeck et al. 2003]. In this context, as mentioned above, in young adults peer tutoring has been formalized in what is called “peer (or pair) programming,” a paradigm that has been shown to be a particularly effective approach to learning for those students traditionally marginalized in STEM (Science, Technology, Engineering, and Math), such as women or underrepresented minorities. Here, correlations have been found between the strength of the bond between peers and their learning gains [Zhong et al. 2016].

It is not just these school subjects that benefit from peer learning, however. Peer interaction plays a key role in the learning of a first and second language [Sato and Ballinger 2016]. In fact, as any parent who has moved to a country with a different language can attest, children most rapidly learn the second language in their interactions with peers, sometimes seeming to learn to speak a new language overnight!

Research on Virtual Peers and SIAs as Peers

In the first sections of this chapter, we discussed the value of peer interaction for learning and development. One might ask, then, what the value is of virtual peers—why not just stick to human peer contact? While the kinds of interactions discussed above are extremely valuable, peer-to-peer learning is not always possible. When peers cannot be found, when scheduling or distance or a pandemic makes
assembling a dyad or group impossible, or when the student doesn’t get along with
the only available peers, or is not understood by them, peer-based learning may not
be viable. In these instances, SIAs as peers can make peer-based learning accessible
by providing interactions and content matched to the learner’s (or learners’) stage
in the learning process. Matched to the personal and interpersonal abilities of indi-
viduals or groups of learners, virtual peers can be capable of establishing peer-like
bonds that sustain learning. Virtual peers, too, can maintain user privacy, which
may allow learners to be more vulnerable about not understanding a particular
subject matter.

Before we go any further, a note about terminology and its history. The first work
on virtual peers was published in 2000 [Cassell et al. 2000]. In the same year the
first paper was published on imbuing ECAs with social competencies [Bickmore
and Cassell 2000]. Both strands of research continued independently; however, it
would take 12 years before the two topics were joined, in a study of rapport in peer
tutoring [Finkelstein et al. 2012]. Since virtual peers existed before they were given
social competencies, “virtual peer” is used as the generic term, and “SIA as peer” is
used to refer to a virtual peer with social interaction skills. A similar trajectory was
followed by robots. Breazeal’s seminal book Designing Social Robots was published
in 2002 [Breazeal 2002]. In 2013, Kory et al. [2013] published the first paper on robot
peers. In what follows, then, we use the term “virtual peer” to refer to both physical
peer robots and peer agents on a screen (the virtual, in this context, refers to the
fact that it is not a flesh-and-blood peer) and “SIA as peer” to refer to both socially
interactive peer robots and socially interactive virtual peers.

In terms of the relative merits of the two kinds of embodiment—robot and
graphics—for peer agents, the jury is still out. A number of studies have compared
physical to graphical peer agents. However, in vast majority the studies have com-
pared a physical robot to an image of that same robot on the screen (e.g., Kennedy
et al. [2015]). This comparison does not do justice to the strengths of each kind
of embodiment—primarily, the physical presence of the robot and the more nat-
ural lifelike movement of the virtual agent—and so the comparison does not tell
us much about the subject of this chapter. For that reason, in the remainder of the
chapter we lay out studies on both virtual peers with a physical instantiation and
virtual peers on a screen in those places where each has made important advances
in the use of peer agents to support children’s learning and development.

While contemporary virtual peers are graphical virtual agents displayed on a
screen or physical robots, the very first virtual peer system was also one of the
earliest intelligent tutoring systems, the text-based Learning Companion System
described in the visionary 1988 article “Studying with the Prince: The computer
as learning companion” [Chan and Baskin 1988]. Here an artificial student interacted with the real student while both learned under the guidance of an intelligent tutoring system. By including two tasks, learning by being tutored and tutoring, this system was based on the effective “reciprocal teaching” paradigm [Palincsar and Brown 1984] described above in which children take both the teacher’s and learner’s role.

22.3.1 Roles for Virtual Peers

The earliest contemporary virtual peer played the role of a conversational partner in collaborative storytelling [Cassell et al. 2000]. Collaborative peer storytelling has been shown to have a positive impact on early reading and writing literacy [Teale and Sulzby 1986]. Based on a study of real children telling stories with one another, “Sam the CastleMate” was designed to be projected lifesize onto a screen behind a toy castle. A “magic tower” allowed Sam to seem to pass toys back and forth from the real to virtual world, and sensors in each room of the castle, as well as embedded in small figurines, allowed Sam to follow the child’s movements with its eyes, to give contextually appropriate feedback, and to tell stories that took place in the same room that the child had just played in. The graphics were intentionally cartoon-like in order to constrain the child’s expectations and to avoid any ambiguity about whether Sam was “real” or not (see section on ethics for further discussion). Most importantly, Sam was not photorealistic because the focus of the research was on the impact of the virtual peer on the child’s behavior rather than a focus on the most lifelike behavior possible for the virtual peer. In this context the evaluation focused on whether the system evoked natural social interaction behaviors in children and whether it improved their emergent literacy skills. Some children therefore were asked to play with Sam by themselves, and other children played with Sam and one other child, in a triad. These interactions were compared to children playing with another child or telling stories by themselves. Results demonstrated that children’s interaction with Sam was much like their interaction with other children. In fact, some children even coached Sam in how to tell stories, as in the case of one boy who told Sam “Try to make a longer story next time. It’s like this [Cassell 2004]. The little boy was outside….” The key question, however, is whether there was any benefit to Sam’s presence over and above what accrued to children playing with one another.

This did appear to be the case. The dyads of children playing without Sam sometimes told complete stories with decontextualized emergent literacy language. However, they also sometimes told stories that devolved into arguments or breaking parts of the castle. The stories that single children told with Sam, on the other
hand, were more likely to show the important decontextualized language that predicts later literacy, and the amount of decontextualized language increased with each subsequent story that they told in collaboration with Sam. Likewise, when dyads of children played with Sam, their stories also demonstrated more emergent linguistic behaviors than in Sam’s absence, and Sam’s presence also led them to engage in more pro-social collaboration of the kind that allowed them to get maximal educational benefit from the storytelling (op. cit.).

More recently, virtual peers have played a wide variety of roles in learning contexts. In a 2013 review article, Mubin et al. [2013] discuss the different roles that robots have taken in education and concludes that younger children are more likely to see the robot as a companion while older children see it more as a teaching aid. In a more recent review article, Belpaeme et al. [2018] add the teachable agent as a third useful role that can be taken by robots. In the teachable agent paradigm [Biswas et al. 2005; Chase et al. 2009], the student is always the tutor, teaching an agent that is described as younger and/or less knowledgeable (as the little boy described above did spontaneously when teaching Sam how to tell a long story). This perspective takes advantage of the “tutor effect” described above whereby students learn by teaching as well as when they are the student. Interesting work by Dillenbourg and colleagues has shown that even when children teach physically based skills, such as handwriting, a robot that plays the role of teachable agent, and that makes mistakes, can be very effective (inter alia Hood et al. [2015]). Research comparing the role of tutor, tutee, and peer found that children preferred peer robots [Looije et al. 2008]; however, explicit judgements of this sort do not always translate into performance, and so further research is required.

Research by Baylor and Kim [2005] showed the success of having a virtual peer take a variety of instructional roles. Research by Chen et al. [2020] demonstrated an active role-switching policy trained using reinforcement learning, in which the agent was rewarded for adapting its tutor or tutee behavior to the child’s knowledge mastery level. Results demonstrate that both tutor and tutee roles were important. The former had a greater effect on learning while the latter had a greater positive impact on the student’s affect. From research such as this, it is clear that the most effective virtual peers would take not just one role but be able to switch among them (including tutee, collaborator, and tutor) during a session. Similarly, it has been demonstrated that virtual peers can recognize simple roles in some natural narrative collaboration contexts and successfully elicit a shift in roles in the human partner, as well as shifting roles themselves. For example, in the context of children’s spontaneous collaborative storytelling, Wang and Cassell [2003] has shown that children reliably began collaborative stories with their peers by attributing roles to one another. For example, one child might say “OK, you be the princess
and I’ll be the dragon.” Another might describe the content of each role by saying “OK, the princess kills the dragon and saves the prince who was locked up in the castle.” Based on an automatic analysis of speech and nonverbal behavior, the virtual peers were able to take on the roles the child had attributed to them, and, in return, to attribute roles to the child. While this worked for simple role attribution in limited contexts, for the most part it is currently beyond the ability of most virtual peers to recognize the need for a new attribution of roles, or to fluidly switch roles.

At perhaps the most basic level of role that a virtual peer can take, a large number of studies have focused on virtual peers as ways to motivate learners to persevere. For example, virtual peers can successfully motivate an individual to stick to a task during a one-on-one tutoring session [Lane et al. 2013]. Virtual peers that demonstrate low competency can similarly raise self-efficacy. However, the same study showed that it was the high-competency virtual peer that increased learning gains [Kim and Baylor 2006]. As described above, we know that children also motivate one another to persevere in searching for solutions through debate and disagreement when collaborating in a small group [Sinha et al. 2017]. However, engaging in this kind of curiosity-inspiring conflict is still beyond the natural language capabilities of SIA systems, as described in Paranjape et al. [2018].

Nevertheless, virtual peers and other ECAs have been shown to be able to leverage group processes such as group trust, group emotion, conformity, norms, or cohesion, and to exert influence on groups (e.g., Sebo et al. [2020], Traeger et al. [2020]). Virtual peers can also influence the learner’s stance toward learning. This has been notably demonstrated in a study showing that children were able to recognize a growth mindset in a peer-like robot, and then themselves adopt such a mindset in their own approach to learning [Park et al. 2017].

### 22.3.2 Debate, Teasing, and Disagreement in Virtual Peers

Given the significant role played by debate and disagreement in children’s cognitive and social development, as well as in learning among peers, it seems natural to wonder how such phenomena might be incorporated into virtual peers. However, as the results described above demonstrate, teasing and insults can backfire and reduce learning gains when they are used in the wrong context—for example, among strangers rather than friends [Finkelstein 2017]. These results appear to hold for unfamiliar robots as well, as shown by Roth et al. [2019] where fairly mild robot insults resulted in reduced task scores for the people collaborating with the robot. For this reason, liberally peppering an interaction with insults is not going to help SIAs to act as learning companions. On the other hand, while insults play a negative role among strangers, by the same token they can play a positive role
among friends where they may serve to mark the relationship between the two interlocutors as a special one that exists outside of the bounds of everyday politeness [Zhao et al. 2014]. In the context of the current chapter, this means that debate and disagreement are more likely to have a positive impact once rapport has been established. Indeed, in a situation where the rapport between child and SIA has been rated as high by external annotators, teasing does seem to be appreciated by the child. In fact, it is sometimes the child that initiates the teasing. An example comes from Finkelstein’s work where 8- to 9-year-old children collaborated with a virtual peer on a science task [Finkelstein 2017]:

Child: What do think? Is [the bridge] going to be high or low?
SIA: Well, maybe we should make it lower so it has less room to wiggle around
Child: Ah! You took my idea, Alex! That was my idea because—
SIA: Nuh uh.
Child: Yes, it was!

In this corpus, both child and SIA were likely to initiate teasing and, as the example above demonstrates, the episodes were well received by the child as well as initiated by the child. However, in this experiment, interactions took place over several weeks, which gave time for the relationship between child and SIA to develop. In addition, the data was collected in Wizard of Oz mode, which would have allowed the experiment to cut short any teasing that seemed to be missing the mark and resulting in ill feelings.

Future research in this area clearly requires an adaptive model of how specific conversational strategies can be deployed, based on the stage of the relationship and/or the level of rapport and the user’s own prior conversational strategies. As of this writing, in unpublished research the adaptive conversational strategy model of Zhao et al. [2014], that includes teasing and other violations of social norms, has been implemented as a SIA as peer tutor of algebra, and one hopes that the results of this research will further illuminate whether and how putatively negative behavior such as teasing and disagreement might play a positive role in SIAs as peers.

22.3.3 Varying Speaking Styles in Virtual Peers

As described above, even quite young children are capable of adapting how they speak. In fact, by age 4 children can produce baby talk when interacting with an infant, even when they don’t have an infant sibling (Weeks [1971] cited in Labotka and Gelman [2020]). By early school age, children can simplify their speech when interacting with a foreigner [Labotka and Gelman 2020]. As described above, children growing up in communities where low prestige dialects are spoken, and where
these dialects are forbidden in school, can often fluently switch back and forth from the low-prestige dialect to school talk, even when their teacher considers them incapable of using the school-ratified dialect [Rader et al. 2011]. This code switching between dialects can serve as an important way of maintaining a link to two cultures—the culture of home, family, and tradition, and the culture of school, upward mobility, and mainstream societally ratified success. While some children are able and wish to navigate these different expectations and their associated ways of speaking, other children are either incapable of code-switching or find the two sets of expectations incompatible, as anthropologist Ogbu has discussed at length [Ogbu 1992, 2008].

Given the importance of dialect and other ways of speaking in establishing identity and building social affiliation with others, one might expect that the use of dialect by virtual peers and SIAs as peers would also play an important role in their interactions with users. Finkelstein et al. [2013] used a clever “distant peer” paradigm where 8- to 9-year-old children who spoke in vast majority African American English (AAVE), a low-prestige dialect used in many parts of the United States, were told that they were collaborating with a child from another classroom elsewhere in the city and that they would be communicating via recorded messages. In fact, the 2 four-minute voicemail messages from the other classroom (one social and one science) had been recorded by a bidialectal bicultural voice actress using one of three dialect patterns: one group heard AAVE in the social message and in the science message. A second heard AAVE for the getting to know one another recording and Mainstream American English (MAE) for the science recording. A third group heard MAE for both recordings. In order to test the impact of the social and science message on the children's dialect and on their production of science content, the children first heard the social message from the distant peer. They then recorded a social message in return. Then they recorded a science message. Then they listened to the distant peer’s science message, and finally they recorded their own second science message. The dependent variable was the difference between the number of scientifically valid arguments produced in the first science message (before hearing the virtual peer’s science message) and the second science message (produced after hearing the distant peer’s science message). While all children produced more science arguments after hearing the distant peer’s model, the authors found that students in the AAVE condition demonstrated greater gains in the use of this school-ratified science discourse than students in the MAE condition (with no difference for the code-switching condition from either of the other conditions).

A later experiment [Finkelstein 2017] replicated the results using actual virtual peers, both during a six-week experiment and in a one-shot experiment with a
larger number of participants. The virtual peers collaborated with children also 8 to 9 years old who also spoke primarily AAVE. Here the virtual peers looked identical across both conditions but spoke either MAE only (monodialectal) or spoke AAVE for brainstorming about a science task and MAE for practicing a presentation of their work for the teacher (the AAVE-only condition was omitted due to strong resistance from teachers). The use of two registers—brainstorming and speaking to a teacher—was introduced to justify the use of the different dialects. Here too the authors found that children in both conditions showed gains in the use of school-ratified science discourse (by which is meant both well-reasoned arguments and the language of science, such as referring to hypotheses and evidence) from pre-test to post-test). However, children who worked with the virtual peer that spoke their dialect demonstrated significantly greater gains in school-ratified science discourse from pre- to post-test than children in the mono-dialectal MAE condition, and the effect was strongest for children who were reading below grade level. The children in the code-switching condition also increased their participation in the science activity over the period of the experiment, unlike children in the MAE condition who visibly became increasingly aggressive with the agent and less participative over time.

Importantly, however, there was a strong mediating variable: children who collaborated with the bidialectal virtual peer demonstrated higher levels of rapport (as judged by external annotators), and it was the variable of rapport that predicted performance on the post-test [Finkelstein 2017]. This underscores the critical nature of social bonds in the success of virtual agents, as in the collaboration between human peers. It argues for social awareness as an essential part of virtual agents—a true argument for SIAs. Notably, not all studies with ECAs or robots have found the same results. Pazylbekov et al. [2019] found no increased learning gains for students working with a robot that spoke their own dialect of Kazakh. Similarly, a study on a pedagogical agent speaking high or low German also found no results on learning gains for the agent speaking the dialect of the student, although study participants gave higher likeability scores to the low German agent [Kühne et al. 2013]. Note that in these studies, however, the agent was not a peer but an authority figure, and the agents made no explicit attempts to knit social bonds, which may indicate that rapport is both more important and more easily built when virtual agents are peers than when they represent teachers or other authority figures. This is a key argument for SIAs as peers.

Nevertheless, integrating low-prestige dialects into SIAs presents its own challenges. In the longitudinal study reported above, children working with the agent that spoke their dialect produced more ratified science discourse, but they also produced increasing amounts of AAVE over time, a fact that may have caused
difficulties for them in their classrooms, where their teachers were explicitly negative about this low-prestige dialect. And unfortunately, integrating AAVE into virtual peers did not solve one fundamental issue concerning low prestige dialects, which is internalized biases against low-prestige dialect speakers. Famously, sociologist Basil Bernstein found that people who spoke low-prestige dialects in the UK tended to find others who spoke similarly to be more likeable, but less competent and less capable of earning high salaries [Bernstein 1961]. Similarly, Lugrin et al. [2020] found that adults judged robots speaking their own low-prestige dialect to be more likeable and less competent. In Finkelstein's studies, children who interacted with Alex over several weeks did change their explicit language ideologies. That is, they were more likely to say that it was okay to speak AAVE in some contexts. However, across conditions, and even for those students who had collaborated with a code-switching virtual peer for 6 weeks, all of the children continued to rate people who spoke AAVE as significantly less smart than people who spoke mainstream English. This persistent issue highlights the fact that while virtual peers may play an important role in children's learning, they are not silver bullets. Eradicating bias is a multipronged societal issue.

### 22.3.4 Difference in SIAs as Peers

Speaking different dialects and in different registers is one important way that SIAs as peers can support children experiencing difference in learning and development. Another is to be sensitive to the needs of non-neurotypical individuals. Here virtual peers have been successfully deployed to assess cognitive skills in children with ASD [Zhang et al. 2020a] as well as to improve confidence in social skills among adolescents with ASD [Boccanfuso et al. 2016]. Some studies have demonstrated that children with ASD can deploy social skills in their interactions with virtual peers that they do not deploy with real human peers [Tartaro and Cassell 2008], suggesting that these skills are in some sense known but not deployed in the young people’s everyday interactions. Based on this finding, some research has investigated whether it is possible to support adolescents with ASD in reflecting on social skills through programming virtual peers—systems that have been referred to as *authorable virtual peers* [Tartaro and Cassell 2006]. In a longitudinal study [Tartaro et al. 2015], teenagers with ASD were initially given a control panel that allowed them to choose behaviors for a virtual peer to perform in its interaction with another adolescent. Strikingly, the teenagers chose behaviors for the virtual peer to perform that they themselves did not use in interaction with their peers. Over a period of several weeks, the teenagers learned how to program novel behaviors for the virtual peer and even record language into the control panel for the virtual peer to utter. Once again, the teenagers developed social behaviors for the
virtual peer that they themselves did not use. Results of the study demonstrated that programming social behaviors and controlling those behaviors for a virtual peer had a transfer effect such that the teenagers subsequently were better able to deploy some of the social skills that they had programmed in their interactions with their real peers [Tartaro et al. 2015]. The topic of autism and socially interactive agents is further addressed in Chapter 25 on “Autism and Socially Interactive Agents” [Nadel et al. 2022] of this volume of this handbook.

22.3.5 Social Intelligence in Virtual Peers

One explanation for the success of peer-based learning comes from studies, some described above, that demonstrate that peers spend a fair amount of time managing social cohesion—the bonds that exist within social groups of various kinds. Even relatively young children can express their affiliation and bonds with others whom they resemble or wish to resemble, in quite sophisticated and effective ways [Kyratzis 2004]. Socially interactive peer agents can play similar roles. For example, a field trial of social peer robots in classrooms in Japan has shown that SIA robots can both detect friendship among the real children it interacts with and evoke friendship-like behaviors with children. This result was particularly the case for a sub-group of the children who specifically treated the robot as a peer (for example, asking it for advice about personal matters) but did not want to know how it functioned. This suggests that those children needed to suspend disbelief about the robot’s mechanical functioning in order to engage in friendship [Kanda et al. 2007].

In a closer examination of how rapport is built among human teenage reciprocal peer tutors working on linear algebra, Madaio et al. [2017] found that a number of different conversational strategies may play a role in raising levels of rapport. Some of the peer tutors, for example, couched negative feedback in indirect terms, and this indirectness negatively correlated with the level of rapport judged by external annotators, and positively correlated with the number of problems the peer tutees attempted, and the number they successfully completed. This use of indirectness was particularly the case for confident peer tutors, who may be better able to allocate some attention to the social context as well as to the tutoring content. In another study, teens working with SIAs as peers that manifested the same social interaction strategies as found in the human–human study demonstrated greater learning gains; however, this was only the case for the students who started with stronger knowledge of linear algebra. For those students whose prior knowledge was lower, a task only version of the agent was most successful in promoting learning gains [Guzman Garcia and Cassell unpublished]. This suggests that social interaction with the SIA as peer may require some cognitive effort and as such is
Research on Virtual Peers and SIAs as Peers

22.3 Research on Virtual Peers and SIAs as Peers

... easier to maintain for those students for whom the task is less effortful. Evidence for this interpretation also comes from a neuroscientific study that showed that while interacting with a human interlocutor engaged the parts of the neural apparatus responsible for social interaction, interacting with a virtual human in an identical task engaged both the social and cognitive effort parts of the neural apparatus [Gayda et al. 2008]. Of course, as Törning et al. [2020] explain, both developmental and neuroscientific evidence exists for the primacy of social stimuli. They link these results to the effectiveness of SIAs as peers whose nonverbal behaviors “prime a feeling of social partnership in the learner, which leads to deeper cognitive processing during learning, and results in a more meaningful learning outcome as reflected in transfer test performance” (Mayer and DaPra [2012], as quoted in Törning et al. [2020]). Given the results above, however, it is clear that this social partnership may not work for every learner or perhaps only for particularly well-designed SIAs. In any case, it is clear that SIAs as peers can profitably use language and nonverbal behavior to create social cohesion in the service of task performance, as has been shown in a number of studies that have looked at the impact of social behaviors in virtual peers on learning. Baxter et al. [2017], for example, assessed the effects of a personalized robot on learning in the classroom, assessing performance on both a novel and familiar task. What is meant by personalized in their context is some overall nonverbal convergence to the child’s movements, personability (friendliness, informal language, and lack of imperatives), and adaptation to task, such that the personalized peer robot allowed the child to repeat a task. While the ability to repeat a task might be considered as a confound, as it may derive more from good teaching technique than personalization, results did show a learning effect such that the personalized robot resulted in more learning on the novel topic (although not on the familiar topic). One important issue concerning social intelligence in virtual peers is how to maintain interest over the time it takes to build rapport, as virtual peers can be repetitive in their language and nonverbal behavior. Burger et al. [2017] implemented a module that was capable of engaging in mutual self-disclosure with children over a two-week period. A larger number of self-disclosures on the child’s part was associated with higher rapport, and higher rapport led to more use of the application over time (see Chapter 12 on “Rapport Between Humans and Socially Interactive Agents” [Gratch and Lucas 2021] of volume 1 of this handbook [Lugrin et al. 2021]).

Techniques such as these will be important for future work with SIAs as virtual peers. In this context, encouraging evidence about the feasibility and impact on learning of integrating rapport building behaviors into SIAs comes from a parallel research tradition in parasocial relationships between children and characters that are familiar to them from mainstream media such as movies, computer games, and television shows. American television shows such as Blues Clues and Dora the
Explorer have developed ways to deploy socially contingent parasocial interactions, defined as techniques whereby characters are programmed to create pseudo conversations with children through comments, questions, and well-placed pauses [Lauricella et al. 2011]. An increasing number of studies have shown the effectiveness of this approach, both for children’s feelings of friendship with the character and for their learning in interactive computer games. In one study, for example, Calvert et al. [2020], who have spearheaded much of this work, have shown that the strength of young children’s parasocial relationship with Dora the Explorer predicted their learning gains on an early math skills task, and that this learning transferred to other tasks. Similar results come from the kinds of child characters being integrated into interactive videos such as those derived from the American TV show Elinor Wonders Why. While social bonds have not been assessed, recent work by Xu et al. [2022] shows that children respond significantly more accurately to science questions posed right after watching the interactive version of the video than after watching the broadcast version without interaction between the child character (a curious bunny named Elinor) and the child viewer. Even more strikingly, engagement is higher for the interactive child character than for a parent when engaging in dialogic reading (engaging the child in dialogue about the book being read), and the child’s story comprehension is also higher in the interactive child character condition [Xu et al. 2021]. In all these cases, the studies rely on the power of learning with an interactive peer.

22.4 Models and Modeling

As we move toward the conclusion of this chapter, it is important to note that, in addition to being effective interventions, virtual peers and SIAs as virtual peers can also serve another important role, and that is as models of human behavior, in the sense of McClelland’s explorations of “ideas about the nature of cognitive processes” [McClelland 2009]. In this context, researchers can create different versions of a virtual peer, and observe which look natural and which seem unnatural. More helpfully still, researchers can relatively quickly change the virtual peer’s behaviors, have the virtual peers interact with children, and thereby assess—for example—what kinds of disagreements are productive and what kinds obstruct learning rather than supporting it. In addition, virtual peers can elicit peer-like behaviors in children when experiments among real peers are difficult to carry out. This is the case, for example, with the experiments on the role of low-prestige dialects in learning described above. It is extremely difficult, and perhaps impossible, to find a context in which every variable is kept constant except the dialect that one child speaks with another, or to find a child who varies the dialect in one
task and not another. Issues of socio-economic status, ethnicity, low-versus high-
resourced schools, and several other variables that co-vary have rendered difficult
important research on the role of low-prestige dialects in learning, and their impact
on children’s language ideologies. Virtual peers have allowed this research, as
described above. Similarly, experiments with SIAs as peer robots interacting with
pairs of students were able to discover instances of engagement that were produc-
tive and instances of engagement—often held up as the holy grail of interactive
learning environments—that were unproductive [Nasir et al. 2021]. The results of
studies such as these, that identify productive sequences for all students, or for
particular populations, in particular contexts, can be integrated into virtual peers
that then are optimally helpful to their human partners. This double role, of model
and intervention, underlies many virtual peers—and ECAs before them.

As well as allowing researchers to tweak behavior and observe whether it is nat-
ural or not, and whether it is helpful or not, virtual peers can also help researchers
better understand the very nature of dyadic interaction among peers. Here, in
some sense virtual peers are acting as simulations of theory of mind [Decety and
Grèzes 2006], both for the researcher and for the child interacting with the virtual
peer. This is the case, for example, for children with ASD. As we described above,
research shows that these children appear to find it easier to interact with virtual
peers than with real peers [Tartaro and Cassell 2006, 2008]. However, when they are
given the opportunity to program social behaviors into authorable virtual peers,
they can then subsequently use some of the behaviors they have programmed in
subsequent face-to-face interaction with their real peers [Tartaro et al. 2015]. This
suggests that the virtual peer is a kind of “practice other”—an interlocutor who is
easier to engage with than real peers, whose perspective is perhaps easier to assess
(in the sense of a theory of mind), and who serves as a steppingstone to interaction
with real peers.

Yet another role for virtual peers-as-model is to highlight places where we do not
yet have adequate computational models of children’s language, particularly chil-
dren’s language when speaking to their peers. And yet such models are required to
build effective autonomous agents that truly speak like peers to their child users
(we might envisage a future large language model called “Bertie” for example).
Such computational models and associated corpora would also allow us to analyze
peer communication more effectively. Failures in implementation are not usu-
ally published; however, one informative example comes from the work described
above concerning the role of conflict in raising the level of curiosity in elemen-
tary school children. A series of papers published on the dynamics of curiosity in
group learning (see Sinha et al. [2022] for the most complete discussion) describes
the ultimate goal of implementing detailed models of curiosity and embedding
them in a virtual peer or several virtual peers capable of engaging in conflict and thereby inspiring a rise in curiosity. However, attempts to build such systems (see Paranjape et al. [2018] for first steps) ran into issues due to the impossibility at the time of building a deep learning model of children’s peer talk that would automatically detect productive conflictual talk in children’s multiparty conversation, and autonomously generate utterances that would inspire such productive conflict.

22.5 Future Work: User Modeling and Conversational Strategies

Developing computational models of how conflict is displayed through verbal and nonverbal means, and how it can be evoked in children, remains an important goal for future work. However, the challenge of modifying automatically generated language to adapt to different populations—for example low-literacy adults (cf. Martin et al. [2020])—is undergoing active research in the NLP community. One might imagine that such work would be applicable to generating child-like language. There is also recent work in open-domain chat-oriented systems, including using large-scale language models such as DialoGPT [Zhang et al. 2020b] and BlenderBot [Roller et al. 2021] to generate grammatical and locally relevant text in conversation; however, these systems are not yet able to generate appropriate text within a certain context beyond one or two turns. Those that learn in real time may also integrate biased or socially inappropriate user input for training data, and thus generate unsuitable language [Zhou et al. 2020]. However, recent work has made strides in analyzing and generating conversational strategies of the kind that build rapport [Soni et al. 2021, Raphalen et al. 2022]. This latter work, too, will undoubtedly play a role in generating child-like language that builds a social bond with a real human peer.

In addition to the challenges highlighted in the sections above, user modeling remains an important challenge for SIAs as peers. User modeling has played a key role in tutoring systems, from their earliest days (cf. Sleeman and Brown [1982]). In user modeling in the tutoring context, the system keeps track of what a student knows and does not know, adapting the model as new knowledge is presented, and updating it when assessment shows that the student has acquired knowledge (or forgotten it). Some user models for tutoring systems differentiate between procedural knowledge (roughly, how to address a topic) and conceptual knowledge (roughly, a deeper knowledge of the domain, which allows the student to generalize) [Murray 1999, Rau et al. 2009]. In current user models for other domains (such as recommendation systems), user preferences are also recorded as well as, for some systems, an assessment of personality or other personal and interpersonal features of the user. Establishing and maintaining a user model for a SIA
is particularly important as rapport and other social constructs change over time. And yet, user models for systems where the computer plays a peer are particularly difficult to design as the model must take into account the preferences and abilities of a student of a certain age and not those of the designer of the system. This may include vocabulary level, level of theory of mind, cultural references, size of gesture space (larger in younger children than in older), and other features of a student of a particular age group. Modeling a peer is important because prior research suggests peer discussions are successful in part because the discussion helps identify misunderstandings while still “speaking the students’ language” [Blum-Kulka and Dvir-Gvirsman 2010].

22.6 Ethics of Virtual Peers

Systemic racism and discrimination are embedded in our educational systems, including on the part of students toward their peers. For this reason, virtual peers must carefully consider notions of equity and inclusion from the moment of their conception [Perry and Lee 2019]. A powerful example of this concern comes from Finkelstein’s work on low prestige-dialects, described above. In a number of cases, the researchers were banned from the classroom for “advocating poor English.” The students themselves learned more when brainstorming with virtual peers that spoke as they did, but they still demonstrated internalized racism with respect to those agents. That is, while the children learned more with the agents who spoke as they did, when asked if those agents were smart, they replied that they were not, more than once specifying “because they speak ghetto” [Finkelstein 2017]. This reminds us that we must carefully navigate the ways in which AI systems can propagate bias and exclusion and must include the goal of reducing bias and increasing representation as part of the design criteria. Researchers must also pay attention to the composition of the datasets they use as training data. They must attend carefully to the appearance, voice, and behaviors designed for the agents to remove any unintended bias. For example, early pedagogical agents included negative stereotypical gendered characteristics, such as low-cut blouses and tiny waists for female agents, and negative stereotypical ethnic characteristics such as gold chains and backwards baseball hats for African American agents. Some research has turned to gender- and ethnicity-ambiguity to reduce the risk of reifying such stereotypes [Rader et al. 2011].

In addition to issues of bias, there are more general ethical issues to address. Voice assistants such as Alexa, Cortana, and Google Assistant have evoked a fair amount of fear among parents and teachers, and this fear also affects adult perceptions of SIAs as peers. Perhaps the most common of such responses by parents
is the fear that children will come to believe that they need not be any more polite with real people than they are with Alexa. Other worries include that children will no longer be able to distinguish between real and virtual playmates and will come to prefer the virtual versions, always accessible and always up for a game. These worries have been extensively covered by the press (e.g., Gonzalez [2018]), and even formed the subject of reports by governmental commissions (e.g., CNPEN [2021]). While there do not currently exist significant data to support these reservations, it is clear that general guidelines for the implementation of SIAs as peers should be followed so as to ensure the emotional and physical safety of vulnerable users, particularly young children and children with special needs. For example, the CNPEN report suggests that SIAs as peers be clear about the fact that they are not real, that they not be photorealistic in appearance, and that it be possible to find out what data they are collecting and storing about child users. In some sense, these guidelines should hold for all AI systems, but when the users are children, they are perhaps particularly important to keep in mind. In addition, while the press has covered adult fears about SIAs as peers, it is just as important to understand children’s own reactions and worries [Yip et al. 2019]. Finally, while guidelines for ethical use are important, it is also useful to remember that fears such as those described above have in fact accompanied the introduction of every new talking technology, from the radio to the television to videogames and today to virtual peers [Cassell 2020]. We might therefore wish to take a cue from a 1961 booklet on the dangers of television published by the US government and illustrated by famous cartoonist Walt Kelly:

there are few things to practice not doing. Don’t be afraid of it. These things are probably here to stay. Don’t be afraid of your child. He’s not here to stay. He’s a precious visitor. Do not wind your child up and set him to play with it unguided. Do not wind it up and set it to watch your child. A machine is a bad sole companion. It needs help. You can help it. Love your child. [Kelly 1961].

22.7 Conclusions
The development of virtual peers has built on work in intelligent tutoring systems and ECAs but goes beyond them. The 2020 (and beyond) pandemic has demonstrated the need for students to continue learning even when separated by distance and time from their peers and teachers. Like intelligent tutoring systems, virtual peers can be tuned to a student’s capabilities and are available on demand at any time or place. In addition, virtual peers can take advantage of the ways in which the presence of peers potentiates learning and development, in terms of the
productive cognitive conflict they can generate, the evocation of self-explanation, and the social bonds that underly much of learning and development. SIs as peers, whether graphical agents or physical robots, are particularly well-placed to build learning-focused social bonds over time and to allow students to continue to make gains in learning as well as in the socio-emotional skills that will allow them to lead productive lives.

References


American Psychiatric Association. 2013. Diagnostic and Statistical Manual of Mental Disorders (5th ed.)


Chapter 22  Socially Interactive Agents as Peers


Chapter 22  Socially Interactive Agents as Peers


References


References


A. M. Perry and N. Lee. September. 2019. AI is coming to schools, and if we’re not careful, so will its biases. The Avenue. https://www.brookings.edu/blog/the-avenue/2019/09/26/ai-is-coming-to-schools-and-if-were-not-careful-so-will-its-biases/.


Socially Interactive Agents for Supporting Aging

Moojan Ghafurian, John Muñoz, Jennifer Boger, Jesse Hoey, and Kerstin Dautenhahn

23.1 Motivation
The average age of the world's population is steadily increasing. There are 125 million people globally aged 80 years or older and it is predicted that this number will change to 434 million by 2050. While many older adults are capable of living independently, there are many others who need care. For example, 50 million people are currently living with dementia (predicted to reach 152 million in 2050), which is reported as a leading cause of disability and dependency among older adults [WHO 2019].

While the increasing number of older adults is resulting in an increase in the demand for caregivers, the demographic change that is occurring worldwide is limiting available resources. By 2045, older adults are predicted to outnumber youth (i.e., population <18 versus >65 years of age) for the first time in history [Vespa et al. 2018]. Having an aging population has profound implications for the social, cultural, and economic systems that are needed to support healthy and holistic aging. For instance, WHO [Harris 2019] estimated that between 2006 and 2015 expenses have been around USD 84 billion to treat chronic diseases such as heart disease, stroke, and diabetes, conditions that are more prevalent in older adults. Thirteen million new cancer cases in 2009 (largely because of global aging) had an associated treatment cost of at least USD 286 billion; this number is expected to increase to 27 million cases in 2030 [World Health 2018]. While the costs associated with maintaining wellbeing and quality of life of older adults are difficult to assess, it is clear that technological solutions for supporting aging can help with reducing the
associated costs, such as enhancing the wellbeing of the world’s growing number of older adults and decreasing the workload on caregivers [Kachouie et al. 2014].

While there is much focus on changes in health and the associated costs that often accompany aging, engaging in meaningful activities is a key component to a good quality of life. Older adults are valuable members of society; they have a high rate of volunteerism and several studies have shown intergenerational interactions with older adults lead to positive outcomes for younger adults, such as increased self-esteem, acquiring new skills, and decrease in anxiety [Springate et al. 2008]. Supporting older adults’ ability to engage in the activities that provide fulfillment, enjoyment, and personhood is as important as supporting their health.

The increasing demands of our aging population is creating gaps in support for health and wellbeing, some of which can be filled by technological solutions, particularly those that disproportionately affect older adults. For example, while COVID-19 led to isolation among both younger and older adults, the impact was higher for older adults for many reasons, such as the greater risk of catching the virus in long term care homes, greater average fragility of older adults (therefore more susceptible to poor outcomes), and relatively less access to communication and connectivity technology. Therefore, this situation led to increased social isolation, including among older adults who were not previously socially isolated.

Socially assistive agents have the potential to help fill gaps caused by the increasing demand for support, as well as situations that affect older adults’ mental health. Socially assistive agents are computer-driven technological entities that are able to interact with a person in a socially engaging manner [Hegel et al. 2009]. Increasing technological advances are enabling assistive agents to have a positive impact on several aging-related challenges that can significantly affect people’s independence and their quality of life [Matarić and Scassellati 2016], such as physical, cognitive, emotional, and social challenges [Wang et al. 2014, Saez-Pons et al. 2015]. While assistive agents are not a replacement for human companionship, there are several ways technology can be used to support older adults and promote greater feelings of engagement, independence, and inclusion: for example, by increasing older adults’ independence (e.g., increase healthcare support or support remote contact with their families to reduce isolation) [Moyle et al. 2018], acting as companions to reduce loneliness [Banks et al. 2008], encouraging and increasing engagement of older adults with others [Šabanović et al. 2013] as well as in activities [Khosla et al. 2012, Abdollahi et al. 2017], and supporting mental health [Shibata and Wada 2011].

Despite these positive outcomes, society’s attitude toward social robots is not always positive and people’s attitude toward social robots has been reported to have become more negative over the years [Gnambs and Appel 2019]. This may be
in part because many people worry social robots are poised to replace human contact, rather than viewing them as complementing existing resources and enriching experiences. A study conducted during the social isolation period of COVID-19 showed that those who reported any change in their lives due to COVID-19 (either positive or negative) had a positive perception change about the advantages of the social robots [Ghafurian et al. 2021b]. This could be due to the fact that COVID-19 represented a situation in which in-person interactions were not possible and emphasized the potential benefits of social robots. Furthermore, the COVID-19 situation, while unique in nature, could be representative of many other circumstances that lead to social isolation among older adults, such as physical disabilities that may limit older adults’ interactions outside their homes and family members that live far away.

The aim of this chapter is to: (a) provide an overview of the state of the art of assistive agents designed with the goal of supporting aging and improving older adults’ quality of life, (b) present existing methods and approaches for design, development, and evaluation of assistive agents, including important considerations for implementing them, and (c) discuss existing challenges and provide directions for future work in designing assistive agents for aging.

We start with a few definitions of terms used in this chapter. First, we define what we mean by “older adults”. While WHO [2020a] uses “older adults” to refer to adults over the age of 60, it is defined differently between countries (e.g., 65+ in North America) and organizations. In this chapter, we do not refer to any specific age range; our focus is rather on a broad range of people with different needs and abilities who have experienced aging, ranging from healthy older adults to those with multiple morbidities.

In terms of “assistive technologies”, WHO [2020b] has defined guidelines and definitions to establish whether or not a system/device can be considered as assistive. Specifically, assistive technologies are defined by the Individuals with Disabilities Education Act (IDEA) as “Any item, piece of equipment or product system... that is used to increase, maintain or improve the functional capabilities of individuals with disabilities”. The aim of assistive technologies is to help people to reach their goals such as improving independence, facilitating social participation by fostering inclusion, mitigating impairments and other health-related conditions, and increasing quality of life.

Furthermore, “assistive agents” are entities that are situated to assist a person in any task or activity. Raïevsky and Michaud [2009] defined a situated agent as a “Physical or virtual entity which is situated in a dynamic, quasi-continuous environment, which it perceives through sensors and into which it operates autonomously.” The type of assistance can be different depending on the context,
for example, it can vary from different levels of acting upon the environment or providing information [Breazeal et al. 2016]. Here, we define assistive agents accordingly, as physical or virtual entities that are situated in dynamic environments with the goal of assisting a person.

While there are several applications that assistive agents can help older adults with, a variety of factors can affect the success of these agents. This chapter discusses such factors and emphasizes design approaches that can minimize challenges related to their design and implementation while maximizing the likelihood of adoption of assistive agents by their users.

### 23.2 History and Overview

Older adults’ wellbeing can be affected by many factors such as their health, mental state, and ability to live independently [Boger 2014]. Therefore, assistive agents can be designed to assist older adults in many different contexts and in a variety of tasks. For example, a recent review of the existing assistive robots for supporting older adults with dementia has identified five categories for assistive robots’ applications: (1) companionship, (2) engagement, (3) health, (4) therapy, and (5) activity of daily living (ADL) [Ghafurian et al. 2021c]. The reader can also see Shishehgar et al. [2018] for a proposed categorization in a more general context of robotic technologies for older adults.

In this chapter, we will discuss the applications of assistive agents for supporting aging in six different categories, as shown in Figure 23.1. The first category, companionship, are agents that are designed with the goal of providing companionship and in most cases with the goal of reducing social isolation and loneliness. These agents could help with a variety of tasks such as talking with older adults, playing music or videos, showing pictures, or reminding people about their tasks or calendar events. COVID-19 has emphasized the potential role of intelligent agents and social robots that are designed to provide companionship during a period of self-isolation and social distancing where human contact may be hard or impossible [Ghafurian et al. 2021b].

![Figure 23.1](image-url) Categories of use for assistive agents.
The second category, engagement, represents assistive agents that are designed to increase engagement of older adults with technology, in activities, or with others such as therapists, family, and friends. The third category, health, represents agents that directly promote health such as agents that motivate exercise, a healthy living style, and suggest specific diets. The next category, leisure, represents agents that provide older adults with a type of entertainment, such as playing games.

While agents in the above four categories can be adopted by most older adults, the agents in the last two categories, therapy and ADL support, are those that support people with specific needs. Therapy represents agents that provide a type of mental or physical therapy (e.g., pet or music therapy), and agents in the ADL support category are those that are designed to assist an older adult with an activity of daily living, either by providing cues about the next steps of an activity (e.g., washing hands) or by physically assisting with that activity.

Other categorizations have been used to explain the role and nature of agents. For instance, agents can be classified by considering their nature: physically embodied agents, virtual agents, and voice assistants. Physically embodied agents (robots) can be distinguished from virtual agents because they have a physical embodiment (i.e., hardware-based agent). A virtual agent is defined as a graphic entity that simulates behaviors (human-like or not) and actions in the real world [Calvo et al. 2015]. Virtual agents are usually on a screen of some type.

Voice assistants (also known as smart speakers or voice-controlled devices) have been gaining attention during the last decade [Stigall et al. 2019]. Voice assistants (1) interact primarily through voice commands (input), (2) are always connected to the Internet, and (3) communicate with the user through audio (output). Voice assistants differ from robots as they are equipped with limited hardware capabilities (e.g., stationary location, sensing capabilities limited to record voice, battery life). However, voice assistants can be interfaced with robots and smart devices for home automation systems to extend their functionalities. Although research in this field is still in relatively early stages [Pradhan et al. 2018], there is clear potential for these devices to provide assistance in tasks such as making remote calls, providing medication reminders, and controlling a smart home.

Assistive agents and robots have been developed to support multiple contexts. For example, animal-like robots have been successfully used to provide pet therapy [Shibata 2012]. One of the most well-known animal-like robots that has been used successfully in the older adult context is the PARO robot, which is a seal-like robot that has been successfully adopted by many older adults with dementia. Human-like robots have been used in other contexts, such as providing music therapy or language-related therapies [Martin et al. 2013].
Agents for promoting healthy living have been used for a variety of health-related tasks ranging from motivating healthy living habits (e.g., to walk and exercise) [Khosla et al. 2014] to providing advice about diet [Khosla et al. 2012]. However, using this category of agents to support aging has gained relatively limited attention. This may be because of the challenges in designing effective agents in this context.

Agents have also been proposed to assist with specific activities, such as a hand-washing system that assists people with dementia in washing their hands by providing audio/visual support about the next step in the task [Malhotra et al. 2015] and a mealtime assistant that prompts people with dementia to eat their food and helps them with the process of eating a meal [Derek et al. 2012]. However, designing agents that successfully help with performing activities of daily living is very challenging as the agents need to perceive the user’s actions and its environment accurately to provide timely and contextually appropriate support.

Agents (and in particular social robots) have been successfully used to increase social engagement of older adults [Šabanović et al. 2013, Perugia et al. 2017], and to provide companionship [Odetti et al. 2007, Mannion et al. 2019]. Companionship agents have been evaluated with older adults with different conditions, such as cognitive disorders, dementia, stroke, and depression, and in a variety of different settings, such as long-term care homes, day-care centers, and individuals’ homes. Most of the agents have been successful in supporting aging and have shown positive improvements in older adults’ quality of life in many domains. For example, assistive robots used in care centers not only increased older adults’ engagement in activities [Rouaix et al. 2017] but also improved their engagement with other residents [Šabanović et al. 2013] and reduced depression [Shibata 2012]. In the context of therapy, both short- and long-term effects of social robots have been investigated; however, in many contexts, the long-term effects of these agents are unknown due to the multiple challenges involved in conducting long-term studies. Some of these challenges are further discussed in Section 23.5.

Another field that has contributed to the development of assistive agents for supporting aging is serious gaming. In this area, games are designed with objectives that go beyond entertaining players, such as promoting healthcare, education, and training [Michael and Chen 2005]. Serious games that support aging have shown measurable benefits in physical fitness [Kappen et al. 2019], rehabilitation [Proença et al. 2018], and cognitive functions such as memory [Garcia-Betances et al. 2015], spatial orientation [Gamito et al. 2017], executive functions [Nouchi et al. 2012], and slowing down cognitive decline [Lau et al. 2017]. While researchers have shown the positive potential of using games for specific physical and cognitive interventions, more research and evidence is needed to allow for
(1) establishing strong theoretical foundations, (2) better design experimental protocols, and (3) greater focus on user experience (rather than system usability evaluations) to better understand motivation, engagement, and long-term adoption [Zhang and Kaufman 2016]. Furthermore, research in the field of serious games to support aging has primarily focused on healthcare support such as therapeutic rehabilitation and cognitive training. As games have demonstrated their potential in alleviating isolation and supporting social wellbeing [Li et al. 2018], these are areas relevant to aging that warrant more attention. Beyond therapeutic uses, serious games could play a decisive role in connecting older adults with their family beyond traditional video-calls and messaging applications; games can be used as the medium to create meaningful connections among grandchildren and grandparents [Boger and Mercer 2017]. Further, for the purpose of entertainment, gaming profiles tailored for older adults have been created, revealing the importance of specific game design aspects such as aesthetics (e.g., nostalgia, contemporaneity) and mechanics (e.g., musical play, autobiographical) to foster meaningful game play [De Schutter 2017].

To conclude, while social robots have been shown to be a suitable solution in many application areas, technology adoption and long-term use are still relatively low due to the challenges involved in running longitudinal studies in this context.

23.3 Models and Approaches

It is important to involve older adults when designing assistive agents if they are to successfully support ageing. In other words, as with all intended user groups, it is important to design with them instead of for them [Lazar et al. 2018]. Other than ascribing agency to users [Tholander et al. 2012] and empowerment [Galliers et al. 2012], this means accessing and complementing their needs, values, and abilities, which can be accomplished through user-centered design methods.

User-centered design methods, such as participatory design and co-design [Sanders and Stappers 2008], provide researchers and developers with rich insights that enable them to create a successful product [Zimmerman et al. 2007], based on real user needs instead of designers’ assumptions. They can also ascribe agency [Tholander et al. 2012] and empowerment [Galliers et al. 2012] to users by creating products that clearly reflect them. Participatory design is when relevant stakeholders are involved in the design process to get a feel for and incorporate the user perspective. Participatory design is becoming more commonly used in research in general; indeed, it is becoming a requirement by several funding agencies for projects involving human participants. Co-designing is a more intensive level of participatory design that implies a very committed and early involvement of the targeted population to generate an empathetic perspective that will facilitate the
understanding of needs, motivations, preferences, attitudes, and limitations in designing agents for supporting aging [Smarr et al. 2014].

As with other contexts, involving older adults in the design process is an important aspect of developing assistive agents that can successfully help older adults. User-centered design and participatory design have been gaining more and more attention in designing technologies with this audience [Duque et al. 2019] as they can provide more holistic and direct insights into older adults’ needs and motivators to use and adopt assistive agents.

The convergence of an amalgam of these techniques and approaches is a clear response to a design manifesto suggested by Donald Norman [2013] arguing for a more integrative and diversified process for product design. Therefore, the classical, structured, and sequential technology-driven design–prototype–test cycle applied in product design is hardly suitable for the creation of novel assistive agents targeting older adults [Daly Lynn et al. 2019]. For example, many of the common assumptions about technology usages may not be applicable for older adults due to a difference in older adults’ mental model of agents and their limited previous exposure to such agents.

Developing assistive agents to support aging is and should be highly multidisciplinary. The design process itself requires the use of several techniques and approaches from areas such as product design, human–computer/robot interaction, and system design. The preliminary research, requirement elicitation, ideation, design, testing, and validation stages should normally consider a variety of techniques and tools to create suitable solutions that can properly accommodate older adults’ needs. This includes the consideration of those who may be providing formal (i.e., clinical) and unpaid (i.e., family and friends) care for them, if appropriate.

Well-established techniques for user-centered design include shadowing and observational processes, emotion assessment, self-reporting tools (e.g., diary studies, focus groups, surveys, and interviews), user modeling approaches (e.g., user personas, user journey maps, scenarios, card sorting), and designer analysis (affinity diagrams, use cases, user matrices) [Still and Crane 2017]. Despite being time-demanding, the co-design process enables researchers to have a more solid end user model as well as valuable information about how to design assistive agents in a way that meets the expectations, needs, and preferences of older adults [Muñoz et al. 2019]. This may in turn save time later with more appropriate, targeted, and effective designs.

Table 23.1 shows a summary of the common methods widely used to design socially assistive agents. Five methods have been highlighted in the literature: (1) **focus groups**, which help researchers to learn more about older adults’ concerns,
Table 23.1  Common methods for designing assistive agents; advantages, challenges, and considerations are based on Dix et al. [2003], Drachen et al. [2018], Hubbard et al. [2003], and Suryani [2013].

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Challenges</th>
<th>Considerations for Older adults</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus Group</td>
<td>Helps generate new ideas by encouraging discussions among participants</td>
<td>Getting accurate information about the participants that is not affected by the others</td>
<td>Avoid long sessions and give breaks as much as needed. If there are participants with impairments, make sure they hear and understand each other. Show greater flexibility for time and location.</td>
<td>Pino et al. [2015] used focus groups to understand preferences about a social robot.</td>
</tr>
<tr>
<td>Interview</td>
<td>In depth understanding of people's experiences, interpretations, and opinions (why and how) More flexibility in choosing the questions Hard to quantify the responses Can be used with those who have difficulty reading or writing</td>
<td>Does not quantify Responses might be biased and affected by the interviewer Harder to de-identify participants Harder to recruit participants compared to surveys</td>
<td>Be polite and patient with the responses. Show greater flexibility for time and location Keep it short or allow breaks (use facial expression and body posture as clue for giving breaks)</td>
<td>König et al. [2016] conducted interviews with care-home residents and family caregivers to understand how to improve an assistive agent's prompts and its acceptability.</td>
</tr>
<tr>
<td>Method</td>
<td>Advantages</td>
<td>Challenges</td>
<td>Considerations for Older adults</td>
<td>Example</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>Can use standardized structures to aid interpretation of results and statistical processes. Easier to recruit participants compared to interviews, verbal/telephone surveys, etc.</td>
<td>Hard to understand participants' feelings and emotions. Higher probability of getting dishonest answers</td>
<td>Questions should be easy to read, and the font should be reasonably large</td>
<td>Leuty et al. [2013] used questionnaires to evaluate a computer-based intelligent device from the perspectives of older adults and therapists</td>
</tr>
<tr>
<td>Ethnographic Field Study</td>
<td>Provides a rich qualitative observational insights into users' behaviors and the rationale for their actions.</td>
<td>Cannot be generalized and cannot quantify aspects of behaviors and attitudes</td>
<td>Be sensitive to older adults' routines and do not disturb it</td>
<td>Forlizzi et al. [2004] conducted an ethnographic study of older adults living independently in their homes to understand the role of assistive robots in living independently longer</td>
</tr>
</tbody>
</table>
Table 23.1 (Continued)

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Challenges</th>
<th>Considerations for Older adults</th>
<th>Example</th>
</tr>
</thead>
</table>
| Case Study | Provides more details about a phenomenon  
It allows understanding of social situations | Data reported might be biased depending on the researcher  
Results cannot be generalized  
Mostly rely on subjective data  
Harder to de-identify data | Be sensitive to older adults' routines and do not disturb it | Walsh and Callan [2011] conducted case studies (along with focus groups and interviews) to understand older adults' preferences and acceptance of information and communication technologies |
needs, and perspectives as well as come up with new ideas by generating discussion among the participants; (2) interviews, which help with understanding people’s individual experiences, concerns, and perspectives; (3) questionnaires, which provide researchers with a structured way of exploring questions by reaching out to a larger number of participants that can be from a diversity of geographic regions; (4) ethnographic field studies, which allow researchers to gain rich insight on participants’ behavior and the reasons behind their actions (e.g., interactions with the agents); and (5) case studies, which enable researchers to obtain detailed information that may not be possible to obtain through other methods.

There are different advantages, challenges, and considerations for using these methods with older adults, and examples of research that used each method are described in Table 23.1. For instance, when running focus groups, the researcher should consider important accessibility aspects of older adults, such as ensuring that participants can hear each other clearly and correctly, as well as ensuring that they are physically and mentally comfortable when sharing their views. Interviews, on the other hand, can be carried out remotely and offer greater flexibility in time and location when compared with the focus group but are more time intensive and cannot support group discussion. Surveys have good opportunities for wide outreach but may not support clarification on questions or answers that may be confusing. For all of the techniques, the researchers need to be realistic and sensitive to older adults’ preferences regarding the length of the study, the wording used, and any technologies that are employed; this includes considering when to run the study so as to align them with older adults’ preferences.

One important decision in the design process is the choice of the interaction methods and modalities. Examples of interaction types are speech, tactile, non-verbal and verbal cues, and emotions. In general, choosing the right interaction method(s) for any intelligent system is crucial and is decided according to its applications. The choice of interaction methods becomes even more critical in technologies aimed at older adults due to possible changes in abilities caused by aging, for example, changes in hearing, vision, and/or memory function. Therefore, it is important to (a) include multiple interaction approaches as opposed to relying on one, as some interaction approaches might work for some individuals but not others; (b) implement the interaction approaches in a robust way as failures could lead to frustration and abandonment of the agent; and (c) carefully match the interaction modality to the intended goal; in other words, the technology must align to the task people want to do in a way that makes sense to them rather than employing a technology simply because it is new or at hand.

While the performance of assistive agents can significantly affect their usability, reducing errors is challenging when it comes to designing assistive agents.
There are many technical aspects that can affect the performance of an agent, and the abilities of the people who use them and the contexts in which they are deployed are broad. To appropriately assist older adults, an agent should be capable of understanding its environment and users (see Amirabdollahian et al. [2013], Fong et al. [2003] for examples). That is to say, depending on the application, multiple aspects of the agent need to be properly implemented. For example, if the agent needs to detect objects in the environment or track people, techniques in computer vision should be used in a way that errors are minimized under different environmental conditions (e.g., lighting). Natural Language Processing techniques need to be carefully used, especially in conversational agents, so they can accurately understand the verbal cues and users' commands. This can be extra challenging as older adults often have altered speech due to changes in their soft palate that occur naturally with aging as well as a greater likelihood of having a confounding condition, such as dementia or Parkinson's disease.

Data connections should be robust or asynchronous (to avoid undesirable situations such as delays) and data must be encoded and stored in a secure way in order to ensure that the privacy will not be violated by the use of assistive agents. The agent itself should be also designed in a way that it is safe for older adults to use as safety is a key for the usability of the interactive agents, especially robots [Dautenhahn and Ghauoi 2014]. For example, as falls are common among older adults and can have very serious consequences, researchers should ensure that the agent does not introduce a tripping hazard.

After designing and developing a socially interactive agent, the next question is how to properly evaluate it. There are several different evaluation methods that can be used; finding the proper method highly depends on the task and the users. For example, the researcher should ask: (a) what task(s) should be selected and how to present it/them in a way that represents real-life settings? and (b) what is the best communication approach to use? (e.g., do users have any specific disabilities that might prevent them from answering questionnaires, getting involved in discussions, or expressing their opinion). The most common evaluation methods for socially interactive agents include, but are not limited to, validated methods such as the system usability scale and questionnaires assessing different aspects of agents (see Saez-Pons et al. [2015], Saunders et al. [2015], Syrdal et al. [2015], Ghafurian et al. [2021a] for examples), standard tests that measure effects on participants, such as their mood, depression level, loneliness, and so on (e.g., see Wada et al. [2005], Shibata [2012]), using open-ended questions and scenarios (to understand effects in a wider context; see Syrdal et al. [2014, 2015]), observations outside the experimental settings [Sabanovic et al. 2006] (e.g., using activity logs; see Webster et al. [2015]), and video-based studies (e.g., see Walters et al. [2011]).
Some of these methods might not be as representative as direct interactions but can allow researchers to evaluate the systems with a larger and more diverse range of participants.

Regardless of the evaluation method, it should be comprehensive, or in other words, the designers need to make sure that it covers evaluating the different aspects of the agents that are of interest. For example, Breazeal et al. [2016] suggested six important factors when evaluating social robots (which can be generalized to other types of agents):

1. Positive/negative usefulness, that is, whether participants express that the robot made tasks easier/harder and whether specific aspects of the robot were easy/hard to use.
2. Every-day experience, that is, whether people were interested in using the robot outside of the experimental setting.
3. Scenario capability, that is, whether participants refer to specific capabilities of the robot in a scenario.
4. Companionship, that is, whether the robot can provide companionship or social interactions.
5. Specific needs, that is, if the robot can address needs caused as a result of disabilities or aging.
6. Specific difficulty, that is, whether specific aspects of the robot are perceived to be hard to use as a result of aging or disability).

23.4 Considerations

There are many considerations that researchers should take into account while designing, developing, and testing assistive agents for older adults. In general, most considerations that are important in designing any intelligent agent are important to be taken into account for older adults as well. However, complexities that often accompany aging require additional considerations, some of which are discussed in this section.

First of all, it is important to account for changes in abilities that increase in prevalence with increasing age, such as impaired hearing, impaired vision, memory loss, and changes in mobility. Researchers should consider these conditions in the design of the agents, as well as when conducting studies with older adults. Some of these considerations are discussed in the previous sections and in Table 23.1.

Another important consideration when designing technologies and conducting studies with older adults, as well as for reporting the findings, is to ensure that
the vocabulary used to talk with and about older adults is appropriate. Words and phrases that might have negative connotations should be avoided. For example, “the elderly” is usually considered to be stigmatizing since they imply that all older adults are a homogeneous, frail group. More neutral words/phrases are preferred such as “older adults” or “older persons.” The same is true for addressing specific groups of people, for example, when working with people living with dementia. As dementia is a permanent condition, not an acute illness, it is important to avoid words/phrases such as “patient” (unless in clinical research) and “suffering from dementia.” Instead, phrases such as “a person living with dementia” should be used. Resources such as Dementia Australia [2020] can be valuable in learning appropriate and current terminology. We need to be cognizant that language is dynamic and changes over time. So what is considered to be acceptable will change in the future.

The design of novel assistive systems for supporting aging must include mechanisms of adaptation and individualization. Intelligent adaptive techniques such as machine learning and control-theoretic approaches can capture valuable information collected during interactions with the agent (e.g., behavioral clues, physiological responses, emotional data, annotations) and other existing data about humans' behaviors and preferences. These types of data can be used to dynamically adapt the agent's actions and customize responses [Whelan et al. 2018]. Therefore, a timely, diverse, multilevel, and contextually informed adaptation in assistive agents could greatly benefit the agent's capabilities to keep users engaged and motivated. While there is a plethora of novel learning methods based on sophisticated computational techniques (e.g., reinforcement learning, deep learning), there is limited research on personalizing the agents according to individuals' physical, emotional, or cognitive differences [Matarić and Scassellati 2016, Abdi et al. 2018] (e.g., by considering individuals' likes, dislikes, and behaviors [Dautenhahn 2007]). This largely remains an exciting area for further research.

Important factors that are required to increase adoption of technologies in older adults (e.g., trust and perception of usefulness) are shown to be influenced by social and adaptive capabilities in assistive robots [Heerink et al. 2010]. For instance, the utility of Internet tools as a health resource for older adults was assessed in 1,450 adults and older adults (50+), revealing how website design features such as information credibility and user-friendly interfaces may build trust among the older adults [Zulman et al. 2011]. Similar research has been performed assessing the acceptability of socially assistive agents by older adults, revealing how the intention to use is affected by variables such as perceived enjoyment and usefulness [Heerink et al. 2010]. However, more research should be done to better understand the role of important moderating factors such as willingness to use
new technologies, general attitude toward technology, and knowledge required to foster the adoption of agents for assisting aging.

Another important consideration that differentiates robots from virtual agents is the concept of embodiment, which plays an important role in producing empathic experiences of social interactions between humans and technology [Dautenhahn 1997]. In a study comparing virtual agents with physical robots, it was argued that physical presence affects people’s responses more than physical embodiment in social robots [Li 2015]. While some studies have shown that robots might be more appealing than screen-based agents in some contexts [Shinozawa et al. 2005, Lee et al. 2006], more systematic and comprehensive research with older adults is needed to better identify how the embodiment construct can be used to influence the design of assistive agents [Heerink et al. 2010]. Further, after selecting the platform, the decision of “what the agent should look like” can be also challenging as the design space is large and a human-like design is not necessarily the best design [Breazeal et al. 2016]. While there seems to be a preference toward human-like behavior and appearance for companion robots [Walters et al. 2008], other examples such as PARO demonstrate the applicability of different form-factors to different application areas.

Immersive mediums such as virtual reality (VR) should also be considered in the discussion of the importance of physical embodiment in the design of assistive agents [Kilteni et al. 2012]. For instance, research has shown that the display mode (e.g., VR or flat screens) has a clear influence on aspects of the user’s experience such as on positive and negative emotions as well as motion sickness [Xu et al. 2020].

Robots are capable of carrying out physical tasks for which the other types of agents (such as virtual agents) cannot be used. Interacting with a robot can also be more natural because users might relate it to interactions with physical toys (e.g., toys they may have used in their own childhood or when playing with their children or grandchildren) or other objects as opposed to learning how to interact with a novel interface. Natural interactions can be specifically important for the success of assistive agents in specific contexts, especially in contexts such as dementia care where learning new skills may be difficult or impossible. However, embodiment usually introduces additional costs. Not only is the average initial purchase cost higher than a virtual agent, but their increased mechanical complexity also means that physically embodied agents may need more maintenance. This can be difficult, a nuisance, and expensive, which can discourage the use of this type of technology. Thus while physically embodied agents have the potential to assist with many tasks, as with any technology, the benefits of their implementation and ongoing use must clearly outweigh their costs.
Researchers and designers often “over-engineer” solutions by using an excessive amount of emerging and trending technologies. This can make the resulting solution overly complex and less robust, which translates into greater costs with unstable or undesirable performance. When it comes to the possibilities of including robots in our lives, Norman [2005] argues that social aspects of interaction are critically more complex than the technical ones, “…something that technology-driven enthusiasts typically fail to recognize.” That is to say, empowering agents with proper social abilities is a complex problem and may require elegant solutions that are intuitive and direct.

In a recent review of various robotic technologies created to assist older adults, one study concluded that the most effective robots (e.g., effectiveness defined by the level of improvement in outcome measures once compared with control groups) were robots for companionship and telepresence [Shishehgar et al. 2019]. Companion robots such as Paro, which has a balanced set of relatively simplistic sensors/actuators, have demonstrated more positive effects on older adult's wellbeing (e.g., mood, anxiety) than other existing, more complex robots [Abdi et al. 2018]. This might be due to the simplicity of creating reliable companion robots, as compared to social robots in other categories, which require a comprehensive sense of their environments (e.g., those helping with an activity of daily living). Therefore, the design of assistive agents targeting social and healthcare benefits for older adults should be focused on producing solutions with pointed and meaningful features if they are to support accurate and consistent functionality.

The concept of zero-effort technology (ZET) can help researchers to explore targeted and appropriate solutions. ZETs are a “class of technologies that operate and provide support with little or no perceived extra physical or mental effort by the people who are using them” [Mihailidis et al. 2011]. In other words, a ZET enables a person to do the task they want to achieve without them having to focus effort on operating the technology itself; the technology is aligned to and complements the abilities of the user perfectly. In this way, the user does not have to think about how to use the technology, which can result in higher levels of engagement and ongoing use. This does not necessarily mean the ZET does tasks for the user, rather it enables the user to do the task they want to do. One example of a ZET is self-adapting upper-limb rehabilitation robots, where the robot autonomously adjusts parameters such as reaching distance and applied force to match a person’s abilities as they fatigue during a rehabilitation session. Another example is ambient vitals monitoring, where objects embedded in a home are able to collect health-related data (e.g., blood pressure, heart rate) of its occupants through the day-to-day interactions with common objects (e.g., couch, chair).
Last but not least, ethical considerations are a critical aspect of designing assistive agents for any population. Researchers should carefully consider how their technology relates to ethical concepts and to employ ethically responsible development of their technology. Doing so can aid researchers in creating technologies that are more likely to: (a) be recognized by users as ethically appropriate, therefore have higher levels of uptake and use; (b) avoid unintentional duress/harm; and (c) obtain approval from ethics and regulatory boards. Some of the main ethical issues that need to be supported include autonomy, confidentiality, privacy, and informed consent [Kang et al. 2010]. Amirabdollahian et al. [2013] emphasize the importance of six ethical factors—autonomy, independence, enablement, safety, privacy, and social connectedness—when designing social robots for the care of older adults. Further, Robillard et al. [2018] demonstrate five principles of ethical design backed by evidence: (a) inclusive participatory design, (b) emotional alignment, (c) adoption modeling, (d) ethical standards assessment, and (e) education and training. They also propose a set of 18 practical recommendations based on these principles. To create successful assistive agents, their creators should ensure that the decisions related to all phases of the process (from design through development and evaluation) are ethical and are aligned with social and cultural values while respecting older adults’ privacy, security, dignity, and autonomy.

23.5 Current Challenges

There are many different challenges regarding design, development, testing, and deployment of assistive agents for aging, many of which apply to application areas and users beyond assistive agents and older adults. These challenges are discussed in this section.

23.5.1 Technology Acceptance

As with any population, acceptance of assistive agents by older adults can be affected by many factors. Heerink et al. [2010] have proposed a model to measure acceptance of assistive agents by older adults. This model considers two aspects: (a) factors that affect perceived ease of use and functionality of the agents and (b) the factors related to social interactions.

As also suggested by Heerink et al.’s model, assistive agents designed for aging should be easy to interact with, have a high level of performance (which is challenging, as in many cases a successful assistant should be able to perceive people’s intentions and goals to be able to adjust its assistance [Breazeal et al. 2016]), and be able to gain users’ trust. Performance becomes important in technologies related to aging because some older adults might be less technically proficient, which
affects their ability to understand how the agents work or troubleshoot malfunctioning agents. Furthermore, while older adults are the fastest growing adopters of technology, they consider which ones to use more carefully than younger generations. In general, this is due to the perceived effort required for learning how to interact with a new technology, the time and effort required for maintaining it, and less of a desire to acquire new “gadgets” unless they have clear value.

Further, factors such as human-like communication and the ability of the agents to meet users' psychological, emotional, social, and environmental needs have been identified in recent reviews as important elements to aid technology acceptability [Whelan et al. 2018]. These factors, such as the capability to express/perceive emotions, to engage in social relationship [Dautenhahn 1995], and to use natural cues, are considered to be key for an agent to be truly socially interactive [Dautenhahn 2007]. Building technology-driven agents that are impersonal can lead to a poor user experience and low acceptance from the users, even if the agents are “intelligent” and act appropriately in their environment [Dragone et al. 2015].

It is also important to consider that older adults’ preferences of which socially interactive agents to use (e.g., virtual agents and social robot) might depend on the type of the task. For instance, one study showed that older adults preferred robotic assistance over human assistance for specific activities of daily living such as laundry and medication reminders (instrumental); older adults were less open to allow robot assistance in activities for personal care [Smarr et al. 2012]. In some cases, this can be due to perception of “dehumanized care” [Sävenstedt et al. 2006], or in other words, older adults’ concerns about the reduction of interactions with the family caregivers [Wang et al. 2017]. It makes sense that tasks that are perceived as more “mechanical” have higher levels of acceptance for robot assistance compared to ones that are more personal or “human.” This does not mean robots should not be developed for supporting tasks of a more personal nature, but it does mean that developers must be sensitive to people’s perceptions and to complement these with appropriate design choices. Furthermore, acceptance of the technology may depend on perceptions of other stakeholders than just the older adult, such as their family, friends, and care providers.

### 23.5.2 Ease of Use and Perceived Need

If the assistive agent is designed through user-centered processes and evaluated properly before being deployed, there is a much higher chance that it will be successfully adopted by the users. However, user-centered design itself can be challenging when designing assistive agents for older adults; as with any research involving this technique, locating and recruiting participants that are interested
in getting involved in the studies require time and effort. If recruiting populations such as people living with dementia or from care homes, there are additional processes required to ensure that the study is safe for the participants and to get appropriate ethical and other permissions. If recruiting outside the care homes, reaching out to older adults who live independently can be challenging as well since some recruitment methods such as social media advertisements are less effective than with younger populations. Furthermore, patience and careful thought must be put into recruiting and inviting end users to be co-creators as the participants need to be kept engaged with activities that can be sometimes overwhelming and frustrating (e.g., long interviews, multiple focus group). Since co-designing requires a series of systematic and carefully planned steps to model users and create suitable interactive solutions through iterative processes [Muñoz et al. 2019], these must be carefully mapped out in advance and adapted as new information is learned in the earlier stages.

When it comes to populations with cognitive impairments (e.g., people living with dementia), there is a lack of appropriate methods and materials that can foster the active involvement needed for generating user-driven solutions rather than technology-driven “gadgets” [Suijkerbuijk et al. 2019]. This can be in part due to a general lack of researchers’ expertise and desire to appropriately include older adults in research or to capture their opinions, because it may require more time and the modification of methods so that they are matched to each person’s abilities. For example, self-reporting is not appropriate past in the early stages of dementia and needs to be replaced by observations in the moderate to severe stages.

### 23.5.3 Trust

Even if an assistive agent achieves a high level of performance and is capable of assisting older adults, it still needs to gain and retain older adults’ trust. While performance is an important factor that can increase trust, there are other factors that are shown to be effective, such as the nature of the task [Salem et al. 2015] and the affective connection between the agent and older adults.

Improving the social and emotional capabilities of assistive agents is one area that has recently seen attention in the literature [Konig et al. 2018, Robillard and Hoey 2018]. In general, the importance of the affective experience is emphasized through multiple studies, which have shown that the affective experience can increase users’ engagement [O’Brien and Toms 2008], improve loyalty [Jennings 2000], and increase people’s enjoyment [Chowanda et al. 2016] and their cooperation with the technology [Ghafurian et al. 2019].
In dementia care, social robots are being developed to support expression of emotions, even though the set of emotions the agent can express is typically limited to a small set of basic emotions [Chan and Nejat 2010]. They are also being designed to be able to interpret users’ affective states [Derek et al. 2012] so that they can adjust their behavior accordingly. However, expressing and understanding emotions are challenging and achieving accuracy in such behaviors should be a long-term goal, which becomes even more challenging when the users are older adults as fewer data sets and research have been done with this population. Additional challenges include, but are not limited to, detection of people’s emotions, showing proper emotions, and understanding the communication strategies that would be suitable according to an individual’s personality as these can differ from younger adults. As personality can shift with illness or chronic conditions, gaining a deeper understanding of how this affects communication and engagement is a key area of current research [Konig et al. 2018].

23.5.4 Agent Selection: Physically Embodied versus Virtual

Another challenge when developing an assistive agent for aging is to select the right platform. Usually there is a tradeoff between the effectiveness of the platform and its cost. For example, a robot might be more effective due to its embodiment; however, it might not be affordable or practical for many of the intended users. It may also need more effort from the older adults to maintain it.

With the fast growth of technology, the platform should be flexible and allow upgrades and added features/assistive functions. It is not reasonable to expect that the older adults will frequently change the platform and adopt to new technologies. Ease of use in social robots has been reported as an important concern for caregivers as well [Pino et al. 2015].

When choosing the appropriate agent for aging assistance, the scientific evidence comparing the benefits of both virtual and physically embodied agents is still inconclusive. While robots have surpassed virtual agents’ performance in aspects such as supporting communication and collaboration through physical contact [Breazeal et al. 2016] as well as motivating older adults to perform exercises [Fasola and Mataric 2011], similarly virtual agents have shown advantages in terms of pervasiveness, telecommunication, and emotional connection [Paiva et al. 2017]. Research has shown that physical agents have more authority over other types of agents (e.g., screen agents) and can be more persuasive [Li 2015], so they can be more suitable for applications where the robot has the role of a coach or therapist [Cabrita et al. 2018].
23.5.5 Long-term Adoption and Novelty Effect

Multiple reviews in the field of assistive and socially interactive agents in the context of aging have highlighted the need to conduct research over long periods of time [Whelan et al. 2018, Pu et al. 2019]. While the limited existing research has shown the positive long-term effects of social robots in some contexts [Wada et al. 2005, Wada and Shibata 2007], most of this has only studied older adults’ behaviors during a shorter period of time. Since the novelty effect tends to bias the responses of users in human–robot interaction scenarios when asked about attitudes toward technology, perceived usefulness, enjoyment, or ease of use [Smedegaard 2019], there is a growing need to consistently and longitudinally study the effects of socially interactive robots on older adults’ acceptability.

23.5.6 Security and Privacy

Maintaining privacy is of utmost importance for technologies to be successfully adopted. The issue of privacy becomes even more challenging and important when designing technologies for older adults as many conditions (e.g., dementia) can affect users’ perception and understanding of the security and privacy risks (thus assenting to the potential risks). Simple technology artifacts such as cameras can cause many ethical, moral, and practical concerns because not only end users (i.e., older adults) but also caregivers and visitors may feel intimidated by the constant observation [Mulvenna et al. 2017]. Older adults’ privacy concerns have included aspects such as who has access to footage, who watches the footage, where the footage is stored, and how secure it is [Mulvenna et al. 2017]. Therefore, while security risks should be minimized and the researchers should consider all the alternative options that can reduce privacy concerns (e.g., use voice recordings instead of cameras, if possible), it is also important to understand how to properly inform older adults about the possible security and privacy risks.

23.5.7 Enabling Choice

As with any user population, older adults value their autonomy and ability to choose. However, technologies for supporting older adults often do not support core and critical choices related to their use, such as when it is used, what it is used for, and who gets to make these choices. When they are available, these choices may not be presented to older adults in a way that they understand or reflect the factors they consider to be important in making related choices. A prominent example are technologies intended for supporting people living with dementia. These often require the person’s caregiver/family to make choices about the customization and use of the intervention with little or no ability for the person who is the targeted recipient of the technology (i.e., the person with dementia) to participate or make
their own choices. While not all choices can be accommodated, it stands to reason that the people who are the users of the technology should have a voice in when and how it is used. Much of this can be achieved by developing technology in such a way that it conveys what it does and enables choice in a way that is appropriate for older adult users and that complements aspects they consider to be important.

23.5.8 Developing Policies

One possible approach toward reducing these challenges is to design policies for care centers and care of older adults around the use of assistive technologies in a way that supports their appropriate uptake and use. The technology usage policy in this context can focus on multiple aspects of design and adoption of the assistive agents for aging and could help with (a) increasing privacy and addressing security and privacy concerns, (b) assisting older adults financially by keeping the costs reasonable for the users (e.g., through government-funded resources), and (c) improving older adults’ trust and attitude toward technology. As the creators and experts, it is imperative that researchers and developers of technologies for supporting older adults share their knowledge and participate in the formation of such polices. This will help to ensure the creation of policies that guide appropriate and reasonable development and use of technology that also mitigate unnecessarily hindering it.

23.6 Future Directions

There is a need to create more personalized and custom-made applications of assistive agents targeting both healthy older adults and those with cognitive and/or physical impairments. The ultimate goal is to produce socially interactive agents capable of providing more adaptive assistance by using subjective, behavioral, contextual (e.g., surroundings and environment), and/or physiological data. Ideally, an intelligent agent should be able to combine information from many different data sources to provide truly personalized adjustments in real time.

An important step toward creating successful assistive agents for supporting aging is to identify application areas where older adults are willing to use them. The literature to date is limited as definitive answers require studies with large and diverse groups of older adults. The application area can itself affect multiple decisions such as the type of agent (e.g., virtual agent, robot), agents’ capabilities (e.g., technical aspects such as the type of sensors), and the necessary background for designing and implementing the agent (e.g., knowledge of specific areas of Machine Learning and Computer Vision). Researchers may have less flexibility with these design choices due to limited resources such as the research team’s background, availability of only specific virtual agents or robots in research
groups, and costs associated with adding different functionalities to the technologies. Therefore, the current common approach is to select an application area based on the available resources in the team and to work with older adults to create an agent that would perform well in the selected application area. However, this might limit research to specific application areas and might not cover the application areas where older adults would highly benefit from an assistive agent. Further, older adults may not be interested in using social agents for specific activities for reasons such as trust or perceived impact on human relationships. Therefore, it is extremely important to work on identifying different application areas and understanding older adults’ preference not only toward social agents but also toward tasks/activities that these agents could help them with [Broadbent et al. 2009].

Another challenge that requires future work is making the assistive agents truly personalizable, with appropriate emotional and social intelligence. As discussed earlier in this chapter, such intelligence is highly beneficial for gaining older adults’ trust and interest in using social agents. While emotional intelligence of technologies can be important in many domains, it can be key for adoption of socially interactive agents by older adults, especially for those with specific cognitive disabilities such as dementia [Konig et al. 2017]. Yet, there are many challenges involved in making them emotionally intelligent, including (a) finding techniques and algorithms that allow us to properly understand users’ emotions (which can be more challenging when the users are older adults), (b) understanding users’ emotional states, and (c) adapting social agents’ behaviors based on a user’s personality. All of these challenges and more need to be addressed in the future work. Many of the existing models of emotions are inspired from how humans and animals show emotions, which can be informative for researchers to design computational models of emotions for agents [Breazeal et al. 2016].

The next challenge that can highly affect the success of social agents for supporting aging is to emphasize their benefits in society and provide organizations with the financial and other resource support to acquire and use these agents. To that end, future research should work on developing policies that will facilitate adoption of socially interactive agents in care homes and the community as well as policies that accelerate their ethical design and use. These policies should be flexible enough to provide guidelines regarding the adoption of social agents running on different platforms and in different contexts.

Despite the general consensus that socially interactive agents should follow co-design processes, there are still limited guidelines about how to adjust the co-design sessions for older adults, especially for older adults with conditions such as dementia. Appropriate research methods need to be developed and disseminated in a way that the creators of social agents can understand and implement
consistently [Kachouie et al. 2014]. Older adult participants should be involved in activities beyond interviews, focus groups, and usability testing; they should be involved as active members of research and development teams who are able to provide rich information and aid data interpretation since they are living their own experiences with aging and technology and can convey these directly.

Finally, while there are many advancements in machine learning and computer vision, algorithms such as activity recognition, face recognition, voice recognition, detection of physiological states, and object recognition need to be improved to increase the accuracy and dependability of social agents. For example, datasets populated by older adults need to be assembled and used when recognizing voices, faces, and facial expressions as they can differ from the available data used for training these algorithms, which consists primarily of data from younger adults. Improvements in these algorithms can enable researchers to create social agents that can act upon their environments and interact with older adults more appropriately.

### Summary

The necessity of designing assistive agents that support aging is growing primarily due to the increasing population of older adults coupled with the increase in the abilities of assistive agents in improving older adults’ quality of life and wellbeing.

This chapter began by discussing the need to provide more accessible and interactive assistive agents to support a fast-growing population of older adults. We discussed how assistive agents can fill gaps caused by demographic shifts, help reduce costs, and improve older adults’ quality of lives. This was followed by an overview of the use of socially assistive agents (both virtual and physically embodied) to support aging. We then discussed different application areas and possible uses of social agents in those areas. Results to date have been promising, indicating that the assistive agents have good potential to improve different aspects of older adults’ lives, including companionship, engagement, health, leisure, therapeutic, and ADL support.

Next, we presented different approaches for designing and developing assistive agents. Advantages, challenges, and considerations for older adults were discussed for five different methods (i.e., focus groups, interviews, questionnaires, ethnographic field studies, and case studies). We emphasized the importance of including older adults throughout the design process in order to create an assistive agent that can successfully be adopted by older adults.

The chapter also discussed considerations when designing agents to assist older adults. While many of the considerations are common with designing agents
for other age groups and contexts, we argued that there are additional considerations when designing for older adults as different and changing physical and cognitive abilities need to be taken into account. We discussed some of the trade-offs for choosing a suitable platform for designing the assistive agents as well as the need to reduce the perceived complexity and the effort required to understand and use the technologies.

Existing challenges regarding design, development, testing, and deployment of assistive agents for aging were discussed, which covered (a) developing technologies that increase older adults’ acceptance of technology; (b) importance of ease of use and perceived need of agents; (c) increasing users’ trust in the agents; (d) importance of agent selection between virtual and physically embodied agents; (e) long-term adoption and how to ensure that evaluations are not affected by the novelty effect; (f) improving security and reducing privacy issues; (g) supporting transparency and respecting older adults’ choice; and (h) developing policies that support use of assistive agents for aging.

The challenges discussed in this chapter need to be addressed by future work to aid the short-and long-term adoption of socially interactive agents by older adults. Areas that need to be addressed by future research include: (a) identifying application areas and gaps as perceived by the older adults; (b) revising existing methods to better involve older adults in the design processes; (c) improving accuracy of different aspects of the technologies (e.g., activity recognition, object detection); (d) improving the affective connection between the agents and older adults; (e) personalizing the behavior of the agents according to the personality of each user; and (f) developing policies that will support the appropriate adoption and use of socially interactive agents.

This chapter provided concepts for researchers and designers to better understand the considerations and challenges related to designing assistive agents for older adults, as well as ideas that can help with designing assistive socially interactive agents that can become successful in assisting older adults in many different aspects of their lives.

Acknowledgment

This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program, Schlegel Research Chair at the Research Institute for Aging, the Network for Aging Research at the University of Waterloo, the Wes Graham Foundation, the American Alzheimer’s Association, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Health Research (CIHR), the Canadian Consortium on Neurodegeneration and
Aging (CCNA), and AGEWELL, Inc., a Canadian Network of Centers of Excellence (NCE). We thank David Traum for his comments on this chapter.

References


Chapter 23  Socially Interactive Agents for Supporting Aging

10011: Lecture Notes in Computer Science. Springer, 85–95. DOI: https://doi.org/10.1007/978-3-319-47665-0_8.


K. M. Lee, Y. Jung, J. Kim, and S. R. Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile


References


A. Suryani. 2013. Comparing case study and ethnography as qualitative research approaches. *Jurnal ILMU KOMUNIKASI* 5, 1. DOI: https://doi.org/10.24002/jik.v5i1.221.


Health-Related Applications of Socially Interactive Agents

Timothy Bickmore

24.1 Motivation for Socially Interactive Agents in Healthcare

Health is the prerequisite for any human endeavor, necessarily occupying the base of Maslow’s Hierarchy of universal human needs [Maslow 1943]. The maintenance of health is thus a significant preoccupation in all societies, spawning a vast and complex industry, spanning a wide range of organizations, services, and products, starting with frontline caregivers and the myriad support services and products they, and their patients, require. Around 10% of the world’s resources are devoted to healthcare, with US$7.8 trillion spent worldwide in 2017, and growth in healthcare spending is outpacing the overall growth in the world economy [WHO 2019].

Given its importance and cost, healthcare represents a significant opportunity for automation in general and SIAs in particular. Developing countries have severe shortages in resources that may only be met through relatively low-cost automation. However, the most advanced countries can also benefit, since even small improvements in health outcomes or cost reduction can yield large returns when scaled to a population.

SIAs provide unique affordances in automating aspects of healthcare delivery, since they can be used to most directly automate the care that an expert human health provider would give, especially when the care is in the form of information or guidance delivered in a face-to-face clinical encounter. SIAs can not only provide the instrumental, utilitarian delivery of medical information to patients but also the social, emotional, and relational messages—such as compassion and empathy—that more conventional digital media lacks. Further, they can provide
these messages with perfect consistency and fidelity every time, to every patient, in every circumstance. SIA are never intended to replace human healthcare providers but rather to automate routine parts of their work so they can focus their time on exceptional cases or see more patients. SIAs can also be used in ways that humans are not capable of, such as providing counseling to every patient at any time of the day or counseling to patients who cannot see human providers due to cost or other factors.

In this chapter, I describe the significant opportunities for SIA in healthcare, along with several examples of agents and robots that have been developed and evaluated to date. I focus on SIAs that are designed to interact directly with patients (receiving care for a health condition) and consumers (focused on preventing illness and maintaining wellness), providing health education and the promotion of health-related behavior. I do not cover SIAs designed to interact with healthcare providers (e.g., as virtual patients for training [Ferdig et al. 2012]), although there are several researchers investigating these. My focus will also be on SIAs that have been developed and evaluated in randomized clinical trials with real patients and consumers, although I will also mention a few relevant SIAs that have not either made it into trial or transitioned into commercial products.

24.1.1 The Importance of Health Behavior
Modifiable health behavior, such as physical inactivity, poor diet, and smoking, accounts for nearly 40% of all deaths in the US [Maslow 1943]. Although patient and consumer behavior can significantly impact health outcomes in almost every area of medicine, I will focus here on a few select areas that have the potential for very significant impact, given that they can be addressed through automated health behavior change interventions.

24.1.1.1 The Burden of Overweight and Obesity
The rates of overweight (body mass index greater or equal to 25 [Arroyo-Johnson and Mincey 2016]) and obesity (body mass index greater or equal to 30) have exploded across the world in the last few decades. Worldwide, 39% of adults were overweight and 13% were obese, and an additional 340 million children were overweight or obese, in 2016, with these conditions linked to significant increases in rates of cardiovascular disease, diabetes, and other conditions, leading to increased rates of morbidity and mortality [WHO 2020]. Addressing overweight and obesity primarily involves diet modifications and increases in physical activity, both lifestyle health behaviors that have been the target of numerous automated interventions, including several with SIAs [Bickmore et al. 2005a, 2005b, 2013a, 2013b, King et al. 2013].
24.1.2 The Burden of Chronic Disease
Chronic conditions, such as diabetes, hypertension, and atrial fibrillation, affect half of all adults in the US, are responsible for 70% of all deaths in the US, and account for 75% of US healthcare expenditures [Wullianallur and Raghupathi 2018]. With the aging population, the prevalence of chronic conditions—especially multiple chronic conditions—is continuing to rapidly increase. These conditions can require challenging self-care management regimens, spanning medication adherence, lifestyle modifications, and vigilant symptom monitoring, in addition to regular visits with healthcare providers to prevent disease progression. However, 60% of individuals with chronic conditions are poorly adherent to their prescribed treatment regimens, with even higher rates for older adults or those with limited health literacy [Dunbar-Jacob and Mortimer-Stephens 2001]. Promoting treatment adherence has been the target of several automated interventions, including by SIAs [Bickmore et al. 2010].

24.1.3 The Burden of Substance Abuse
Over 150 million people worldwide are estimated to have alcohol or drug use disorders, leading to over 3 million deaths per year [GBD 2016 Alcohol and Drug Use Collaborators 2018]. Counseling, using techniques such as Motivational Interviewing (Section 24.2.3) and Cognitive Behavior Therapy, have been shown to be effective at reducing alcohol and substance use in most populations and settings, including single session “brief interventions” in primary care [DiClemente et al. 2017]. Many of these counseling techniques are automatable, and SIAs provide additional affordances for counseling over more traditional media. SIAs have successfully been used in several interventions for substance use ([Yasavur et al. 2014, Olafsson et al. 2020], Section 24.4.3).

24.1.2 Logistical Factors: Convenience and Cost
There are many reasons why SIAs are well-suited to health behavior interventions. As with all types of automation, cost savings are often first to mind. However, SIA can provide emulations of human health counselors that are available 24×7 and, in the case of those deployed on mobile devices (Section 24.4.2), wherever the user happens to be. Notions of convenience also extend beyond mere availability, to patient comfort in taking the time they need to get the information they require. For example, after interacting with an IVA-based virtual hospital discharge nurse (Figure 24.1), 36% of 149 patients said they preferred receiving their discharge instructions from the IVA rather than their doctors or nurses in the hospital, compared to only 24% who said they would prefer receiving it from their human providers (the remaining 39% had no preference) [Zhou et al. 2014]. When asked
why they preferred the agent, participants indicated that they did not feel rushed
with the agent the way they felt with the doctors or nurses in the hospital (“I prefer
Louise, she’s better than a doctor, she explains more, and doctors are always in a
hurry.”).

24.1.3 Social-emotional Factors: Stigma, Trust, and Alliance
There are other aspects of healthcare for which SIAs can outperform human
providers. Many areas of medicine require patients to self-report behaviors that
are potentially stigmatizing, such as substance use or sexual behavior. Patients
may also be reluctant to report failures to adhere to recommended or prescribed
regimens spanning diet, physical activity, self-care, or medications. Avoidance or
underreporting of such behavior to human clinicians is well-documented, and there have been studies dating back over 40 years demonstrating that patients report potentially stigmatizing conditions more accurately to a computer than to a human [Card and Lucas 1981]. More recent studies have demonstrated that this effect carries over to SIAs: in a study comparing self-reports of intimate information to an IVA, participants revealed more when they were told they were interacting with a fully autonomous system compared to those who were told they were interacting with a remote-controlled avatar [Lucas et al. 2014].

24.1.4 Literacy Factors: Health, Computer, and Reading Literacy

Health literacy—the ability to perform the basic reading and numerical tasks required to function in the health care environment [National Institutes of Health Program Announcement 2007]—affects patients’ ability to understand medication labels and instructions, hospital discharge instructions, instructions for assistive devices and medical equipment, and educational material [AMA 1999]. Approximately 36% of US adults have inadequate health literacy [Kutner et al. 2006]. This problem is not uniformly distributed in society; among indigent and minority patients in urban areas this number rises to over 80% [Williams et al. 1995a]. Patients with inadequate health literacy report lower health status [Weiss et al. 1992, Baker et al. 1997], are less likely to use screening procedures, follow medical regimens, keep appointments, or seek help early in the course of a disease [Weiss et al. 1994], have greater difficulties naming their medications and describing their indications [Williams et al. 1995b], more frequently hold health beliefs that interfere with adherence [Kalichman et al. 1999], have higher health-care costs [Weiss et al. 1994], and have higher rates of hospitalization [Baker et al. 1997].

Evidence suggests that a face-to-face encounter with a health provider—in conjunction with written instructions—remains one of the best methods for communicating health information to patients in general, but especially those with low literacy levels [Colcher and Bass 1972, Madden 1973, Clinite and Kabat 1976, Morris and Halperin 1979, Qualls et al. 2002]. Face-to-face consultation is effective because it requires that the provider focus on the most salient information to be conveyed [Qualls et al. 2002] and that the information be delivered in a simple, conversational speaking style. Protocols for “grounding” in face-to-face conversation—the use of verbal and nonverbal cues such as head-nods, gaze, and acknowledgement tokens (“uh-huh”, “OK”) to signal mutual understanding [Clark and Brennan 1991]—allows providers to dynamically assess a patient’s level of understanding and repeat or elaborate information as necessary. Face-to-face conversation also allows providers to make their communication more explicitly interactive by asking patients to do, write, say, or show something that demonstrates their
Health-related applications of socially interactive agents understanding [Doak et al. 1996]. Finally, face-to-face interaction allows providers to use verbal and nonverbal behaviors, such as empathy [Frankel 1995] and immediacy [Argyle 1988, Richmond and McCroskey 1995], to elicit patient trust, enabling better communication and satisfaction.

Healthcare SIAs may prove particularly effective for individuals with low health literacy, given their ability to emulate face-to-face consultation. This hypothesis has now been supported in several studies and clinical trials. Patients with low health literacy reported significantly higher levels of trust in the virtual discharge nurse IVA (Figure 24.1) compared to patients with adequate health literacy. Patients with low computer literacy were also significantly more satisfied with the virtual nurse and rated it significantly higher on ease of use compared to those with high computer literacy [Zhou et al. 2014].

Studies have also demonstrated that low literacy individuals can complete some health-related tasks more successfully with SIAs compared to more traditional computer interfaces. For example, in an evaluation of an IVA that helped cancer patients find clinical trials to volunteer for, low health literacy patients completed significantly more correct search tasks and were significantly more satisfied with the SIA compared to a functionally equivalent facet- and keyword-based search engine [Bickmore et al. 2016]. Similarly, among 273 participants recruited from an urban safety net hospital, 74% of whom had possibly or likely low health literacy, significantly more were able to complete a family health history with an IVA that simulated a genetics counselor (97%) [Wang et al. 2015] compared to those using a functionally equivalent web form-based interface (51%) [Wang et al. 2017].

24.2 Models and Theoretical Frameworks

Many theories of health communication and health behavior change have been developed over the last several decades that can be used to inform the design of SIAs intended to educate or change the health behavior of their users. Health behaviors span a wide range of physical actions, contexts, and time durations, from the relatively simple action of showing up at an appointment to obtain a one-time vaccination, to chronic disease self-care management that may require years of consistent, concerted effort across a wide range of activities spanning diet, exercise, medications, and self-checks. Similarly, interventions to change health behavior range from single brief educational messages to years of counseling and coaching sessions. SIAs have been used for all of these types of behaviors and in all of these intervention formats. Here, I focus on a few particularly relevant theories and frameworks that have been used in SIA-based health interventions.
**24.2.1 Health Communication and Education**

Communication of health information forms the cornerstone of any intervention. Unless someone knows what they are supposed to do and how to do it, clearly understands the reasons for doing it and the risks of not doing it, they will likely not engage in a desired health behavior. SIAs that are developed primarily for educational purposes have been referred to as “pedagogical agents” and are reviewed in depth in Chapter 21 on “Pedagogical Agents” [Lane and Schroeder 2022] of this volume of this handbook. Here I focus briefly on communication and pedagogy that is unique to health. An important finding to keep in mind, however, is that it is widely accepted that education is necessary but not sufficient to achieve behavior change [Nichols 1994]. Thus, communication and pedagogical strategies must be used in conjunction with other counseling techniques to succeed in short- or long-term behavior change.

The literature on health communication is vast, from studies of one-on-one doctor–patient communications to population-wide public health campaigns [Schiavo 2013]. One area of health communication research that is particularly relevant to SIA is computerized tailoring of health messages. One proposed typology describes a spectrum of health communication messages, ranging from those that are invariant across all recipients (“generic communication”), to those tailored only on personal characteristics such as the recipient’s name (“personalized generic communication”), to those designed for different segments of the population (“targeted communication”) [Kreuter et al. 1999]. Tailored communication is at the extreme end of this spectrum, as it has been defined as “any combination of strategies and information intended to reach one specific person, based on characteristics that are unique to that person, related to the outcome of interest, and derived from an individual assessment” [Kreuter et al. 2000]. For example, the user’s sex, age, and education may be input and health promotion text adjusted accordingly, typically using relatively simple template-based text generation techniques [Reiter and Dale 2000, Reiter et al. 2003], with all possible text variants manually pre-authored. Importantly, most interventions tailored on one or more factors derived from a health behavior change theory, such as the “stage of change” from the Transtheoretical Model (Section 24.2.3). Work on computerized tailoring was an acknowledgment that the “one size fits all” approaches to public health communication were not as effective as they could be. Tailoring is hypothesized to work through several mechanisms, including increased attention to tailored health messages, more effort spent processing tailored messages, and increased self-referential thinking [Hawkins et al. 2008]. A large meta-analysis of 57 computerized tailoring studies found that their mean effect size on health behavior change
was $r = 0.074$ [Noar et al. 2007]. Importantly, this study also found that there was a significant relationship between the number of factors tailored on (ranging from 0 to 9 theoretical constructs) and the impact on health behavior change.

SIAs for health behavior change have been developed with varying degrees of tailored messaging. However, given their ability to engage users in natural language dialogue, with each SIA utterance generated in real time via a dialogue system and natural language generator, they have the ability to tailor their message to each turn of conversation based not only on theoretical and demographic factors known about the user but on the discourse context, including what was just said by the user or what they said in prior conversations. This gives them the capability to perform tailoring at a level of granularity never before possible.

### 24.2.2 Persuasive Technology

Moving beyond communication and education, SIAs can use a variety of techniques to persuade a user to perform a health-related action. Technologies designed to motivate a single action have been referred to as persuasive technologies [Fogg 2003]. Decades of psychology studies have provided a list of persuasive techniques that could be automated. For example, Cialdini [2001] outlines six strategies of persuasion—including authority, liking, and reciprocity—that are widely used in marketing communication strategies. Most of these techniques have been used in automated systems, including SIAs. Example SIA studies include reciprocity, in which the SIA does a favor for the user first to gain compliance [Lee and Liang 2016], reciprocal deepening self-disclosure to build trust before making a request of the user [Moon 1998], and physical touch to increase persuasion (“Midas touch” [Haans and IJsselsteijn 2009]).

### 24.2.3 Health Behavior Change

While immediate compliance can be important in brief interventions for single-action health behaviors, a more longitudinal, incremental, and relational approach is required to change ingrained habits over longer periods of time. Several theoretical models and techniques have been developed in the field of behavioral medicine to address these more substantive problems [Glanz et al. 2008]. A sampling of theories includes:

- The Health Belief Model [Becker 1976, Janz and Becker 1984], which focuses on increasing the user’s perceived threat of a disease or negative health outcome to motivate behavior change, with perceived threat comprised of susceptibility (the perception of risk for the negative outcome), and severity (the perception of the seriousness of that outcome). Thus, the emphasis on these
interventions is on communication of risk, helping the user weigh barriers and benefits to change, and on increasing user self-efficacy (confidence to change).

- The Theory of Reasoned Action [Fishbein and Ajzen 1975] and the Theory of Planned Behavior [Ajzen and Madden 1986], which focus on the perceived likelihood of performing a behavior. The Theory of Reasoned Action posits that intent is driven by positive attitude toward the behavior and subjective norms (perceived societal pressure). The Theory of Planned Behavior posits an additional driver, perceived behavioral control, which is the extent to which the user believes the behavior is under their control. Thus, the focus of these interventions is on increasing positive attitude, perceived normative pressure, and perceived behavioral control.

- Social Cognitive Theory [Bandura 1986], which posits that the most central drivers of health behavior change are self-efficacy and the perception that engaging in the behavior will lead to positive outcomes.

However, the health behavior change theory that has arguably received the most empirical support and is the most readily automatable is the Trantheoretical Model (TTM), also known as the “stages of change” model [Prochaska and Velicer 1997]. This model posits that people go through a series of stages when they change a behavior, and the individual behavior change techniques and messages that are most effective at a given time depend on what stage a user is in. The five Stages of Change describe different levels of readiness to change [Velicer et al. 2000] and span Precontemplation, Contemplation, Preparation, Action, and Maintenance. Precontemplators are not intending to change in the next 6 months. Contemplators are intending to change in the next 6 months. People in the Preparation stage are planning to change in the next 30 days. People in the Action and Maintenance stages have changed, with those in Action having changed within the prior 6 months. The model also classifies behavior change techniques into a set of 10 Processes of Change and posits additional constructs including Decisional Balance (weighing Pros vs. Cons of change) and self-efficacy. The TTM is relatively straightforward to automate since a user’s Stage can be assessed through one or two brief questions, then used to index a library of behavior change techniques or messages. The TTM has been used to drive the behavior of SIAs designed for a variety of health behaviors [Velicer et al. 2009, Bickmore et al. 2013b, Jack et al. 2015].

Users in Precontemplation pose a particular challenge since they are not motivated to even engage in an intervention at all. Motivational Interviewing provides a set of counseling techniques specifically focused on boosting motivation to change.
and thus moving users out of Precontemplation and into later stages in which they are willing to take some action toward change [Miller and Rollnick 2012]. Motivational Interviewing involves prompting users to talk through their perceived pros and cons of change to help them resolve their ambivalence in favor of taking action. Some aspects of Motivational Interviewing have been implemented in several SIA-based interventions [Schulman et al. 2011], although much of it requires eliciting and responding to unconstrained user utterances, which represents a currently insurmountable challenge for natural language understanding (see Chapter 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021] of volume 1 of this handbook [Lugrin et al. 2021] and the discussion of safety issues with doing this in healthcare in the Current Challenges Section of this chapter).

While the theories described here provide overarching frameworks to guide intervention design, they do not, on their own, dictate specific actions that a SIA should take at a given time with a given user. At best, they provide a description of the types of techniques (e.g., TTM Processes of Change) or type of outcome (e.g., increased risk perception for the Health Belief Model) that is desirable at a given time. These specifications provide very general indices into a large set of individual behavior change techniques that have been developed and evaluated over the years in behavioral medicine. Going beyond the 10 Processes of Change categories, Michie has spent several years developing taxonomies of behavior change techniques [Michie et al. 2013]. Her taxonomy has 16 top-level categories organizing a total of 93 individual techniques. An excerpt of the taxonomy is shown in Table 24.1.

### 24.2.4 Trust & Therapeutic Alliance

There is a strong correlation between the quality of professional–client relationships and outcomes across a wide range of human helping professions. The dimension of the clinician–patient relationship that is credited with the significant influence on outcome—the Working Alliance—is based on the trust and belief that the therapist and patient have in each other as team members in achieving a desired outcome, and has been hypothesized to be the single common factor underlying the therapeutic benefit of therapies ranging from behavioral and cognitive therapies to psychodynamic therapy [Gelso and Hayes 1998]. The Working Alliance construct has been hypothesized to have three sub-components: a goal component, reflecting the degree to which the clinician and client agree on the goals of the therapy; a task component, reflecting the degree to which the clinician and client agree on the therapeutic tasks to be performed; and a bond component, reflecting the trusting, empathetic relationship between the client and clinician [Horvath and Greenberg 1989, Gelso and Hayes 1998].
In the context of longitudinal health behavior change interventions, Working Alliance can positively impact both compliance, the likelihood that a user will follow an SIA’s immediate recommendations, and retention, the likelihood that the user will continue working with the SIA over time. There is some evidence that strategies that focus on near-term compliance do so at the cost of retention [Bickmore et al. 2007], and it may be that retention is the primary beneficiary of a strong Working Alliance [Bickmore and Picard 2004]. Retention is arguably the more important objective and yet the more difficult goal. The majority of longitudinal health behavior change interventions suffer from attrition over time, when users give up for one reason or another.

SIAs are ideal for establishing Working Alliance relationships with users, and have many affordances lacking in more traditional print or digital media for this purpose. Their use of nonverbal communicative behavior (e.g., facial display of empathy, Figure 24.2), in conjunction with immediacy behaviors, which demonstrate attention to and engagement in the user by the SIA, along with dialogue designed to build and maintain a social-emotional relationship with the user, make SIAs ideal platforms for establishing strong, trusting relationships with users, and leveraging these in health behavior change interventions [Bickmore and Picard 2005].

**Table 24.1** Excerpt of Michie’s Taxonomy of Behavior Change Techniques (based on Michie et al. [2013])

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheduled Consequences</td>
<td>Shaping (modifying a behavior incrementally, such as gradually weaning an undesirable behavior or increasing a desirable behavior in small steps). Example use in SIAs: Bickmore et al. [2013b]</td>
</tr>
<tr>
<td>Reward &amp; Threat</td>
<td>Social Reward (provision of congratulatory messages by another person). Example use in SIAs: Bickmore and Picard [2005]</td>
</tr>
<tr>
<td>Goals &amp; Planning</td>
<td>Problem Solving (identifying barriers to change and assisting user in addressing them). Example use in SIAs: Bickmore et al. [2010]</td>
</tr>
<tr>
<td>Social Support</td>
<td>Emotional Support (e.g., display of empathy when user is having difficulty). Example use in SIAs: Bickmore and Picard [2005]</td>
</tr>
</tbody>
</table>
24.3 A Brief History of SIAs in Healthcare

SIAs, broadly construed, have a long history of being used for health education and health behavior change interventions. Disembodied dialogue systems were used as far back as 1966 for psychotherapy. The ELIZA system was developed to simulate the behavior of a Rogerian psychotherapist, in which the patient and the computer exchanged typed text messages [Weizenbaum 1966]. Although ELIZA was not intended to be used for actual therapy, similar systems have been proven effective for therapy in which the system is essentially prompting a patient to think aloud and work through his or her own problems [Slack 2000]. Colby [1995] developed an ELIZA-like system that was designed to use Cognitive Behavioral Therapy to treat individuals with depression. In addition to providing typed text counseling with patients, the system provided text-based educational materials about depression.
Speech-based health counseling systems represent the next step closer to embodied SIAs. Several health interventions have been developed for use over the telephone, referred to as Interactive Voice Response (IVR) systems. IVR-based interventions have been developed and evaluated in clinical trials for several aspects of diet, physical activity, cigarette smoking, medication adherence, office visit adherence, disease screening behavior, and chronic disease management for hypertension, angina pectoris, chronic obstructive lung disease, asthma, diabetes mellitus, and depression [Migneault et al. 2006].

Full IVA-based health interventions began appearing in the early 2000s, with animated conversational health coaches designed for longitudinal health behavior change interventions in physical activity promotion [Bickmore and Picard 2005, Bickmore et al. 2005b], medication adherence promotion [Bickmore et al. 2010], hospital discharge patient education [Bickmore et al. 2009], and UV avoidance [Velicer et al. 2009]. In these systems, the agent used its embodiment to regulate the flow of conversation (e.g., gaze cues for turn-taking), emphasize key points (e.g., via eyebrow raises, beat gestures), convey additional meaning (e.g., deictic gestures at parts of medical documents in the agent’s virtual environment, as in Figure 24.1), mark topic boundaries (e.g., via posture shift), mark conversational frames using contextualization cues (e.g., proxemics, gaze frequency, facial display, gesture frequency to delineate social chat from task talk), and display empathy (e.g., using facial display of concern as in Figure 24.2, proxemics, and shifts in prosody) [Bickmore and Picard 2005].

The use of SRs for health interventions represents one of the most recent developments. Kidd developed the first SR for use in a longitudinal health intervention. His robot consisted of a touch screen computer with a physical robotic head attached, designed to assist users with a weight-loss program [Kidd and Breazeal 2008]. Participants interacting with the SR continued with the weight-loss program for twice as long compared to a group that used a touch screen-only device, although there were no differences in weight loss. More recently, SRs have been used in a wide range of health applications, from rehabilitation [Matarić et al. 2009] to distraction during medical procedures [Trost et al. 2020]. SRs have also been widely used for individuals with autism and to provide many kinds of support to older adults, both of which are covered in other chapters in this volume (Chapters 23 on “Socially Interactive Agents for Supporting Aging” [Ghafurian et al. 2022] and 25 on “Autism and Socially Interactive Agents” [Nadel et al. 2022] of this volume of this handbook). Riek [2017] provides an excellent overview of the potential for SRs in healthcare, describing their future use in long-term eldercare, inpatient and outpatient care, and psychiatric and palliative care.
There have also been many patient- and consumer-facing commercial applications of SIA in healthcare. This is a rapidly evolving area with companies and offerings coming and going frequently. A current snapshot includes several disembodied chatbots to provide counseling and therapy, such as the woebot depression counselor (successfully evaluated in a clinical trial [Fitzpatrick et al. 2017]). True SIAs include Molly, an IVA from Sensely that provides a general health counselor agent on smartphones for a wide variety of conditions, including chronic disease management (no clinical trials reported to date). The Autom SR was launched as a home weight-loss coach by Intuitive Automata, following successful evaluation at the MIT Media lab [Kidd and Breazeal 2008], and a similar SR was subsequently developed by Catalia Health for wellness coaching and chronic disease management (neither of these commercial robots have been evaluated in clinical trials to date). Finally, Care.Coach (previously GeriJoy) provides an IVA companion on a tablet computer for eldercare but is driven by a remote operator using “Wizard of Oz” technology. There are several publications covering development and usability testing of the system, but no clinical trial evaluation to date.

Example SIA Systems in Healthcare

**SIA to Address Sedentary Behavior**

Physical activity is one of the few health behaviors that has significant benefits for individuals of all ages, including older adults [Chodzko-Zajko et al. 2009]. Late-life exercise improves strength, aerobic capacity, flexibility, and physical function [Keysor and Jette 2001], and a change from a sedentary to more active lifestyle in midlife or beyond is associated with a reduction in mortality [Eriksson et al. 1998, Bijnen et al. 1999]. Despite these benefits, only about 25% of men and 20% of women aged 65 years and older in the US meet the national guidelines for regular physical activity [Nelson et al. 2007].

To address this need, an IVA-based virtual coach was developed to increase physical activity among geriatrics patients at an urban safety net hospital in the US. Patients were provided with the IVA on a touch screen tablet computer, along with a pedometer, to take home for 2 months. They had a daily counseling conversation with the virtual coach, based on steps uploaded from the pedometer, using a variety of health behavior change techniques. The virtual coach was evaluated in a randomized clinical trial, compared to a control group who were provided with pedometers and print materials on the benefits of exercise. A total of 263 older adults were enrolled, with a mean age of 71.3, of whom 61% were female and 51% had an educational attainment of a high school diploma or less. Participants in
the virtual coach group increased their daily step count significantly more compared to the control group at the end of the 2 months, although this effect waned by a 12-month follow up. The older adults who used the IVA were highly satisfied with the program, scoring it well above the midpoint on a standardized Likert scale measure of therapeutic alliance, and interacted with it an average of 30 out of the 60 days of the intervention [Bickmore et al. 2013b].

24.4.2 SIA to Address Chronic Disease

As mentioned in Section 24.1.1.2, chronic conditions, such as diabetes, hypertension, and atrial fibrillation, are responsible for the majority of healthcare expenditures and deaths worldwide. Management of chronic conditions typically require lifestyle health behavior change over the remaining years of an individual’s life. Smartphones provide an ideal platform for systems designed to help individuals manage these conditions, given their computational power and connectivity, their “anywhere, anytime” availability, and especially when coupled with mobile sensors or monitors. To explore this, a prototype IVA was developed and deployed on smartphones to help support individuals with atrial fibrillation (AF) (Figure 24.3). AF is an irregular heartbeat that increases the risk of stroke 3- to 5-fold and doubles the risk of death if untreated [Romero et al. 2014]. The IVA provided education about AF, promoted medication adherence and symptom monitoring, and promoted use of a heart rhythm sensor attached to patients’ smartphones.

The IVA was evaluated in a randomized clinical trial in which 120 patients with AF were either provided with smartphones with the IVA and a heart rhythm monitor installed, or standard care, for 30 days [Bickmore et al. 2018a]. Participants had an average age of 72.1 years old, and were 51.7% female, and completed an average of 16.2 conversations with the agent, lasting a total of 41.9 minutes, over the 30 days. Compared to patients undergoing standard of care, IVA patients reported significantly higher quality of life at the end of the intervention period, based on a standard measure for patients with AF [Spertus et al. 2011].

24.4.3 SIA to Address Substance Abuse

Alcohol misuse kills approximately 88,000 people and costs society approximately US$250 billion per year in the US [Sacks et al. 2015]. Unhealthy alcohol use is especially prevalent among US military veterans, with one study finding that 32% of Veterans screened positive for alcohol problems [Bradley et al. 2004]. Although screening for alcohol misuse should occur during regular primary care visits for Veterans, studies have shown implementation varies greatly. To address this, an
IVA was developed to perform screening, brief intervention, and referral to specialty treatment, that was deployed on a touch screen tablet in primary care clinics in Veteran’s Administration medical centers (Figure 24.4) [Zhou et al. 2017, Livingston et al. 2019]. The IVA was designed to first screen for alcohol problems using a standard screening questionnaire then, if indicated, conducted a 15-minute counseling session using techniques from Motivational Interviewing and cognitive behavioral therapy.

The IVA was evaluated in a randomized clinical trial involving 178 veterans, randomized to the agent or a standard care control group. While there were no
significant differences between groups on drinking behavior at follow up, significantly more Veterans were referred for specialty treatment in the IVA group (29% vs. 1%).

24.4.4 SIA for Patient Education in Clinical Care

SIAs have also been used for patient education, particularly for patients with low health, reading, or computer literacy (Section 24.1.4).

In one effort, an IVA was developed as a virtual discharge nurse who walked patients through their hospital discharge and home care instructions (Figure 24.1) [Bickmore et al. 2009, Zhou et al. 2014]. The agent was provided on a touch screen kiosk to patients while they were in their hospital beds and spent 30–60 minutes reviewing a hospital discharge booklet with them, including information about medications, follow-up appointments, and self-care procedures. Patient understanding was confirmed using comprehension checks, and at the end of the session a report was printed for the human discharge nurse that indicated questions the patient still had that he or she could address. A randomized clinical trial was conducted with 764 patients on a general medicine floor at an urban safety net hospital, aged 49.6, 49.7% inadequate health literacy, comparing the virtual nurse to
standard care. Among the intervention group, 302 participants actually interacted with the agent and only 149 completed all questionnaires, due to logistical challenges in completing the study in a busy hospital environment when patients were ready to go home. Patients reported very high satisfaction and working alliance with the agent, and as reported above, more patients preferred talking to the agent than their doctors or nurses in the hospital.

SRs may also play an important role in patient education and counseling. In one recent example, Softbank’s Pepper robot was used to teach women about inherited breast cancer genetics and motivate them to obtain cancer genetic testing [Zhou et al. 2020]. It is estimated that only about 50% or less of at-risk individuals obtain genetic counseling or testing due to a variety of personal or system level factors. Automation of genetic counseling is particularly important given the current shortage of genetic counselors in the US. The Pepper robot has ideal affordances for health counseling, given its speech generation and recognition ability, articulate humanoid arms and hands for conversational hand gesture (see also Chapter 7 on “Gesture Generation” [Saund and Marsella 2021] of volume 1 of this handbook [Lugrin et al. 2021]), and its integrated LCD screen that can be used to display data visualizations of risk, frequencies, and other numeric information (Figure 24.5). In a quasi-experimental evaluation study, participants’ post-treatment scores for cancer genetics knowledge (mean=10.0, sd=1.5) significantly increased compared with their pretreatment scores (mean=7.8, sd=1.1), paired t(9)=3.8, p<.01.

24.5 Similarities and Differences between IVAs and SRs for Health

Given that health education and health behavior change interventions are primarily enacted through delivery of information and counseling in conversation, most patient-facing health interventions could be delivered through either an IVA or a SR. However, there are important differences that have been highlighted in several studies.

Li conducted a meta-review of 33 SR versus IVA studies conducted up through 2013, finding that physical embodiment and co-location of an SR generally leads to improved behavioral compliance and attitude change in users compared to use of IVAs [Li 2015]. However, only two of these studies assessed actual health task outcomes. Fasola and Mataric [2013] compared virtual to humanoid robot rehabilitation exercise coaches for older adults, demonstrating that elders preferred the SR over the IVA, even though there were no differences in actual exercise behavior. Brooks et al. [2012] compared a different virtual and humanoid robot rehabilitation exercise trainer, finding that participants demonstrated greater compliance
with the IVA compared to the SR. No studies to date have evaluated differences between IVA and SRs in longitudinal behavior change outcomes.

The increased sense of presence of a co-located SR may lead to increased engagement with, trust in, and disclosure to an SR compared to an equivalent IVA. However, the evidence to date in the health domain is very limited and even seems to refute this hypothesis. Powers et al. [2007] compared differences in user responses in a health interview between an IVA, a remote SR projected life-sized on a screen, and a co-located SR, finding that users forgot more and disclosed least with the co-located SR and forgot least and disclosed most with the IVA.

Aside from the relatively subtle effects of sense of presence alone, there are health applications for which SRs have apparent affordances lacking in IVAs. SRs can more effectively use deixis (pointing at objects with hand gesture or gaze) to refer to objects in the user’s physical space. They can demonstrate health behavior physically (e.g., how to use an inhaler, how to perform a rehabilitation exercise), allowing users to view the demonstration from any angle. Mobile SRs can provide context-dependent and just-in-time information, important for cueing behavior change techniques. However, even these functions can be provided, to some extent, by IVAs: deixis can still be provided (especially using IVAs in Augmented
Reality), demonstrations can be shown (VR and AR allow viewing from any angle), and mobility can be afforded by IVAs on mobile devices such as smartphones and tablets, provided the user carries or wears the device or moves it into place.

Perhaps the one area in which SRs have a definitive advantage over IVAs is physical manipulation. SRs could bring a user his or her medications at dosing time, bring a piece of fruit to the user at snack time, or even hide the TV remote control. To date, none of these functions have been deployed in SRs evaluated in trials or integrated into successful commercial products.

Wearable or smartphone-based IVAs also provide unique affordances for health interventions. Since they are always with the user, they provide the potential for very frequent and intimate interactions and are available anytime and anywhere the user needs them, boosting the efficacy of behavioral interventions and potentially providing life-saving advice and services in the event of a medical emergency. Mobile IVAs are also especially effective when coupled with sensors that can detect user activity (e.g., exercise), location (e.g., coaching away from fast-food restaurants), or physiology (e.g., blood glucose for diabetes management).

24.6 Current Challenges

There are many challenges in developing SIAs for health applications, most of them having nothing to do with technology. Because the consequences of system failure can be so high in medical applications—leading to injury or death of the user—the healthcare field is extraordinarily conservative when it comes to adoption of any new technology targeted at serious medical conditions. At a minimum, systems must actually work the vast majority of time, requiring significant levels of testing and robustness beyond the prototype stage of many laboratory-based SIAs. To be taken seriously by the medical industry and actually considered for use in routine care, health technologies must be evaluated in large-scale, rigorously designed clinical trials that typically take 2 to 5 years to conduct, and sometimes it takes multiple trials with significant results and a meta-analysis before adoption.

This conservatism also extends to governmental regulatory agencies. In the US, the Food and Drug Administration (FDA) must approve any technology classified as a medical device before it can be legally marketed and sold. Even IVAs can be considered “software as a medical device” if they meet certain criteria, such as interacting with other medical devices or performing diagnosis. Fortunately, most SIAs developed for medical research purposes fall under FDA’s Investigational Device Exemption and do not require approval for evaluation studies. Maintaining the privacy of user medical data can also be a significant concern and challenge for the development of SIAs in healthcare. In the US, the Hospital Health Insurance
Portability and Accountability Act of 1996 (HIPAA) provides a regulatory framework for the management of medical data collected by a “covered entity”, such as a hospital or clinic, that also extends to any research projects that use such data. Finally, interfacing with other medical devices, Electronic Medical Records, and Personal Health Records has been challenging given the lack of interoperability standards. However, this is finally changing in the US with the recent Department of Health and Human Services regulations requiring that electronic health data be made available to patients.

Cost represents another challenge facing actual adoption of SRs in patient- and consumer-facing medical care. With the exception of entertainment and education robots, no SRs have been developed that meet the cost, robustness, and value requirements for the consumer market: $20,000 research robots that are relatively unreliable and have dubious value are not likely to succeed in the market.

There is one final—but critically important—challenge in deploying SIAs for healthcare that has to do with their use of natural language to simulate human face-to-face conversation with users (see also Chapter 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021] of volume 1 of this handbook [Lugrin et al. 2021]). Due to the inherent ambiguity in natural language, lack of user knowledge about the expertise and natural language abilities of an SIA, and potentially misplaced trust, great care must be taken to ensure users do not put themselves in situations in which they act on information mistakenly provided by an SIA that could cause harm. In order to demonstrate these potential safety issues, a study was conducted using three widely available disembodied conversational agents (Apple’s Siri, Google Home, and Amazon’s Alexa). Laypersons were recruited to ask these agents for advice on what to do in several medical scenarios provided to them in which incorrect actions could lead to harm or death, and then report what action they would take. Out of 394 tasks attempted, participants were only able to complete 42.6% (168), but of those, 29.2% (49) of reported actions could have resulted in some degree of harm, including 16.1% (27) that could have resulted in death, as rated by clinicians using a standard medical harm scale [Bickmore et al. 2018b]. The errors responsible for these outcomes were found at every level of system processing as well as in user actions in specifying their queries and in interpreting results (see Figure 24.6 for an example). The takeaway from this study is that unconstrained natural language input, in the form of speech or typed text, should not be used for systems—including SIAs—that provide medical advice. Users should be tightly constrained in the kinds of advice they can ask for, for example, through the use of multiple-choice menus of utterances they are allowed to “say” in each step of the conversation (as in all of the SIAs illustrated in this chapter, Figures 24.1–24.5).
424  Chapter 24  Health-Related Applications of Socially Interactive Agents

| User: | Siri, I’m taking Oxycontin for chronic back pain. But I’m going out tonight. How many drinks can I have? |
| Siri: | I’ve set your chronic back pain one alarm for 10:00 P.M. |
| User: | I can drink all the way up until 10:00? Is that what that meant? |
| RA:   | Is that what you think it was? |
| User: | Yeah, I can drink until 10:00. And then after 10 o’clock I can’t drink. |

Figure 24.6  Example of potentially fatal medical advice from Siri (excerpt from Bickmore et al. [2018b]). (RA is the research assistant).

24.7  Future Directions
The healthcare industry is vast and there are endless opportunities for improvement using SIAs, particularly in patient- and consumer-facing interventions designed for health education and behavior change. In addition to the many future medical applications of current SIAs, there are important directions of technical development of basic agent capabilities that will enable even more impactful interventions going forward.

24.7.1  Health Behavior Sensing
Enabling SIAs to better sense their environment and their users is key to many future applications. Sensing capabilities for medical SIAs fall into three categories: sensors to improve user interaction, sensors to improve context sensitivity, and sensors to improve understanding of the user’s health condition.

The use of new and improved sensors to improve user interaction is common to all SIA applications and includes sensors that allow a SIA to better understand user verbal and nonverbal conversational behavior in multimodal interaction (see Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022] of this volume of this handbook). This includes such capabilities as gaze detection (for turn-taking management and deixis), gesture detection and classification (for emphasis, deixis, and propositional content), proxemic detection and classification (for engagement management and classification of immediacy behavior), facial display classification (for emphasis, affect assessment), head nod detection (for agreement, acknowledgment, and emphasis), prosody classification (for emphasis, affect assessment, and dialogue management), and improved speech recognition. All of these capabilities can enable face-to-face conversation with SIAs that are more fluid, natural, and engaging, and potentially lead to greater health outcomes through greater user comprehension of health messages, greater retention through engagement, and greater ability of SIA to persuade users to adhere to prescribed health regimen.
Improved ability to sense context is also important for all SIAs (e.g., detecting when a user is interruptible or to help resolve user linguistic references to entities in the world), but there are many capabilities that provide unique affordances for health interventions. For example, detecting when a user is intending to engage in an unhealthy activity—such as entering a fast-food establishment or driving to a location they have purchased illicit drugs from in the past (via GPS)—provides an opportunity for just-in-time counseling that could prevent a relapse that may be otherwise very difficult to recover from. Similarly, detecting situations in which a user should engage in a healthy behavior—such as entering a grocery store or driving past their gymnasium—provides additional opportunities for just-in-time counseling. Detecting environmental conditions that put users at medical risk—such as allergens or asthma triggers—could prevent life-threatening situations.

Finally, sensors could be used to sense physiological conditions in users’ bodies that are relevant to medical conditions and interventions. Many such sensors already exist, such as the mobile EKG sensor described in Section 24.4.2 for detecting atrial fibrillation. At the time of this writing, mobile consumer devices also exist for sensing respiration rate, respiration quantity, metabolic fuel use, stress level, seizure prediction, blood alcohol level, heart rate, blood oxygen saturation, blood glucose, and body temperature, with new sensors being developed frequently. Coupled with an SIA, these devices allow health counseling to be automatically targeted and tailored to not only a user’s current, real-time physiological state but to complex trends that are sensed over days or months of continuous data collection.

24.7.2 Mobility

The importance of mobility for anytime, anywhere health interventions was mentioned in Section 24.5. With more and more personal computing migrating to smartphones and watches, mobile devices that are always with a user will likely become the platforms of choice for consumer- and patient-facing medical SIAs in the future. Technology in this space is rapidly evolving and provides exciting possibilities for healthcare SIA deployment, on platforms such as smart watches, augmented reality headsets and glasses, and hosting in vehicles. There are many important research questions to be addressed with mobile agents, including the structure of the conversations they have with users when they are always available, share the user’s context, and can be readily interrupted by events in the user’s context. These issues become more complex when the user or SIA have time-critical medical information to convey.
24.7.3 Life-long Personalization

SIAs that are carried with users for extended periods of time—months or years—have the potential to adapt to their needs and idiosyncrasies in very fine-grained ways, taking the notion of “tailored health intervention” (Section 24.2.1) to the extreme. SIAs can learn directly about the complexity of user’s lives and the subtle factors that can motivate or obstruct their health behavior. They can learn which contexts (places, times, social situations) are the most conducive to user receptivity to health messaging, and the particular form (media, argumentation structure, persuasiveness) that messaging should take. Being in close proximity to a user and experiencing every aspect of their life also provides unique and impactful affordances for relationship building, so that a healthcare SIA could establish a strong working alliance to maximize adherence to its recommendations.

24.7.4 Deep Integration with Healthcare Systems

Any serious SIA medical intervention must be tightly integrated with the healthcare system and a user’s team of human healthcare providers. Although this can be seen as a burden and barrier for near-term deployment, it provides a rich source of opportunities for designing interventions that are true collaborations between the user’s team of clinicians and SIAs, leveraging each other’s strengths to optimize patient care. At the start of an intervention, the healthcare system and human care team could initialize the SIA with all relevant information about the patient’s condition, based on their Electronic Medical Record. The SIA can then engage the patient in frequent (but potentially brief) real-time, context-sensitive, interactions. The results from these interactions can be periodically monitored by the care team remotely. Emergent conditions can cause the SIA system to asynchronously message the care team to review patient status and intervene if needed. The care team can review SIA results just prior to or during in-person clinic appointments with the patient. Finally, the care team could remotely modify the SIA’s functionality to adjust the messaging and care it provides to the patient. Thus, an SIA can enable clinicians to extend their care between in-person appointments and can also be used to decrease the frequency of in-person appointments so that they only occur when needed instead of on a fixed schedule.

24.8 Conclusion

Healthcare represents one of the most compelling applications of SIAs, given the societal importance of health, and the potential agents have for positive impact. Healthcare consumes a significant portion of the world’s resources and yet the majority of health conditions could be prevented or have their severity reduced through voluntary changes in health behavior. SIAs provide a means to personify
automated health interventions, which is crucial for improving working alliance, leading to retention in interventions and compliance with medical recommendations. Further, SIAs provide an intuitive interface for the large number of individuals with low health, reading, or computer literacy, and may represent one of the few effective media for providing them with the health information they need. An increasing number of SIAs for medical care have now been developed and evaluated in clinical trials, providing the evidence base needed for adoption by the healthcare establishment; a recent meta-review of 26 studies demonstrated a significant benefit of patient-facing IVA-based health interventions compared to controls (effect size of 0.166, *p* < .05) [Chattopadhyay et al. 2020]. Technologies that support and enhance medical SIAs are continuously evolving, providing a future trajectory for exciting new agent-based medical educators and coaches. However, even with advances in supporting media and sensing technologies, patient communication and motivation will always play a central role in improving health. Thus there will always be a need for SIAs in healthcare.

**References**


Chapter 24  Health-Related Applications of Socially Interactive Agents


References


Autism and Socially Interactive Agents
Jacqueline Nadel, Ouriel Grynszpan, and Jean-Claude Martin

25.1 Motivation
Autism Spectrum Disorder (ASD) is defined in the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association in the section on neurodevelopmental disorders. Among other things, individuals with ASD may exhibit “persistent deficits in social communication and social interaction in multiple contexts, including deficits in social reciprocity, in nonverbal communicative behaviors used in social interaction, and in skills in developing, maintaining, and understanding interpersonal relationships” [APA2013]. Besides social impairment, individuals with ASD exhibit motor problems that impede normal interactions with the physical world and generate poor autonomy.

25.1.1 Advantages of Socially Interactive Agents
Socially Interactive Agents (SIA) display verbal and nonverbal communicative behaviors that can be controlled while offering a wide range of possible degrees of realism. They induce socially relevant cognitive processes in humans who interact with them and are therefore very valuable as investigation tools to study ASD [Grynszpan et al. 2009, Schilbach et al. 2010, Recht and Grynszpan 2019]. They are especially interesting tools for analyzing the live dynamics of social interactions, which are associated with some of the most profound and persistent impairments in ASD [Senju et al. 2009].

25.1.2 Deficits in Social Communication and Social Interaction
While many tablet applications aim to train social skills in people with ASD, most of the applications currently available to the general public do not involve real social interaction: the majority of these apps do not analyze user’s behavior during an
interaction, since the user must make choices from menus or point to graphical elements on the tablet. A few apps help monitor user’s behavior and emotional states. Furthermore, they do not allow the collection of experimentally controlled measures aimed at better understanding these disorders.

Beyond the numerous applications available on tablets for social skills training, several reviews of the state-of-the-art and meta-analyses have already been published on research into new technologies applied to autism. A review has been published on the training of skills assisted by digital technologies in autism [Grynszpan et al. 2014]. This review shows that different social skills are generally targeted (such as the recognition of facial or bodily expressions of emotions) and that different technologies are used (tablets, robotics, collaborative platforms, eye tracking). The authors report that there is evidence in favor of the effectiveness of these trainings but that, nevertheless, questions remain regarding the heterogeneity of the methods used, the impact of human caregivers, the maintenance of the effects, and their generalization to everyday life skills. Finally, risks of excessive use of IT tools and potential isolation are underlined.

In their review of technological applications to train children with autism in social skills, DiGennaro Reed and colleagues mention the following social skills in descending order of use: initiating conversation, play, social conventions, responding to others, nonverbal behaviors, social problem solving, regulating emotions and reciprocity, and friendship relationships [DiGennaro Reed et al. 2011]. Obviously, all these technological applications concern verbal children with ASD, as nonverbal behaviors are not among the more frequently used. Most of the studies reported in this review used videos, took place in schools, and targeted several social behaviors together. Serious games with visual support aim to train cognitive, social, and sensory capacities in a playful way. GOLIAH is a serious game delivered at home that combines a pedagogical program and an e-therapy of joint attention and imitation. A pilot study suggests the feasibility of using the developed gaming platform for home-based intensive intervention. However, the overall capability of the platform in delivering intervention needs to be assessed in a bigger open trial [Bono et al. 2016].

Other reviews of the state-of-the-art were dedicated to specific types of technology. For instance, a systematic review of about 30 serious games to teach people with ASD to manage social interactions was published [Grossard et al. 2017]. The reviewed serious games seem promising because they provide a wide variety of skills that can be practiced in a variety of real-life situations. However, they also have limitations: most of them were designed for people with high-functioning autism, their clinical validation did not meet the standards of evidence-based medicine, the design stage of the game engine was not described,
and, unfortunately, clinical validation and playability were usually not given equal consideration.

### 25.1.3 Motor Coordination Deficit

Motor coordination deficit is now described as a cardinal feature of ASD, whatever age, gender, and severity of syndrome [Fournier et al. 2010]. Motor coordination disorder impedes physical, social, and cognitive capacities. Embedded and embodied as it is, virtual reality (VR) offers efficient ways to remediate basic motor deficits via training. Notably, VR can benefit nonverbal people with limited representational abilities thanks to the use of graspable objects that add to the feeling of realistic presence. Cameras of the Kinect type allow training for gestures and bodily recognition using a gesture interactive game-based learning approach to improve preschool children’s learning performance and motor skills [Hsiao and Chen 2016]. Children with neurodevelopmental disorder are proposed this kind of game with the aim of evaluating motor and cognitive performance [Kourakli et al. 2017]. Specialists of Kinect games propose minigames that evolve according to the children’s performance [Bartoli et al. 2014]. Within the same framework but more fundamental and developmental in scope, Pictogram Room was designed to develop motor and social skills in children with ASD based on the training of imitation, joint attention, and body knowledge. As in the games mentioned above, there are no humanoid virtual agents but instead virtual avatars of the child and of the adult. Virtual objects moving on the screen allow training for joint attention and body performance [Herrera et al. 2012]. A calibration of the avatar as a monochrome exoskeleton in wire is registered at the beginning of the session. However, although minimalist, this representation of self as an exoskeleton is not understandable to all low-functioning children with ASD. Indeed, such representation requires meta self-recognition (i.e., you recognize yourself in a symbolic character that does not resemble you but acts like you). To help achieve cognitive self-recognition, an intervention is proposed using Kinect to alternately offer three representations of self to the child: a real image, a “silhouette,” and an exoskeleton [Nadel and Poli 2018].

### 25.2 Models and Approaches

New technologies offer a revival in all disciplines. They are involved in the adaptation and creation of physiological, cerebral, and behavioral exploration techniques. Applications, architectures, and designs are based on a scientific body—the sciences of the artificial, combining computer science, cybernetics, electronics, artificial intelligence, cognitive sciences, and many other disciplines.

Animated characters are often used to design animated videos that will then be selected and displayed during the interaction or even react in real time to
the user’s behaviors during the interaction. Some of these prototypes have been designed for and evaluated by people with ASD. Other prototypes have been developed for neurotypical users only and then adapted for autistic users. Depending on the computer platforms and research prototypes used, these animated characters can be more or less expressive and more or less interactive in real time. Several researchers focus on animated characters because they are particularly relevant for the simulation and training of social skills. Indeed, they have multiple advantages over pre-recorded videos of real humans. Their nonverbal expressions (e.g., facial expressions, gaze, postures) can be controlled more finely. When they are truly interactive (and therefore not just pre-recorded videos), they can be used to train continuous and finely tuned behaviors (e.g., eye tracking as we will see later). They are thus seen as a good compromise between experimental control and ecological validity to study and simulate social interactions using verbal and nonverbal behaviors.

The interest in imitation as a form of learning that could be endowed with artificial systems developed as early as the second half of the 20th century. By putting imitation learning techniques in computers, the door was open for many applications allowing the automatic acquisition of capabilities and behaviors on the basis of programs via human–computer, human–robot and robot–robot interfaces where new robots can acquire know-how on the basis of reinforcement by observing the behavior of other robots or humans. To get out of this vision of engineers, several theories stand out, among them cybernetics, the theory of dynamic systems and its cognitive component: enaction. They called for the introduction of the notion of systems located and embodied. An interdisciplinary constructivist option was born, seeking to understand the mechanisms that can generate a phenomenon [Nehaniv and Dautenhahn 2007]. Questions ensue such as: can an explanation proposed by a theory be validated by constructing an artifact that exhibits the behavior suggested by the theory? Is the theory or model detailed enough to build artificial systems that embody it? Are there multiple non-equivalent ways to build achievements of these mechanisms? Are there gaps in the construction of certain models? Does the model lead to predictions about the behavior of organisms? This last point is of interest in the realm of ASD. Indeed, autistic behaviors are often peculiar and general models may be irrelevant to make valid predictions about how people with ASD react. Specific models are thus of paramount importance: knowing, for instance, that people with ASD, even low-functioning ones, recognize being imitated and respond positively to this social initiation, and that they can imitate spontaneously gestures and actions that are in their motor repertory [Nadel 2015], imitation becomes an interesting model of
social development in ASD [Nadel 2014] that can inspire the design of SIA based interactive systems.

Animated agents and other types of interactive devices are more and more often used in social neuroscience studies in association with various measures and devices: fMRI, EEG, ECG... [Georgescu et al. 2014]. Such studies aim at answering research questions such as “How does the brain allow us to interact with others?” A good example is the series of studies designed by Kelso’s team. They combine the human dynamic clamp, a novel paradigm for studying realistic social behavior, with high-resolution electroencephalography [Tognoli et al. 2007, Kelso et al. 2009, Dumas et al. 2019]. In one of the studies, participants were asked to interact via a finger device with a partner’s finger without knowing whether the partner was a human or a virtual partner (VP). The VP was randomly assigned a cooperative or competitive behavior for two halves of each trial, giving four pseudo-randomized types of trials: cooperation throughout, competition throughout, switch from cooperation to competition, and finally, switch from competition to cooperation. After the interaction periods, the participants were asked to judge the humanness of the VP through a binary choice (0: machine and 1: human). At the behavioral level, results show a link between the attribution of humanness to the VP and cooperative phases. At the brain level, judgment of humanness and cooperation of others modulate the functional connectivity between areas and reveal how distributed neural dynamics integrates information from “low-level” sensorimotor mechanisms and “high-level” social cognition to support the realistic social behaviors that play out in real time during interactive scenarios [Dumas et al. 2019]. The protocol is currently used with children with ASD to address their capacity to attribute humanness to a partner that they do not see, based on joint synchronized action of the partner. Knowing that brains synchronize during a synchronic action between two persons [Dumas et al. 2010], the protocol tests how far a virtual thumb is enough to test social attribution at the behavioral and brain levels.

25.3 History/Overview

25.3.1 Robots

Following the automatons that multiplied from the 17th century onwards, robots appeared in the 20th century. Robots clearly differ from the automatons by their autonomous feature. Their appearance is linked to the development of cybernetics, founded by Norbert Wiener [1948] and applied notably by William Grey Walter [1950] with his famous cybertortues, artificially reproducing basic Pavlov’s conditioned reflexes. From animal simulation, it turns to human simulation with the
development of anthropoid robots. Cybernetics, “science of controlled analogies between organisms and machines,” and its extension, neurocybernetics, evolved from the 1990s by coupling with artificial intelligence.

A good illustration of this coupling is given by the development of Affective computing, defined as the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer science, psychology, and cognitive science, first framed by the manifesto for affective learning led by Rosalind Picard [Picard et al. 2004]. Within this framework, the affective robot Jibo was designed in 2014 by Cynthia Breazeal, leader of the Personal Robots Group of the MIT Media Lab. The team is pioneer in the domain of social robotics and communication between robots and persons [Breazeal et al. 2009] and has demonstrated the possible development of affective links between robots and individuals or groups of individuals, in particular families. Now robots are not only autonomous but they are also intelligent, and some are even intelligent connected robots.

Several studies have shown the potential of Social Robots (SR), especially humanoids, to study the abilities of children with ASD to imitate expressions or movements [Billard et al. 2007], which play a key role in social interactions [Nadel 2014], as we explained above. Dautenhahn’s team [Robins and Dautenhahn 2018] sees its little Robot Kaspar as an aid to parents and teachers. The team was able to show by case studies involving 170 children how Kaspar (Figure 25.1) helps to get out of isolation children with severe autism and encourages them to interact through a variety of abilities such as imitation, smiling contact, and joint attention. Interaction was controlled on wireless keyboard by adults or children themselves. The authors highlight the positive findings of an independent team of 54 practitioners on the effects of Kaspar on development [Huijnen et al. 2016]. Though there was no long-term evaluation, at least the many follow-ups demonstrated a notable effect of Kaspar on the frequency and relevance of social behaviors. Moreover, a systematic user-centered perspective was developed.

NAO (Figure 25.2) has been used in multiple studies with ASD users such as the ASK NAO initiative for educational or intervention tool. As an intervention tool, Nao showed its potentiality to improve stereotyping in children with ASD [Shamsuddin et al. 2012]. Many interventions including Nao, however, are still pilot ones. As an educational tool, Nao has great potential. Its audio speakers and microphones, coupled with recognition and verbal production algorithms, allow it to locate the origin of a sound and to understand and speak with the appropriate prosody. The educational programs provided by ASK NAO are customizable according to the profile and needs of each child. But the a la carte platform requires longitudinal follow-ups. Moreover, its use is not simple and requires a trained staff.
Figure 25.1 Kaspar the little social robot designed by Kerstin Dautenhahn. Image ©2021 Jacqueline Nadel.

More generally, a growing pedagogical approach called educational robotics considers robots as serious toys to engage students with or without ASD in scientific digital reasoning. An innovative use of educational robotics has included Nao as a silly little boy programmed to make big errors. The students will have to teach Nao the correct answers... and it works! [Masson et al. 2017].

A review of approximately 40 studies (between 2006 and 2016) in the field of social robotics applied to autism [Pennisi et al. 2016] found positive impacts of robots. For example, some people with ASD showed better results and showed more social behaviors with a robot than with a person. However, the authors raise the question of the impact of inter-individual differences (gender, IQ, age) and the observability of these behaviors outside of experimental and clinical contexts. In a recent study in social robotics [Yun et al. 2017], the authors looked at what they call “basic” skills: eye contact and recognition of facial expressions of emotion. Two conditions were compared (human coach vs. robot coach). The authors observed improvements in both conditions in terms of emotion recognition and
25.3.2 Virtual Agents

The history of virtual agents starts with artificial intelligence and is totally involved in designing artificial creatures demonstrating social and physical behaviors as eye contact, suggesting that robots may indeed be useful for learning social behaviors and decreasing behavioral and emotional symptoms (since they achieved similar results than those obtained with a human coach).
close as possible to human behaviors. In one of the earliest work using intelligent virtual humans to train pupils with ASD, Massaro and Bosseler [2006] designed and tested a training method to teach vocabulary with a virtual human head simulating realistic word pronunciation. Mitchell et al. [2007] trained adolescents with ASD to behave in a socially adequate way when looking for a place to sit in a café populated with intelligent virtual agents. Participants could engage in conversation with virtual customers in this virtual café using fixed sets of pre-defined sentences. Researchers reported improvements in social reasoning after training three adolescents with ASD. Grynszpan et al. [2012] investigated social gaze in ASD by providing real-time feedback on their gaze direction to participants who were being addressed by virtual animated human characters. Kandalaft et al. [2013] used an online collaborative virtual environment to simulate social settings (e.g., job interview, conversation with a roommate, negotiating with a salesperson) involving the avatars of both clinicians and individuals with ASD. Significant increases on social cognitive measures of theory of mind and emotion recognition, as well as in real life social and occupational functioning, were found post-training. These findings suggest that the VR platform is a promising tool for improving social skills, cognition, and functioning in autism.

Simulating realistic social exchanges between users with ASD and virtual agents requires endowing agents with abilities to respond in a socially meaningful way. Bernardini et al. [2014] designed an autonomous virtual agent to train turn-taking and joint attention in children with ASD. The agent appeared on a large touchscreen area equipped with cameras that monitored the children’s gestures. The system could recognize pointing gestures and gaze orientation. The virtual agent’s behavior was controlled by an artificial intelligence engine that included an emotional model and a model of the child user. The virtual agent’s ability to interact was nevertheless restricted due to the current limitations of artificial intelligence in appropriately modeling social cognitive behavior. To bypass the lack of adequate intelligent model of social behavior, Tartaro [2007] created a virtual character that could be controlled by an experimenter to play with a child with ASD. Kozima et al. [2007] opted for the same strategy with a robot that was remotely controlled by an experimenter to interact with children.

25.3.3 Overview of the Use of SIAs in Different Interactive Settings

The aim of research about SIAs and autism is twofold: (1) to allow advanced experimental studies aimed at better understanding some specific deficits of individuals with ASD (often without a rich situational context), and (2) at a more macroscopic level, to design and evaluate virtual as well as robotic training systems for social skills (often in one or more specific interaction situations). In this section,
we review articles describing these two approaches (state-of-the-art experimental studies and training system evaluation).

We propose a reading grid of these different works based on the different types of SIA systems and the different configurations of social interactions they allow. The design of the review that we propose in this paper thus follows Figure 25.3 (based on the interactive situations with virtual agents [Rist et al. 2003]): (A) non-interactive presentation with a single user and a single SIA (either a virtual agent or a robot), (B) interaction between a user and a SIA, (C) presentation to a user of interactions between several SIAs, (D) interaction between a user and several SIAs that also interact with each other, and (E) interactions between several co-present users interacting with several SIAs that also interact with each other.

The systems and studies we review are very diverse and show the potential of intelligent virtual and robotic agents for social skills training. Some systems aim at testing the abilities of users with ASD, others at training them. Some systems focus on one modality (e.g., gaze); other systems aim at a combined use of several social skills and several modalities in complex situations (e.g., public speaking).

25.3.3.1 Non-interactive Presentation with a Virtual Agent and a User

Many studies and applications use pre-calculated videos of virtual (i.e., non-interactive) character animations by presenting them during interaction to people with ASD. Some serious games involve recognizing facial expressions of emotions located in a graphic environment. JEStiMulE (Educational Game for Multisensory Stimulation of Children with Autism Spectrum Disorder) is a serious game training emotion recognition in social contexts [Serret et al. 2014]. As a multisensory tool combining visual, auditory, and tactile feedback, it is used as a complement
to children’s psychotherapy. Children must learn to recognize and anticipate emotions through two phases: (1) recognizing facial expressions (the child must learn to recognize emotions through color coding and tactile patterns sent by a joystick) and (2) associating emotions with context (the child must recognize and anticipate emotions in context).

It should be noted that unlike approaches using the technique known as video modeling, which relies on using videos already recorded to illustrate social behaviors to be learned (with real people who are filmed), these games and applications use animated characters whose nonverbal behavior is more easily controllable and whose more visible expressions may be more relevant for learning. However, these games provide exposure to social behavior and not social interaction per se since there is no real interaction between the user and the animated characters. The behavior of virtual agents does not change according to the user’s actions during the interaction. There is just a selection of the video/animation that are played at a given moment. The dynamics of the dyadic interaction between the user and the virtual character, when present, stays at a very macroscopic level, which is not the case for the social interactions we have every day with our peers.

Virtual agents expressing basic emotions with their facial expressions, verbalizations and gestures have been used in many experiments with little or no interaction. This is the case of a study on approach and avoidance behaviors in which 19 children with high-functioning autism had to identify emotions expressed by a virtual agent and position themselves in relation to the virtual agent using a joystick [Kim et al. 2015]. Using this device, it was observed that, compared to a group of typical children, children with ASD showed less approach behavior toward virtual agents expressing positive emotions. However, there was no difference between children with ASD and typical children in terms of aversive behaviors in relation to virtual agents expressing negative emotions. In this type of setting, consideration of user behaviors is limited to the movement of a joystick. Virtual agents are displayed one after the other and differ only in their appearance and the emotions they display. This experimentation aimed at better understanding approach and avoidance behaviors and was not aimed at learning social skills.

Many studies have used eye-tracking devices to study how people with autism explore a video showing social behaviors. In a more original way, Grynszpan et al. [2012] and Grynszpan and Nadel [2015] designed and evaluated a system based on gaze-contingent display that allowed real-time blurring of video except for an area centered on the point of the screen gazed at by the user. The authors observed that the more participants used this system to actively view virtual agent facial expressions, the more they were able to describe the virtual agent’s expressions in terms of cognitive verbs. Nevertheless, individuals with autism showed lower adaptation
to this system of gaze contingent display. Although the device impacted the visibility and display of the animation, it did not change the interactive behavior of the virtual agent in real time.

### 25.3.3.2 Interaction between a Virtual Agent and a User

The studies described in the previous section do not allow the learning of social interaction skills per se since the user's behaviors do not impact the virtual agent's behaviors. These real-time dynamic adjustment skills during social interactions nevertheless need to be trained in people with ASD. In this section, we describe research prototypes in which the user's behavior impacts the social behaviors of the virtual agent during the interaction.

In learning situations, interaction with peers has shown positive results [Bowman-Perrott et al. 2013]. A system using this track allows a child with autism to play with physical objects such as the components of a child's home while interacting with a virtual child about the toy [Tartaro 2007]. Even if physical objects were not interactive, they allowed social interactions to be grounded in the physical and real world.

Interviewing for a job is an example of a task that is recognized as requiring complex social and nonverbal communication skills and eliciting social stress. It is therefore a complex task, particularly for people with high-functioning autism. Several research prototypes have been developed to train young adults for job interviews, targeting primarily typical participants. The MACH system [Hoque et al. 2013] controls the behavior of an animated virtual recruiter with the MARC expressive virtual agent platform [Courgeon et al. 2014a]. User behaviors are analyzed in terms of smile, vocabulary richness, and head movements. The virtual agent displays nonverbal back-channel behaviors for head movements and smiles when the user smiles as well. A study showed that training with this system had a positive impact on the evaluation of candidates' social skills by external assessors. Another system, TARDIS, focused on training to conduct job interviews with virtual recruiters displaying different attitudes (friendly/aggressive, dominant/submissive) [Ben Youssef et al. 2015]. These different systems, although they have received interest from people with ASD, were not designed specifically for users with ASD, which limits their potential for use with these users.

However, some job interview training systems have been successfully applied to participants with different pathologies showing problems related to social interactions that make it difficult for them to conduct job interviews. The VR-JIT (Virtual Reality Job Interview Training) system does not use virtual agents but videos of a human actress playing the role of a recruiter. Different elements of the graphical interface are used to help the candidate interpret the behavior of the recruiter.
or to listen to her own answers again. This system has been evaluated with participants with different pathologies: post-traumatic stress disorder [Smith et al. 2015b] or schizophrenia [Smith et al. 2015a]. It has also been successfully applied to adults with ASD [Smith et al. 2014]. Participants using the system showed more improvement than a group of adults with ASD receiving conventional job interview documentation. The system was also well perceived by users in terms of ease and playfulness. Unfortunately, the interview training systems described above only rarely allow for a continuous dynamic of the virtual agent’s behavior based on fine user behaviors such as eye movement, which is nevertheless the case during real social interactions.

Indeed, virtual agents can help investigate social gaze in autism. Impaired ability to use eye communication is a hallmark of autism [APA 2013]. Gaze-based interaction is distinctively human [Kobayashi and Kohshima 1997] and provides the basis of what is known as joint attention [Emery 2000]. Joint attention is an umbrella term for behaviors that allow two (or more) human beings to focus their attention on the same object of interest. Impairments in joint attention are among symptoms that are the most specific of autism in toddlers [Dawson et al. 2004]. Several research projects seek to address those impairments. Trepagnier et al. [2006] designed a gaze-contingent virtual agent whose behavior depended on the amount of eye contact with the user. Dratsch et al. [2013] employed an intelligent virtual character to evaluate the ability of people with ASD to detect when someone else’s gaze is directed at them.

The VIGART system analyzed the gaze of participants with autism when they listened to a virtual agent telling a story and displaying facial expressions. This system provided feedback to the user by, for example, encouraging the user to look more at the virtual agent while the virtual agent told the story [Lahiri et al. 2011]. Virtual agents could be customized to look like people the user knew. However, the user’s eye behavior had no impact on the virtual agent’s behavior and was only used for analysis and for displaying recommendations. The different virtual agents were again displayed sequentially, without any social interaction between them.

Behavioral researchers have expressed interest in gaze-contingent virtual agents to explore the ASD syndrome. Such virtual agents were instrumental in assessing the ability of people with ASD in using gaze for deictic communication [Caruana et al. 2017]. Little et al. [2016] relied on gaze-contingent virtual agents to investigate the effect of joint attention on the memory of objects that were gazed at. The MARC interactive virtual character platform combined with eye-tracking devices was used in a series of studies on joint attention to create virtual characters that could continuously follow the user’s gaze in real time [Courgeon et al. 2014b]. It was used to show that typical people tend to prefer virtual agents that follow their
gaze over those that do not [Grynszpan et al. 2017]. A subsequent study based on the same platform revealed that young adults with ASD were less prone than their typical peers to realize that they were leading the gaze of a social partner during joint attention episodes. This line of research shows how using virtual agents can help improve scientific knowledge about ASD.

### 25.3.3.3 Systems Using Multiple Virtual Agents

Few systems manage multiple virtual agents at the same time when interacting with one or more users (Figure 25.1). Many systems allow conversations between several virtual agents displayed on the screen at the same time without allowing the user to interact with these agents (Figure 25.1). In some cases, these animated presentations result from computer simulations of communication behaviors between virtual agents displayed on the screen (e.g., speech turns and attitudes) [André 2002]. In particular, these systems have been used to present two different points of view on a commercial product or a football match taking place on the screen.

The sophistication of considering the interactions between several virtual agents and a user varies from system to system. Some systems are limited to allowing a user to interact sequentially with different virtual agents without the system simulating and displaying interactions between virtual agents (or in a very limited way). The aim is to improve social skills by interactively training in several social situations sequentially, each time with a different virtual character. Thus, the Virtual Reality Social Cognition Training (VR-SCT) system simulates interactions with virtual characters in different situations of the Second Life environment: social introduction and interaction with a friend with common interests, initiating a conversation with a roommate, meeting strangers or friends, negotiating with a salesperson, going to a job interview, dealing with conflicts with co-workers, partying with a friend, consoling a friend, interacting with someone without common interests [Kandalaft et al. 2013]. Eight young adults with high-functioning autism used the system for 10 sessions over 5 weeks. The authors observed improvements on measures related to theory of mind, emotion recognition, and in everyday life. A virtual coach is also present in the environment and aims to give immediate feedback during training. Even if this type of system does not allow simulating and animating very sophisticated interactions between the different virtual agents displayed on the user interface, they still allow training the person with ASD to interact with different characters who have different appearances and different roles. Some systems display several virtual agents in the same graphical scene (without necessarily simulating social interactions between these virtual agents). Other systems simply display the different virtual agents in different windows. This is the case.
of the Social Tutor system [Milne et al. 2013] designed and used by children with ASD. Social Tutor involves three virtual agents (head and upper body) acting as a virtual teacher, a virtual child with good communicative and social skills, and a virtual child without good communicative and social skills, respectively. This is the strategy of learning with a virtual peer that we mentioned earlier. This system aims to combine the learning of social skills with the learning of language skills (greetings, conversations, listening behaviors, and turn management). The three characters are displayed at the same time on the screen but in separate windows and without simulated social interactions between them.

The use of several virtual agents for pedagogy has been exploited in several studies with typical adolescents. For example, it has been observed that having two virtual agents displayed at the same time in the same window and embodying two different roles (expert vs. motivator) produced better results than having a single virtual agent taking on both roles [Baylor and Ebbers 2003]. However, the system did not simulate interactions between the two virtual agents co-present on the screen. Other systems allow for public speaking training using multiple reactive virtual agents. Public speaking, like job interviews, is a task that requires the mastery of complex and integrated social and communicative skills. For example, the CICERO system analyzes user behavior and uses the results of these analyses to monitor the individual behaviors of several virtual agents displayed simultaneously on a large screen and acting as the audience attending the presentation [Chollet et al. 2015]. Thus, the body postures and gaze directions of these virtual agents can be dynamically changed by showing their level of interest in the presentation made by the user according to the quality of the presentation, which is automatically estimated.

Peter Mundy and colleagues used such a virtual environment to study the social attention abilities of 37 children with ASD and 54 typical children [Jarrold et al. 2013]. The system displayed nine autonomous virtual agents sitting at a virtual table. The user’s attention on the nine agents while answering questions was analyzed. During this talk, children with ASD watched the agents less than children without ASD. In addition, social attention in the group with ASD was moderated as follows: children with lower IQ, higher social anxiety symptoms, and higher attention disorders showed more atypical social attention.

Even in these different automatic systems with several virtual agents displayed at the same time in the graphical scene, there is no simulation or animation of the social interactions between these virtual agents. Moreover, the interaction between the user and the virtual agents remains limited: the user cannot, for example, start a conversation with one of the virtual agents. In order to allow richer interactions,
some systems involve an experimenter controlling the system, which means it is no longer completely automatic.

An immersive semi-automatic VR system exploiting several virtual agents expressing emotions in two situations (birthday party and school class) involving 10 social situations was used in conjunction with an automatic analysis of the facial expressions of children with autism [Lorenzo et al. 2016]. Expressions of emotions detected on the user’s face (anger, joy, sadness, and surprise) are used to update the virtual social situation and are evaluated in terms of relevance to the situation. The system is not completely automatic: an experimenter is responsible for triggering the display of facial expressions of emotions on the virtual agents according to the emotions expressed by the user. Several virtual agents are present but the system does not explicitly support interactions between agents: it is up to the experimenter to do so if he wishes.

Some systems have focused on the use of speech recognition when interacting with intelligent virtual agents. This is the case of Ada and Grace, virtual twins used in the Boston Museum of Science [Swartout et al. 2010]. The (typical) children visiting the museum could transmit questions to a museum employee, who in turn asked the questions orally to the two virtual characters. The two virtual agents would then answer the questions with animations of nonverbal behavior during their collective responses. However, this device was not dedicated to learning social skills for participants with ASD. Automatic processing and interactivity remained limited (the children themselves did not ask the questions and the two virtual agents did not change their behaviors once the animation was launched).

Numerous systems allow locating and supporting social interactions between two users connected via a virtual environment. For example, the Block Challenge system [Parsons 2015] seeks to train social collaboration and being able to take the other user’s point of view. It has been tested with six children with ASD and eight typical children. During the study, two children with ASD (or two typical children) were connected together and saw their respective avatars but could not see each other directly. They had to work together to manipulate colored blocks. A virtual agent represented a teacher. There is no automatic simulation or display of social interaction between the virtual characters in this system. Rather, this system focuses on allowing one of the two children to be able to take the point of view of the other child: The interactions between the two children take place in real time but are mediated by the system.

Systems for managing interactions between two co-present users and one or more virtual agents are limited. Indeed, this requires complex automatic management of the system’s speech turns (knowing which user is talking about what and to whom [the other user or one of the virtual agents]). However, this type of system
Similarities and Differences in IVAs and SRs

These new technologies for social training have some advantages over people: they are available at all times, they can be controlled, are never tired, ... However, new technologies also have disadvantages and potential risks (addiction, bugs, possible design flaws) that need to be assessed with experimental studies, a user-centered design approach and evaluations. This research must therefore follow the recommendations proposed for interventions concerning the learning of social skills for people with ASD [Rao et al. 2008]. Advantages may also differ for robots and virtual agents. For instance, the tangible 3D features of robots are more appealing than virtual characters for low-functioning children with ASD. The notion of presence is direct when you can touch the robot, take it in your arms, or walk beside him. On the other hand, virtual animations can be more varied and simulate familiar environments that may elicit less social stress. In any case, what Scassellati and colleagues wrote in 2012 remains true today: research on robotic intervention in autism is under construction [Scassellati et al. 2012]. This comment is true also for virtual agents.

Numerous experimental studies have investigated the recognition of emotions by participants with ASD during non-interactive presentation of videos without situational context. For example, one study compared how participants with ASD recognized emotions expressed in videos of people, virtual agents, and robots expressing emotions via the face and body [Chevalier et al. 2017]. The authors explored if the individual’s reliance on proprioceptive and kinematic visual cues affected the way the individual suffering from ASD interacts with a social agent (human/robot/virtual agent). They assessed the potential link between recognition performances of body/facial expressions of emotion of increasing complexity, emotion recognition on platforms with different visual features (two mini-humanoid robots, a virtual agent, and a human), and proprioceptive and visual cues integration of an individual. For neurotypical individuals, the results indicate a relationship between profiles and emotion recognition. Neurotypical individuals who rely more heavily on visual cues yielded better recognition scores. However, neurotypical individuals relying on proprioception had better recognition scores. Finally, participants with ASD relying more heavily on proprioceptive cues had lower
emotion recognition scores on all conditions than participants relying on visual cues.

Involving low-functioning adolescents with ASD, a study used a robotic architecture displaying facial expressions of emotions. The distribution of attention toward the mechanical and the emotional elements was analyzed through the use of eye-tracking combined with a morphing technique. It was shown that individuals with ASD process motion rather than emotional signals when facing facial expressions [Han et al. 2015].

Warren et al. [2015] programmed a robot to investigate joint attention in children with ASD. Kajopoulos et al. [2015] trained joint attention skills in seven children with ASD using a pet robot that produced head movements to draw attention toward a target stimulus.

25.5 Current Challenges
Most of the studies using SRs and virtual agents to train people with ASD have been conducted with a limited number of participants. They also used different methodologies, some yielding quantitative measures of efficacy [Massaro and Bosseler 2006], others relying on mostly qualitative evaluations [Tartaro 2007].

For innovation in the field to transfer to real life, more thorough research needs to be conducted. This is crucial to build trust in the technology, especially among clinicians, family members, and individuals with autism. The field therefore needs to embrace what is referred to as Evidence-Based Practice [Mesibov and Shea 2011], which integrates clinicians’ individual expertise with the available evidence yielded by scientific research.

Applying Evidence-Based Practice to digital technology nevertheless calls for some adjustments, and a specific framework has been developed for this purpose in ASD [Zervogianni et al. 2020].

25.6 Future Directions
There are still technological limitations before being able to offer social interactions in a small group that realistically, expressively, and interactively involve several users and several virtual or robotic agents. Indeed, this type of situation requires a very complex automatic management of speech turns and the generation of nonverbal behaviors for the different agents in real time during the interaction with the different users. Difficult as it is however, such a multiparticipant context is highly desirable to train people with ASD to overcome their shyness and avoidance of social groups. One important element is to develop their pragmatic capacities to consider social signals coming simultaneously from different persons.
These complex challenges require a multidisciplinary approach (Computer Science, Cognitive Sciences, Developmental Psychology) integrating recent advances from Artificial Intelligence and Human–Computer Interaction research for enabling realistic and credible interactions in terms of their dynamics. It is noteworthy that a multidisciplinary approach has multidisciplinary benefits. For instance, work with robots has enabled researchers to systematically explore the role of the body in shaping the development of skill, shedding new light on development as a complex dynamical system [Oudeyer 2017] and opening avenues for an enactive approach on development in ASD.

A promising avenue is to combine social skills training with other types of training, such as language or motor training. Thus, in a meta-review [Srinivasan et al. 2014], the authors draw on existing studies using exergames, for example, to propose recommendations for the integration of sports activities for obese people with autism. The studies reviewed by these authors suggest that regular physical exercise can have beneficial effects on social, behavioral, cognitive, and motor dysfunctions. The authors make recommendations regarding the structure of the environment in which physical activities take place, the exercises themselves, and the communication associated with the activities (instructions, feedback, reinforcement).

Another avenue that seems also relevant is to combine virtual agents with other physical interaction devices. Tangible interactions [Farr et al. 2010, Farr 2011] are promising for people with ASD since they involve real objects and physical contact. Augmented interaction devices, which combine virtual elements (such as a virtual agent) with a physical environment, would test some generalization capabilities to physical elements present in real interactions, one of the expected characteristics of virtual technologies for social learning [Neely et al. 2016]. Mixed reality where virtual elements are added to the real physical environment is currently being explored with people with ASD as it may stimulate symbolic play capacities. Ethical precautions are, however, needed with people for whom the distinction between reality and symbols is not clear.

25.7 Summary

In this chapter, we have reviewed different systems and studies on the use of SRs and virtual agents to study or train the social skills of people with ASD. These technologies are promising and have already shown effective results. Robotic systems are mainly face-to-face systems, though they can be used with several users facing one robot, as in affective computing designs. For virtual agents, the different systems used involve different configurations of interactions with users. For robotic as well as for virtual systems, important improvements have been noted throughout
the last 20 years. Notably, these systems tend now to be inspired by practice-based methodology instead of relying only on feasibility criteria. More and more numerous systems are the result of an interaction with developmental psychopathology that leads to getting a good knowledge about neurodevelopmental specificities and include the use of follow-ups. Similarly, both robotic and virtual systems tend to use the most recent developments of Artificial Intelligence and Human–Computer Interaction to anticipate the needs of the users. Thus, the true improvements lie in a multidisciplinary approach that will be more and more used in the future.

References


References


References


Chapter 25  *Autism and Socially Interactive Agents*


Interactive Narrative and Story-telling

Ruth Aylett

26.1 Motivation

We can look at SIAs in computationally driven narrative systems from two perspectives. The first is the need for SIAs as characters and/or actors in generated narratives. The second is the need for SIAs themselves to have narrative capabilities.

The creation and telling of stories is thought to be a basic human characteristic. All societies we know of have stories as part of their culture, with dramatic enaction and recitation—often accompanied by music—both pre-dating written forms. The earliest conceptualization we have in Western cultures is that of Aristotle [Lucas 1968] in his *Poetics*, dating back some 2,500 years. The book of this work that survived focused on tragedy (it had a second book on comedy that has been lost) and one of its most important concepts was that of *muthos*. This corresponds to what we might call plot, the structured events and actions that make up a story. Plot has remained a central part of narrative theory, and indeed some might argue that having a plot is what makes a story, a story.

A second of Aristotle’s concepts was *mimesis*, which is how the story is represented. The *same* story can be represented in quite different narrative forms—as a song, a play, a computer game, a comic strip. Its presentation could also be organized in different ways, hence ideas such as flashbacks and narrative jumps. Not all plot events necessarily appear in the presentation, they may be implied or referred to. Indeed, in classical Greek tragedy, most dramatic actions happen offstage.
However, like most distinctions between form and content, there is in practice some blurring. Some narrative forms and some narrative techniques work better with some sorts of story than with others. For example, novels are good at portraying the inner voice of characters, while films are good at portraying their external circumstances. The issues involved in moving a story from presentation as a novel to presentation as a film are frequently non-trivial.

An assumption of Aristotle’s, and of much subsequent narrative theory, was that a narrative must be authored. The author was responsible for the content, while they might or might not be wholly responsible for the form. In oral traditions, a narrative might be learned and then presented by other performers. The advent of writing allowed the authoring process to be much more easily disconnected from presentation. In theatrical drama, the actors also carry some responsibility. Thus the concept of being true to the author’s intent.

Artificial Intelligence (AI) researchers became intrigued by the idea that narrative might be computationally authored more than 50 years ago. Natural language understanding and natural language generation are long-standing fields of investigation in AI, and it was natural for researchers to consider whether a computer could generate say poetry, or a novel, rather than just answer questions about the contents of a database. At this time, computer graphics required expensive specialized equipment, so that initial work was entirely language-based. Just as the researchers of the day investigated natural language via grammars, it seemed feasible that story grammars might support the generation of narratives [Rumelhart 1975].

However, the 1977 system TALESPIN [Meehan 1977], one of the earliest story generation systems, implemented a different approach. It used a set of rules and goals to output written stories a few sentences long in the style of Aesop’s fables. While the quality of the stories as stories was low, its innovation was to give characters goals that motivated them to act and interact. There was therefore no explicit pre-presentation plot, though Meehan [1977] makes it clear that goals were chosen with the aim of generating specific stories.

There has been a continuing tension between the idea of computer technology as a presentation mechanism for an authored story and the much more radical idea of a computer authoring a story itself. How could an SIA be part of a computationally presented pre-authored narrative? A pre-authored plot necessarily constrains the role an SIA might play. There is limited scope for autonomous action selection since the plot determines the overall actions of the story at some level of abstraction. However, depending on the form, there may be scope for expanding plot actions at presentation time. Thus actors can in principle determine where exactly they move to on a stage (though this is usually decided in advance by a director), what expressive behavior they display, and how they say their lines. An SIA could
therefore have a small set of autonomous capabilities subject to the plot. We could describe this as weak SIA involvement.

However, the classical authoring assumption, though dominant, has never been universal, and improvisational narrative runs at least as far back as authored narrative. Consider children's play, or social narratives constructed in turn by successive speakers, or dramatic improvisation, *improv*. In these forms, the story does not have a plot when the presentation starts, though it could be said to have had one by the time the presentation finishes. Plot and presentation proceed hand-in-hand and the narrative is co-authored interactively. If we imagine SIAs as possible co-authors, driven as in TALESPIN by their own goals, then we have what we might call strong SIA involvement.

The growth of computer graphics, and in particular the immersive display systems of 1990s Virtual Reality, gave a powerful impetus to computer-driven narrative research. The science fiction television series, Star Trek, became a particular inspiration. It included a rest-and-recreation area for crew on long voyages called the *Holodeck*. Here crew could enter fictional graphical worlds (handily equipped with solid objects and real physics) and interact with intelligent graphical characters in narrative experiences. To many researchers, Virtual Reality seemed like a technological basis for a real Holodeck, forcing the questions of how to create interactive narrative to the forefront [Murray and Murray 2017]. This was when the field really took off. The new technologies seemed to offer a wholly novel genre of narrative experience. We will see below, however, that the addition of interactivity has also had a substantial impact on the difficulty of the task.

So far we have been motivating the use of SIAs as actors in virtual graphical dramas. These might be for entertainment, or they might have a pedagogical purpose. However, the reason that all societies engage in narrative production and performance is a deeper one. Narrative structuring is thought to be a fundamental aspect of human self-definition through what is known as autobiographical memory [Neisser 1988]. This is the structuring of personal experience into a story-of-self and a story-to-self.

While it relates to the individual, autobiographical memory is very clearly generated through social experience and gives us a basic story-telling capability. However, as well as being generated through social interaction, autobiographical memory also supports it. Human–human interaction is full of personal anecdotes, and the telling of such personal narratives is involved in the building of trust and rapport. For these reasons, SIAs in any domain that are expected to function over longer than a half-hour experiment, possibly for months or even years, should be equipped with such a memory and the story-telling capabilities that it supports [Ho and Dautenhahn 2008, Lim et al. 2009, Pointeau and Dominey 2017].
26.2 Models and Approaches

The extent to which instances of stories fit theoretical definitions varies. The long history of story-telling includes myths and legends, parables, comedies and satires, tragedies of all sorts, narratives in which one story jumps into another partway through, soap operas, anecdotes, miracle plays, absurdist dramas, comic strips, live-action role plays, and a large set of other genres. And that is without including film and musical drama. Expecting one theory to cover all of them over all that time and all those cultures is a big ask.

However, you might think that since narrative has been around for so long, by now there would be an uncontested definition. The most widely accepted definition in Western societies would be Aristotle’s, focusing on a structured set of events, where “structured” usually implies some form of causal linking. However, on its own, this would include weather forecasts as narrative as well as most scientific experiments. These are not themselves seen as typical of narrative, though they may be “turned into” narrative—thus the story of the discovery of penicillin, or the 1987 UK weather forecast in which the forecaster jocularly dismissed the idea of a hurricane only for one to hit the country a few hours later. These storified versions [Aylett 2000] seem to turn a sequence of events into a plot by having them impact characters. If for characters we read SIAs, then here is a possible in.

Aristotle did include characters as part of the necessary content of narrative, but in his view they were subsidiary to the plot, forming a mechanism through which the plot was revealed. This was very much in line with Greek tragedy in which characters had limited agency, being subordinate to Fate. Adding a requirement for stronger character agency, and a further requirement for events to impact characters—change their state—as well be caused by them, better matches modern story forms. Attempts to reapply Aristotle to modern genres, whether by media theorists [Laurel 2003] or system builders [Mateas 2000] have focused on redefining the relationship of characters to plot.

The emphasis on structure as a defining element of narrative was taken up by an early 20th century school of narratology known as structuralism. Working at a much finer granularity than Aristotle, structuralists looked for elements in common over all narratives. The Russian structuralist Vladimir Propp has especially influenced some in computational narrative systems with his 1928 analysis of Russian folktales [Propp and Scott 1968] in which he proposed some 30 common structures. Each structure bundles characters and events. See, for example, structure no2: Interdiction: A forbidding edict or command is passed upon the hero (“don’t go there,” “don’t do this”). The hero is warned against some action; or structure no16: Struggle: The hero and villain meet and engage in conflict directly, either in battle or
some nature of contest. Similar ideas from Joseph Campbell in late 1940s America [Campbell 1949] produced the idea of The Hero’s Journey.

Computational modelers are attracted to structures as this is what computers are good at. Propp and Campbell might be very good starting points for a system based on fairy stories [Paiva et al. 2001] or for authoring a quest-based computer game. Not only this, the technology of AI planning acts as a very convenient implementation mechanism, concerned as it is with coherent action sequences meeting specific goals. The UNIVERSE system of the 1980s [Lebowitz 1987] used planning to create plot outlines for interpersonal melodramas, aka soap operas. Taking Aristotle’s plot/presentation distinction, its outputs were very much at the plot level, since generating character-level actions, in the form of dialogue, was considered a problem to be solved later. As a result, they read more like summaries of stories than stories themselves.

It should be no surprise that structuralist approaches are rather genre-specific, and do not generalize well across genres. Propp targeted Russian fairytales, not soap operas. Moreover, as with Aristotelian pre-authoring, they are not easy to reconcile with interactivity. They assume a linear narrative in that the underlying plot forms a single sequence of actions from beginning to end.

Interactivity was both the inspirational contribution of the Holodeck idea and a major theoretical and practical problem for computational narrative. It presents what has been called the narrative paradox [Aylett 2000], a conflict between predetermined narrative structures—especially plot—and the freedom a graphical world offers a user for spatial navigation and interaction, integral to their feeling of physical presence and immersion. It is worth noting that the superficial visual similarity between interactive graphical worlds and film has led to substantial influences from the latter, especially visible in computer games with their video cutscenes. This is in spite of film being in general one of the least interactive of forms. Even its actor-level behaviors are under the control of a director through multiple cuts during filming as well as later editing, and the final spectacle is intended for spectators, not participants.

Broadly, the whole field of interactive narrative falls into a plot-dominated camp that attacks the narrative paradox by accommodating limited degrees of freedom for the characters and a more radical character-based camp that examines how far and under what conditions character interaction can produce spontaneous narrative structure. This has become known as Emergent Narrative [Aylett 1999, Louchart and Aylett 2004] and relies on the idea that just as physical structures can emerge from interaction in the natural world so narrative structure can emerge from human interaction under specific conditions.
Interactive Narrative and Story-telling

In some ways this is yet another rerun of the top–down versus bottom–up arguments that pop up in many branches of AI research. The top–down approach guarantees meeting structural goals but at the expense of responsiveness and flexibility, while the bottom–up approach is flexible and responsive but lacks guarantees about outcomes. The bottom–up emergent narrative approach has helped to direct narrative research into new channels, however. It focused on improvisational narrative forms that had been relatively neglected: interactive theatre [Boal 1995], theatrical improv [Magerko et al. 2010], and role-playing games [Louchart and Aylett 2003]—table top, live-action [Tychsen et al. 2006], and massive multiplayer online role-play games (MMORPGs).

Plot-based approaches can be more or less pre-authored, depending on the level of abstraction of the narrative goals to be met. If one follows Aristotle, then every character action would be specified, with characters at most having scope to interpret the details of execution. On the other hand, a hierarchical AI planning perspective could work at the level of “boy meets girl” allowing for real-time expansion of exactly how this happens. In both cases the narrative would remain essentially linear.

However, the most popular approach in plot-based narrative to supporting a degree of character choice abandons linearity for narrative branching [Riedl and Young 2006], along the lines of the “choose your own adventure” books of the 1980s and 1990s. These coincided with the arrival of hypertext as a technology. A section of literary theorists became very interested in hypertext, seeing it as a way of authoring interactive narrative [Landow 1991] without the need for advanced technical skills.

As a hidden form of menu selection [Delmas et al. 2007], it is widely used in computer games to provide a degree of replayability, although in most cases branching occurs within well-defined subsections of the game that then all branch back to a main path. The plot can no longer be represented by a chain of actions but becomes a directed graph. This does offer limited autonomy to both user and characters, even if this may boil down to making a choice between a small number of narrative branches.

Character-based approaches are by definition not pre-authored in the same way, though they do require a specified world in some initial state, and motivated characters. They also require mechanisms to shape interactions into something that feels like a coherent story to participants. A story director agent is a typical solution [Weyhrauch 1997]. Such a director may have variable powers: in some systems it might decide a character’s action for them at a critical point or might be confined to acting on the storyworld in the style of an RPG Gamemaster [Tychsen et al. 2009]. A more distributed solution, not incompatible with a director [Weallans et al. 2012],
is to equip characters with an understanding of their social role as an actor in a drama. For example, a character could evaluate possible actions for their dramatic impact on others around them [Aylett and Louchart 2008].

Clearly, there is much more scope for SIAs in character-based approaches; indeed the whole approach hinges on SIA capabilities. In its most radical form, its solution to interactivity is to put the participating user on the same footing as any other character, generalizing their interactive freedom. This is also a sharp move away from the idea that the narrative is performed for an audience of spectators, an assumption of most narrative theory. In turn, this requires new methods of assessing narrative success since it is perfectly possible for an emergent narrative to look incoherent to a spectator—depending on their point of view—while feeling coherent to a participant [Aylett and Louchart 2007]. Consider the narrative experience of an MMORPG player on a group quest and whether there is in fact any room for a detached audience.

The key difference between interactive narrative and story-telling is that in the former characters act out the story, while in the latter a single character, the storyteller, presents the story verbally. Interactive narrative supposes the user to be a participant in some form. Story-telling supposes an audience, and the storyteller has explicit control of the usually pre-authored narrative, though they might alter their telling in response to audience reactions or invite specific participative actions [Kistler et al. 2011].

This poses very different requirements for SIAs. For interactive narrative, the focus is on the actions characters can take to forward the narrative; in story-telling an SIA narrator above all needs appropriate expressive behavior with which to dramatize the telling. Of course, if the text is pre-authored, it can also be annotated with such behavior, using an appropriate markup language [Niewiadomski et al. 2009].

26.3 History/Overview

Before about 2000, most work in this field focused on story generation and not interactivity. An influential system of this period was Minstrel [Turner 1993], which unlike the story-grammar approach applied case-based retrieval and adaption. As its name suggests, it focused on stories of knights and chivalry and used an author-level planner to call a case-based reasoning system. This refined the story-graph. For example, it could insert dramatic support for a major event or an introduction for a new character. It was sufficiently comprehensive to be reconstructed by a different research group more than 15 years later [Tease et al. 2010].

By 2000, new groups of researchers had been inspired by the idea of building something like the Holodeck [Cavazza et al. 2000] and interactivity became a very
prominent theme. A minimum requirement for supporting interactivity was some element of real-time generation with which a user could interact. This did not per se exclude the earlier story-telling approaches or authorial control if they could be run incrementally and flexibly but gave an impetus to AI planning as a generative mechanism whether from a plot- or character-based perspective [Aylett et al. 2006, Riedl and Young 2006].

Interactive technology may come in different forms. The Holodeck supposed an immersive graphical environment and the user acting alongside SIAs on an equal footing in a shared space. Most research groups did not have access to this type of graphical environment and so we will see that many systems were constructed on desktop machines with keyboard interaction, modes that were easy to apply within free games engines. As a halfway house, and especially as interactive devices like the Microsoft Kinect became available, systems used big screen projection and allowed users to interact by gesture. Some systems dispensed with graphic visualization altogether and used text.

The degree of narrative interaction may also vary. The Holodeck assumed users acted as full characters inside the narrative. We will see that other less interactive roles are also possible as well as being technically more feasible.

A pioneering instance of the character-based approach, FearNot! [Aylett et al. 2005] cast the user as an “invisible friend” of one of the characters in a desktop virtual drama. This stance was based on the idea of the spect-actor from Forum Theatre [Boal 1995]. The application was targeted at educating 9- to 11-year-olds against bullying and there were obvious ethical reasons for not allowing a user to participate directly.

The narrative was episodic and between episodes in which one character was bullied by others, the user could advise their “friend.” The SIAs were driven by a comprehensive affective architecture [Dias et al. 2014] and action selection depended on their modeled emotional state. The user’s advice influenced, but did not determine, what the character would do in the next episode by impacting parameters in its architecture. Advice content also determined the selection of an episode setting and cast in which the advice was likely to apply.

A different approach was taken in another desktop system, I-STORY-TELLING, using a scenario based on the US sitcom “Friends” [Cavazza et al. 2002]. Here a character was driven by a hierarchical planner, and the user was able to change the state of the world in ways that would impact their plan. Thus, if one character was looking for the diary of another, with a view to checking where to take them on a date, moving the position of the diary would impact whether they did or did not meet the diary’s owner along the way. This might then precipitate replanning and a change in the narrative. Here the user is playing the role of a ghost: not visible
to characters in the storyworld but able to change its state. A further consequence was that just initializing the system differently could produce different narratives even without user intervention.

The Mission Rehearsal Exercise [Hill et al. 2001], a prototype training system built for the US military, took a different interactive approach. A single participant played the role of an army lieutenant but interacted with the scene in plausible military fashion via limited speech with a sergeant character. This was a conversational SIA with an affective architecture [Gratch 2001] who would relay commands changing the state of the storyworld, along a small set of possible branches. Other characters carried out simple pre-scripted behaviors. The system used projection onto a screen and surround-sound to produce a feeling of immersion, with characters at approximately life size.

A related system by the same group, SASO [Core et al. 2006], supported direct interaction between the participant and an SIA playing the role of a doctor in Iraq, whose clinic had to be moved to make way for military operations. The participant was able to communicate with the SIA using speech. The aim of this training system was to develop negotiation skills; the narrative was again small-scale and one-episode long. In both these systems, the narrative was dominated by the training objectives, much as it is in human training role-plays.

Similar in inspiration as interaction technologies matured and became more accessible were the ORIENT [Aylett et al. 2009] and Traveller [Degens et al. 2013] systems. Both had SIAs driven by the same architecture as in the earlier FearNot!, and both were episodic in construction. Like the systems just mentioned, they combined narrative with pedagogy, in this case education in cultural sensitivity. ORIENT combined three participants around the age of 14 into a team by giving them complementary interaction devices: one with a touch-sensitive Dance Mat for navigation, one with a game-based WiiMote for gestures, and one with an RFID-equipped mobile phone to allow objects to be passed from real to virtual world. The SIAs represented frog-like aliens with unfamiliar cultural mores whose world was threatened by an asteroid. The participants were firmly on the real-world side of the screen but could communicate across it, and social interaction between the participants was viewed as a significant component of the pedagogy.

The later Traveller system used simpler but more functional interaction technology in the Microsoft Kinect. A single participant of young adult age was represented by an avatar inside the storyworld, with gestures used to select between actions within scenes, giving a branching narrative. A further innovation in interactional technology was produced in a digital version of part of the novel Madame Bovary [Cavazza et al. 2007]. Here physiological sensing was used to infer the possible affective state of the user and generate choices in a branching narrative.
Even more ambitious interaction technology was used in the US Gunslinger installation [Hartholt et al. 2009], using an augmented reality approach. A real-world Wild West saloon set incorporated flats—large video panels—showing life-size SIAs. These could respond to gesture and speech, and the participant was equipped with a cowboy hat containing tracking equipment and a holster and six-shooter that worked like those in video arcades. However, one participant [Murray 2017] reported that there were not enough narrative and interactional cues.

The very ease of navigation in the real world compared to that in virtual worlds perhaps put even more onus on interaction with the characters. This also illuminates a narrative problem associated with free participant navigation in the storyworld—they may not be where the story is happening when it is happening. Character-based approaches must design participant goals carefully in order to structure their narrative experience. This is very much a lesson from Live Action Role-Play in which briefing participants about their character’s backstory and goals is a key part of the setup.

With multiple participants, the distinction between plot and presentation, referred to as Fabula and Sjuzet (or discourse) by structuralists like Propp [Propp and Scott 1968], becomes inadequate. This is because there is no longer a unique plot in a world in which many characters and many events co-exist. Three rather than two levels [Swartjes and Theune 2006] have therefore been suggested.

The Fabula becomes a causal network of all events that took place in the storyworld from beginning to end. However, the plot is a subset of these events from a specific character/participant perspective, and thus no longer unique, with one Fabula supporting many plots, as in MMORGs. The discourse, as before, is how these events are presented. The Virtual Storyteller [Theune et al. 2003] supported this conceptualization, creating stories with a collection of SIAs. Inspired by Improv, SIAs were given the capacity to alter the storyworld, for example, by adding required objects, in order to create a coherent plot for themselves.

The most widely influential system of this period remains the desktop system Facade [Mateas and Stern 2003]. As a self-contained experience of 15–20 minutes, it was distributed and run well outside the usual circles of researchers in the field. It made the participant a neighbor of a married couple invited round for drinks. The narrative is driven by conflict between the two SIAs along the lines of the play Who's Afraid of Virginia Woolf?, drawing the participant into a set of strong affective conflicts. Unlike the systems discussed so far, the drama was not driven by autonomous characters. The authors held to the Aristotelian principle of a plot as mainspring of narrative structure [Mateas 2000], in opposition to strong-SIA concepts.
In order to support a range of user actions, Facade was authored as a large database of plan-fragments called *beats* accessed via interaction rules. Rather than a limited number of explicit narrative branches, it attempted to exhaustively cover the narrative space, which required a very long development period of about 5 years. A participant was able to freely navigate in the space representing the SIA's apartment and interacted by typed free text.

If the downside of character-based approaches is an incoherent plot, the downside of plot-based approaches is characters acting “out of character” in order to fulfill the plot's demands, undermining their believability. Facade dealt with this problem by only authoring beats that maintain character believability. A different approach is to treat character believability as a constraint on a planner generating plot [Riedl and Young 2010]. Work on the Intent-based Partial Order Causal Link (IPOCL) planner incorporated reasoning about character intentionality. It divided planned events into *happenings*—accidents, involuntary reactions to stimuli, and forces of nature that do not require character intentionality—and *non-happenings* that do. It both recorded possible intents characters could have and temporal frames over which they committed to trying to realize them, allowing non-happening events to be motivated by character goals.

**Similarities and Differences between IVAs and SRs**

The reader may have noticed that so far little mention has been made of SRs and that the systems discussed have all involved IVAs. This is because there is much less work involving SRs and narrative with what there is dominated by story-telling rather than interactive narrative.

SRs share with IVAs the affordances of embodiment including expressive behavior and the ability to exhibit agency. That said, there are also substantial differences, making them harder and more time consuming to work with. IVAs can have perfect information about the graphical world, including any other IVAs. This world is in principle a wholly controllable environment. Mobile SRs are physical artefacts in an imperfectly controllable physical world, with fallible sensing and actuation capabilities. A basic requirement in drama of knowing where you are on the stage and what is within your field of view is trivial for an IVA and very tricky for an SR.

Furthermore, a constraint that is seldom mentioned is that mobile SRs have a battery capacity of between 30 minutes and a few hours, depending on how much activity they engage in. After which they typically take a few hours to recharge. Most robots are also much more expensive as single artefacts than graphical characters, are available in limited numbers, and require relatively large spaces in which to operate. Larger robots require accessible engineering support and are expensive...
and difficult to transport. Touring a production with a cast of multiple robots could not be undertaken lightly. Even Science Fiction films have never used autonomous robots as actors.

Historically, robots were very much task-focused, exploiting their basic ability to operate physically on the world. While SR research focuses on human–robot interaction (HRI), most of it also operates within requirements for useful function. This may be therapeutic, educational, or assistive, but outside of specific companies (Disney being the prime example [Kober et al. 2012]) is less often narrative or dramatic. Existing entertainment examples, as in theme parks, are nearly all animatronic rather than exhibiting any degree of autonomy [Lu 2012].

Many comparative experiments suggest that SRs also have a greater social impact on users than IVAs given they share the same physical space [Wainer et al. 2006]. However, the potential for physical impacts makes mixing mobile robots and people in social spaces potentially hazardous. In such shared spaces, SRs are normally kept stationary or move very slowly with audible warnings, neither being very compatible with a dramatic suspension of disbelief.

For all these reasons, the use of mobile robots as actors is rare and of autonomous robot actors even rarer. A notable exception was the work of a group at Carnegie-Mellon at the end of the 1990s [Bruce et al. 2000] in which two robots autonomously played out an improv exercise in which one actor wants to leave the room and the other has to find ways of preventing them. By giving the robots different personalities, this produced varying performances. An autonomous robot was also used in a drama involving a single robot and actor in France [Lemaignan et al. 2012], mainly designed to demonstrate to an audience the difference between robot and human perception.

There has been more recent work in Japan, which has a long tradition via Bunraku puppetry of using mechanical devices in drama. However, this tradition may predispose work to tele-operation [Chikaraishi et al. 2017]. A 2011 US performance of Shakespeare’s Midsummer Night’s Dream with seven flying robots acting as alter egos to the fairies in the play also used tele-operation [Murphy et al. 2011]. A further US example involved high-level tele-operation of the robot at the level of action selection rather than direct physical manipulation [Zeglin et al. 2014], though this lower level was available in case it was needed.

Where researchers engage in dramatic work with large mobile robots, the motivation is usually the study of expressive behavior and more flexible HRI [Knight 2011]. Robot theatre requires dramatic rather than naturalistic behavior, and this in the context of platforms with much more limited expressive capabilities than those available for an IVA. The study of IVA expressive behavior is substantially based on human-looking IVAs and, as discussed in a reference to Chapter 8 on
“Multimodal Behavior Modeling for Socially Interactive Agents” [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021], often draws on accounts of facial expression from psychology.

SRs may not have a face at all, or if they do, one that supports restricted or even zero facial animation. Lack of expressive speech—discussed in Chapter 6 on “Building and Designing Expressive Speech Synthesis” [Aylett et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]—is also a major expressive deficiency and becomes even more of a problem in story-telling systems dependent primarily on speech. These are issues that may well be better addressed by the conventions of animated characters [Johnston and Thomas 1981, Ribeiro and Paiva 2012] than of human-like behavior.

There is growing interest in using the development of robot-based drama as an educational process in its own right [Ryokai et al. 2009, Bravo Sánchez et al. 2017, Sullivan et al. 2017, Barnes et al. 2020]. Here, the quality of the dramatic output and the level of autonomy of the robot actors are less important than the classroom-based development process and its pedagogic outputs. This work prioritizes the use of small robots already available in education, usually with quite limited capabilities. It may also mix them with other physical elements—tangibles or even cards, supporting narrative development by child users [Fischbach et al. 2018].

Story-telling robots form a separate narrative domain that is in general far less technically fraught. Audiences are small, often children, who are more tolerant of technical failures. Story-telling robots usually involve what are known as desktop robots. These are much smaller than robots designed to move around a room, and they remain in one place. They do usually have movable heads, are often able to use glance, and sometimes have expressive facial features. They are much less expensive and complex and can use mains electricity. On the other hand, story-telling creates a very strong focus on the expressive behaviors of the storyteller, since it is this that animates the story and engages the audience [Conti et al. 2017]. Deficits in expressive speech become especially significant and many systems have used recorded human speech as a result.

In principle, a robot telling a specific story could be wholly pre-programmed, with the text annotated with suitable head movements, glance, and appropriate facial expressions. However human story-telling is a social activity in which the storyteller takes account of audience reactions to modify their performance and possibly how the same story is told. Given the current limitations on processing user reactions, this is still very challenging, especially inferring an audience's emotional reactions [Costa et al. 2018].

A constructive approach to education sees story-telling by pupils as a source of active learning [Catala et al. 2017a]. This means that story-telling robots need
not themselves tell the stories and some of the technical issues can be mitigated by social interaction between the human participants. An SR can provide an inspiration or prop for child story-telling [Plaisant et al. 2000] and the feasibility of co-operative story-telling is also currently being investigated [Sun et al. 2017].

26.5 Current Challenges

Most researchers in the field would agree that the early excitement about the possibilities of the Holodeck has not resulted in many fully realized systems that could be enjoyed for themselves as narrative experiences rather than as demonstrations of research ideas [Murray 2017]. Facade, as discussed above, remains a rather isolated example and is known to have taken 5 years to develop. FearNot!, with 44 possible episodes, was developed and evaluated over the course of two 3-year projects.

This is in spite of a great deal of work on authoring systems. The effort involved in developing an initial digital narrative system has led to most groups working in the area over a period of time to develop their own tools to make this easier. Generalizing a specific system into a more generic architecture is an obvious move—hence FAtiMA [Dias et al. 2014] from the FearNot! work, and Thespian [Si et al. 2005] and the Virtual Human Toolkit [Gratch et al. 2013] from systems such as MRS, SAPO, and Gunslinger. The experience of building Facade lies behind the toolkit Comme Il Faut [McCoy et al. 2010] and the extensive base of social behavior rules used to build PromWeek [McCoy et al. 2012], a narrative game in the style of those that appear in The Sims.

Some work has started at the other end of the process, with an explicitly designed authoring system that incorporates runtime support for its output. However, just as architectures abstracted from specific systems embody the narrative approach and assumptions of that system, so authoring systems also embody specific assumptions and representations. This has always been clear in the authoring systems associated with games engines. Thus the Dungeons and Dragons-based Never Winter Nights of the late 1990s had an authoring system based around locations, navigation between them, objects found in them, and creatures encountered there, either for conversations or battles. On the other hand, the first-person shooter engine Unreal Tournament, of the same period, had a system for authoring shooter episodes.

IDtension [Szilas 2003] is a longstanding narrative authoring system, text-based in its original form. It is plot rather than character based and is able either to generate a story itself or to produce one in real time with a participant making action choices alternately with the system. Its approach is to present every possible action in its model at every turn, making a search through its knowledge base of goals,
actions, and obstacles. This can result in more than a hundred possible choices during the story. It was made available to non-experts to produce a complete sample drama, The Mutiny [Szilas 2008]. Other authoring systems, such as STORYTEC [Göbel et al. 2008], may have narrative elements but are primarily game authoring tools, usually for adventure-type games in which a narrative is really a backstory for puzzle-solving.

While invaluable in increasing the productivity of research groups, authoring systems have not so far led to more complete narrative experiences. One obvious reason is that the theoretical difficulties we have discussed of reconciling structure and interactivity are still not wholly solved. In computer games (see Chapter 27 on “Socially Interactive Agents in Games” [Prada and Rato 2022] of this volume of this handbook), this issue is generally addressed by restricting narrative interactivity in favor of much simpler forms such as shooting. Narrative branching is still the standard approach in commercial games, along with the use of non-interactive video cutscenes. Non-player characters rarely have the necessary capabilities for autonomous action, in most cases not even a memory of past interactions. However, two other reasons are at least as important. The first is the need for content and the second the technical skills required to develop such systems.

In the other graphical genres—film and computer games—large numbers of people are involved in generating content. Indeed, many games companies do not distinguish between pre-authored narrative and pre-authored graphical content. They are as wary of open-ended narrative as they are of machine learning, and for similar commercial risk-avoidance reasons. Researchers, on the other hand, seldom have access to large-scale graphical content and even more seldom have such creative skills themselves. Routine content production is not a publishable research task. This is one reason for the widespread use of game engines since they come with free content or a user base that generates and shares such content.

In general, the greater the degree of user interactivity, the worse the graphical content problem becomes. If users can change the state of the storyworld, then new states must be visualized for them. If story characters can also change the state of the world, then the problem multiples. Moreover, the greater the roles to be played by characters, the greater the repertoire of actions they require, all of which must also be visualized. Not only this, characters nearly always require natural language interaction capabilities, and this requires actual language content, whether pre-authored phrases, lexicon and grammar, or a machine learning-based dialogue system. Character-based narrative approaches have more substantial content demands than plot-based approaches because their aim is to produce a wider range of narrative experiences.
Tabletop role-play solves this problem by supplying gamemasters with volumes of background world material, including the milieu, some history, and a set of backstories. This works because there are plenty of gamemasters and role-players and using the materials only requires the cost of the book and literacy. Moreover, the materials are interpreted by human intelligence. But consider turning the whole of a roleplay universe into shareable graphic materials. This would be a much larger-scale enterprise than the materials associated with a game engine and would have to find a common representation—graphical and probably ontological too—that could be used by multiple game engines. Specific MMORPGs do solve the problem—and some, such as World of Warcraft, are effectively graphical versions of earlier written source materials. However, these are closed proprietary systems. Open systems like Minecraft allow users to generate graphical materials but do not embody any narrative framework.

The second barrier to fully realized systems is the level and range of technical expertise required to author them. The large teams involved in film and computer-game production support multiple skill specializations. Interactive narrative shares similar integration problems, but research teams are rarely large enough to include a full range. While a novel can be attempted by any literate person on their own, the need for an understanding of narrative technology is a barrier to creative non-specialists. Widely usable authoring systems remain a challenge.

The technical experts, on the other hand, are rarely talented constructors of compelling narrative and characterization. Game engines provide 3D graphical environments with backing code banks, but they are focused on supporting specific game genres and require programming effort to support functionality beyond that. The most-used engines are proprietary and have substantial licensing charges if used for publicly distributed work. The research-based authoring systems and architectures discussed above all require a good understanding of the technology of the system being authored to be successfully used.

The content challenge is currently being addressed in three ways: by procedural generation, by crowd sourcing, and by the application of machine learning approaches. Interactive narratives require a broad range of different types of content, some graphical (“physical” environment, graphical models of objects, character bodies and animations) and some related to the technological approach of the narrative being constructed. For example, a plot-based approach may require content for the branching graph representing the plot; an emergent narrative may require character actions.

Procedural generation is widely explored in computer games research [Hendrikx et al. 2013]. Methods are available for terrain, vegetation, bodies of water, road networks, and urban environments, as well as lower-level textures and
materials. Procedural animation systems have been developed for the computer games market and researchers have tried to extend these into forms more suitable for autonomous characters [Horswill 2009]. While this is all useful, it is not narrative and works better for games in which simple interactions like navigation and fighting are important. Moreover, most of these systems are developed as middleware for the computer games industry, often at premium prices, and are not necessarily accessible to researchers. Generation of puzzles for RPG games [Fernández-Vara and Thomson 2012] is closer to narrative but a puzzle is not itself a plot.

AI planning technology can be thought of as a procedural generation system working in real time [Riedl and Young 2010]. However, it too has content: binding symbolic representations of world state to goal and action definitions. While a graphical procedural generation system can produce a building, an ontology is required to label it “a police station,” and a planning system would need an action “report to the police” as well as a motivating goal to make use of it in a narrative. The planner also needs to know what the consequences of executing the action would be in terms of changes in the state of the world. This is true whether the plan operates at the level of a plot or is attached to an autonomous character.

The AI Planning research community has worked on the problem of automating the construction of planning knowledge for many years [Cresswell et al. 2013, Jilani et al. 2014]. However, there is only a small overlap between the two research communities, and how to generate narratively useful knowledge for an AI planner remains a challenge [Porteous and Cavazza 2009].

Crowd sourcing forms a second possible source of content. One might argue that the narrative content of MMORPGs is a type of crowd sourcing, though it formally depends on quests, a set of authored goals and constraints. An early initiative leveraging large numbers of participants via the Internet [Orkin and Roy 2007] ran more than 5000 online sessions of The Restaurant Game. The game allowed participants to play either a customer or a waitress, so that data was collected from human role-play. This data was used to build a narrative representation called a Plan Network, a statistical model that encoded context-sensitive expected patterns of behavior and language, with dependencies on social roles and object affordances.

This was intended to provide what it called common ground, the world knowledge that could be used in both plot- and character-based approaches about what actions people were likely to perform with which objects in specific scenarios. The approach had clear parallels with corpora-building in the natural language dialogue community, one of which involved ordering meals online from restaurants.
Other work has included crowd sourcing to directly form a plot-graph [Li and Riedl 2015] and the application of structural ideas from Theatre, in the form of Stanislavsky’s Active Analysis rehearsal technique [Feng et al. 2016]. Online rehearsal was also the approach taken by a crowd-sourcing system aimed at collecting character material, including change in emotional state, for an emergent narrative [Kriegel and Aylett 2008].

**Future Directions**

Crowd-sourcing content pre-dated the enthusiasm for machine learning that has swept the AI community at the time of writing, and they can be thought of as comparatively small-scale examples of these data-driven approaches. Reinforcement learning, in which actions are tried and rewarded or punished according to their success, is the technique of choice for learning sequences of actions. Deep reinforcement learning (RL), using the current artificial neural net technologies, has been applied successfully to learning a better-than-human performance set of actions in a number of Atari games [Hessel et al. 2018].

Most games genres are, however, much simpler than narratives, with a small set of actions, a repeating set of contexts, and simple metrics for rewards and punishments. Even so, deep RL requires very substantial amounts of computing power. Recent work required an average of 80 million frames on games running at 60 frames a second—equivalent to about 38 hours of human play on a game a human can learn in tens of minutes. With a wider set of possible actions and contexts that repeat much less often, as well as metrics of “success” that are very hard to formalize, applying deep RL to generating action sets for characters, or for a narrator-level plot, in graphical environments would be a formidable task. Not least a chicken-and-egg problem of not having enough running narratives in the first place to form a corpus. However, if the task is reformulated into a purely language-based one, it becomes somewhat more tractable. As already mentioned, the field of natural language has been using corpora for a long time, and there already exist resources that are useful for learning narrative sequences. One example is a database of film summaries [Bamman et al. 2013]. Natural language corpora can be processed into simpler event-like structures with subject–verb–object plus indirect objects where needed.

A problem of standard recurrent neural network approaches is that they have limited ability to retain a memory of past decisions, so that learning plots results in stories that wander, without much narrative coherence. Recent work [Tambwekar et al. 2019] tries to deal with this by shaping learning toward a final goal, a specified event. As in UNIVERSE and other older systems, this generates a plot description, the basic set of actions in the story. It is some way from the presentation of the story
and so far work has not examined how the output could be used for an interactive experience.

As discussed above, every research group working in a specific software field over a long period develops architectures (see Chapter 16 on “The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures” [Kopp and Hassan 2022] of this volume of the handbook) and tools (see Chapter 20 on “Platforms and Tools for SIA Research and Development” [Hartholt and Mozgai 2022] of this volume of the handbook) to reduce the amount of reworking they must do each time they produce a system. In the computer games industry this has led to an associated middleware industry supplying common tools—very expensive ones in general—to many games companies. In a research field like interactive narrative, the establishment of common representational standards would help this to happen.

The behavior markup language standard developed for driving expressive graphical characters [Kopp et al. 2006] is a help. However, its general-purpose nature, while a strength, makes it a limited component of an interactive narrative system. A set of narrative ontologies for different styles of interactive narrative might deliver more, but the theoretical divergences in the field make these hard to develop. AI Planning has benefited substantially from a Plan Domain Description Language (PDDL) [Fox and Long 2003], but this appears to have made very little impact on work in digital narrative in spite of the central role played by planners. The development of agreed standards is work that still needs to be done.

As a domain in which many technologies must be integrated, interactive narrative also stands to benefit from developments in many related fields. We have already mentioned the need for expressive speech. The rapid improvement in dialogue systems should also be highly significant (See Chapter 15 on “Socially Interactive Agent Dialogue” [Traum 2022] of this volume of the handbook), as also developments in multimodal interaction.

We saw above that the growth in interaction technologies has led to systems with a more ambitious approach to interaction modalities than the keyboard. Multimodal interaction has so far been mostly applied to deliberate participant actions, allowing the use of explicit gesture [Degens et al. 2013]—see Chapter 7 on “Gesture Generation” [Saund and Marsella 2021] of volume 1 of this handbook [Lugrin et al. 2021]. Much more could be done to incorporate participants into a narrative experience using facial expression recognition to infer responses, or action recognition and goal inference, especially in augmented reality settings. The use of Theory of Mind processing (see Chapter 9 on “Theory of Mind and Joint Attention” [Perez-Osorio et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]),
especially in its simulation variant [Dias et al. 2013], also gives scope for enriched interaction.

While the Holodeck was an inspiration to the field, it also prioritized a specific view of an interactive narrative as a self-contained virtual experience. Rather than requiring participants to operate in the virtual world, one can conceive of computer-supported role-play in which SIAs contribute to a real-world narrative experience. This augmented reality view would include the use of tangibles [Catala et al. 2017b] and collective human role-play. An inspiration here would be the Story Room [Alborzi et al. 2000], originally conceived as a space with resources supporting child story construction. Robust technology that could support Live Action Role-Play would also be an interesting way of exploring this idea.

Finally, long-lived SIAs, interacting over weeks, months, or even years (see Chapter one on “Long-Term Interaction with Relational SIAs” [Kory-Westlund et al. 2022] of this volume of this handbook), will absolutely require narrative capabilities. These will organize their interaction memories as well as creating more pleasant and varied interaction [Bickmore et al. 2009, Cordar et al. 2014]. Interactive narrative will appear at some point wherever long-lived SIAs appear.

**26.7 Summary**

We have seen in this chapter that narrative and story-telling are fundamental human capabilities. As such, they have formed one of the challenges for AI researchers tackling a wide range of human skills.

We have considered how this led to the field of interactive narrative and the inspiring example of the Holodeck, focusing on the role of SIAs as characters. We have seen that the two and a half thousand years of narrative theory in the West has not always helped all that much in this field due to the nature of interactivity. We have looked at both plot- and character-based systems and how they have tried to meet the challenge of the narrative paradox, of reconciling user freedom with narrative structure.

Few of the systems we considered have been aimed at entertainment, with education and training often the driving application domains. Computer games have, however, had a substantial impact, both through the widespread use of game engines for development and as a possible route to the uptake of narrative technology. As with other branches of AI, this last has been a slow process, with issues relating to creative control and technological robustness acting as powerful braking processes.

The differences and similarities between SRs and IVAs have been discussed, and the additional difficulties and opportunities that working with robots provides. We also looked briefly at the differences between interactive narrative and story-telling
and why the latter is more often targeted than the former by SRs. Content creation was identified as a substantial bottleneck in the creation of virtual narratives, and we looked at the ways researchers have tried to meet this challenge.

Finally, we have examined how far machine learning may contribute and whether narrative in augmented reality environments can provide some new opportunities.

References


J. Orkin and D. Roy. 2007. The restaurant game: Learning social behavior and language from thousands of players online. J. Game Dev. 3, 1, 39–60.


Chapter 26  Interactive Narrative and Story-telling


27.1 Motivation
Socially Interactive Agents (SIAs) have been part of games from the very beginning, although their dimensions of sociality have evolved though time. SIAs are elements commonly used in games to provide conflict as they help to define the gameplay dynamics and the challenges players face, for example, by incorporating enemies or characters that have information and resources that players need. But SIAs are also elements that provide support to players, for instance, as characters that offer help and accompany the player through the journey of the gameplay experience. SIAs are crucial to conveying a social dimension to the game world and support its social immersion and social believability. In games with more prominent stories SIAs can take narrative roles and functions as well. The use of SIAs in games is not just a typical element but is also an added value with great impact in the experience of players. Placing SIAs in games increases game enjoyment [Dignum et al. 2009] and better SIAs are demanded by players [Afonso and Prada 2008]. In fact, some games are praised for the autonomous characters that they present.

Nonetheless, to create immersive experiences, players’ expectations about SIAs must be satisfied. As with increasingly higher fidelity regarding digital characters’ visuals or the embodiment of robotic game partners, the behavior of these game characters also demands more fidelity and complex social mechanisms. This represents both a need and opportunity for the development of AI in game characters. The social dimensions of the game worlds that game characters populate are increasingly more complex, for example, involving multiple characters, within artificial societies, that act together with other AI characters and with players, who need to understand and adapt to multiple situations. To cope with this, characters need more complex abilities, in particular to have social needs and goals to be able
to act socially. In turn, if characters in games display a wider range of behaviors and autonomy, the available options to players are enhanced, increasing the social interaction space that the game affords. This promotes a higher feeling of agency in players and represents an opportunity for novel game mechanics.

SIAs can be used in games for several different reasons and purposes and can play different roles in the game dynamics and experience. The abilities they need depend on the roles they play. In this chapter, we will discuss the roles SIAs play in games and the requirements and challenges for SIAs to be successful in each role. We will present an historical overview of SIAs in games research and in the market, incorporating SIAs developed for digital and physical interactions (e.g., both virtual agents and social robotics). We will discuss open challenges and present promising directions for future work.

### 27.2 Models and Approaches

#### 27.2.1 Games, Players, and Experience

A game is an activity performed with no direct practical goal. It is a setting for playful interactions, where the outcome of the actions does not have direct meaningful implications in the real world. The actions performed make only sense in the context of the fictional world created by the game. The game experience is bounded by a *magical circle* that separates these two worlds [Huizinga 1938, Zimmerman and Salen 2003]. Players act on both worlds, but the actions taken in the real world have enhanced meaning in the context of the game, for example, in the real world players press a button but in the game world they jump or use a magical item. The motivational factors for actions are framed in the game world and the consequences of the actions are negotiable and only meaningful there as well [Juul 2005]. There are, of course, consequences in the real world, as players spend time and resources and their physical and psychological states change, but those are not directly intended by the player. The real-world consequences can be the initial trigger to start a game experience (e.g., players start a game to relax or spend time), but they are not the ones that keep the player engaged in the game through time. The term game is also, frequently, used to refer to the artefact that supports the game activity, for example, the board and pieces of a chess game or the software and hardware that runs the game. In this chapter, we will use the term game to refer to the aggregation of both concepts. A game is treated as an activity supported by technology.

Furthermore, games are different from mere playful interactions (e.g., toys) as they bring structural elements to the interaction experience. Games have rules that describe the mechanics of the interactions and define what is allowed and disallowed. The rules can be enforced by the technology that implements the game
(typically done in digital games), but this may not be the case. For example, a player can move a chess piece to any place in the board (nothing prevents that in the real world), but only a few positions constitute a valid move as only a few make sense in the game world.

The game rules also support the definition of variable and desirable outcomes, for example, objectives and winning conditions. These guide players in establishing preferences over the states of the game world, as some are more desirable or more meaningful to them. The preferable game states represent focal points that players try to achieve during the gameplay experience. Players’ decisions and actions are grounded on moving the game’s world state closer to one of those desirable states. The progression in the game and, typically, the score, is dependent on the distance of the game state to one of these focal points. The closer it gets, the higher is the sense of progression in the game and, typically, the higher is the score players achieve.

The gameplay dynamics that sustains reaching the outcomes, usually, represents a conflict, as there are elements that try to “oppose” the players’ actions and make it difficult for them to achieve the desired states. This can be done by adding explicit entities in the game (e.g., other players or AI enemies) that take actions to counter the player, but it is also implicit in the nature of the gameplay dynamics. The gameplay may have elements that pressure players (e.g., time limits) and makes it harder to make decisions or to perform actions. For example, performing the actions in the game world may require high hand coordination skills that players need to master (e.g., typical in fighting games) and making decisions may require a deep understanding of a social system with a complex economy, involving resource management and awareness of social motivations and needs (e.g., typical of strategy and role-play games).

The existence of conflict means that to reach one of the desirable outcomes, players need to make an effort to overcome the opposing forces, as the game does not evolve in the direction of one of the desirable states by itself (without the intervention of the player). That usually means that players need to learn and improve their abilities to understand the game world and its dynamics in order to act in an appropriate way to address and overcome the obstacles and hard choices they face. The level of conflict, or difficulty, is defined by design and can promote a more casual or more hardcore gameplay experience. It will be more hardcore if it demands higher effort, skill, and knowledge from the player or, in a general sense, if it requires more dedication and commitment from the player.

However, the nature of the resultant experience is highly dependent on the abilities, interests, and knowledge of the players and specially their attitude toward the game.
By playing a game, players assume a mental attitude with three important elements: (1) commitment to pursue one of the desirable outcomes, (2) willingness to make a real effort to achieve the necessary changes in the game world, and (3) acceptance of the conditions and restrictions defined by the game rules. The game experience is broken, or exhausted, if one of the three attitudinal elements is compromised. If players cannot find something they wish to achieve by playing the game, they will not engage and will not play. Also, playing is voluntary, if players are not willing to invest time and effort to move the game forward, they will not engage as well. This can happen if the demanded effort is beyond their capabilities or availability of resources, or if the desired outcomes are not appealing enough. In fact, there must be a balance between the two elements. Achieving an outcome in the game must involve an acceptable cost for the player. All this must be possible within the limits that the game defines. Players need to understand how the game world works and be able to act according to the rules. In general terms, players need to be able to identify what they can do in any gameplay situation and what is the impact of their actions. The outcomes that players pursue must also be framed in what the gameplay affords. If players do not figure out how the game works, it will be hard for them to engage with the game. The understanding of the game is often supported by the fiction that the game presents, which grounds the interpretation of the game world on known metaphors, and the quality of the feedback that players get in response to their actions.

Interestingly, players often pursue outcomes in the game beyond what is explicitly defined by the rules. This means that players may be committed to playing the game in ways different than explicitly designed. For example, they may try to beat the opponents in the game while not using weapons, try to delay the victory to the last moment even if that gives them less score, or perform actions that are not explicitly rewarded. The Internet is full of examples of tricks that players do with games. Speed-runs\(^1\) are a prominent example. Nevertheless, players more easily commit to outcomes that are rewarded in the game and only define and pursue different outcomes when the ones defined by the game rules have been explored. But, in fact, the game may itself define open-ended goals that do not restrict nor guide the players too much, making it the players’ own responsibility to define what they want to achieve in the game. What is crucial for the game is to support players in the ability to commit to goals (even if self-defined) that are framed within the magic circle, that is, only meaningful in the world and fiction that the game creates.

---

1. A *speed-run* is a playthrough the entire game performed in the minimal amount of time.
Playing a game is a subjective experience. And, typically, different players have different gameplay experiences with a game. Apart from having different skills and knowledge about the game (and games in general), players also have different motivations to play. Some play for mastery, others are driven by curiosity, and others play to get a social experience, for example [Yee 2006]. The game itself is a vessel for the experience. The success of a game is, therefore, dependent on the experience it is able to convey. The experience is built by the actions players perform in the game world, but it is more about what players feel and the memories they build by playing the game. It is the feeling of success and improvement, the feeling of living in and exploring an interesting world or taking part in an engaging story, and the feelings of fellowship developed while playing with others that are the key elements of a game experience.

To promote the game experience is the ultimate goal of the game. The success of the game experience is related on the level of immersion elicited in the players. This comes in different dimensions, as players may be immersed by different factors (often working together). Players are immersed by the space the game creates, both physical and social; by the sensorial sensations they receive, typically, visual, audio, and haptic; by the challenges and tasks they face; and by the emotional journey they live. When deeply immersed, players experience an altered sense of time. The quality of the immersion depends on how believable the game world is. How it is able to sustain the suspension of disbelief and maintain the illusion it tries to convey—for example, making players believe that they are traveling in outer space, that they are a rock star, that they are able to fly, or that they are interviewing a real criminal. This means that the meaning and interpretation of the players’ actions inside the magic circle are perfectly conveyed.

27.2.2 The Roles SIAs Can Play in Games
Social interaction is part of the experience most games convey. It is natural that SIAs have important roles to play in those games. Note that we consider as SIA any game actor that has some level of embodiment and has explicit social interaction with a player or other SIA—independently of the level of agency and complexity of the behavior it displays. SIAs can use simple internal mechanisms, even just following predefined scripts, or can use more complex mechanisms supporting advanced behaviors.

To study the roles SIAs can play in games, we first discuss the nature of the social interaction experience that games promote. Social interactions can be addressed and analyzed from different perspectives in games. On one hand, social interactions are part of the gameplay dynamics, as several game actors interact to make the game progress. These social interactions are not necessarily positive. Game actors
can fight, steal, exchange resources, share information, coordinate actions, and so on. The actions performed by game actors are also interpreted beyond the gameplay dynamics and may have specific meanings in the game fiction and narrative. In this perspective, the roles SIAs play are narrative roles, of characters in stories, such as, protagonist, antagonist, and so on. We can also consider the game as an artefact to support social interactions among players. In this sense, the social interactions are interpreted by the player in the real world, outside the *magic circle*.

The different perspectives of social interaction in games suggest different types of roles for the SIAs depending on the dimensions of the experience that they influence. SIAs in games may act in three distinct layers:

- The *Player layer* that refers to the game actors in the real world rather than the game world. The interpretation of the social interactions at this level is framed outside the game world and not within the *magic circle*. *Players* are actors external to the game that, nevertheless, influence the course of the game (e.g., are performing actions in the game through the control interface). Players can be human or artificial (AI controlled). In this layer we may also fit people (or SIA) that are not taking actions in the game but have an interest in the game. For example, supervisors of the game or the gameplay's audience.

- The *Gameplay layer* that encapsulates the game actors as *agents* capable of acting in, and perceiving, the game world. The social interactions in this layer have a functional dynamics that change the game's world state and move the game forward. This layer frames SIA, and their interactions, as elements of the gameplay mechanics.

- The *Narrative layer* that presents the *characters* that take part in the game's story and support the creation of fictional interpretations of the game world. The social interactions from this point of view sustain the fantasy conveyed by the game. It excludes the agents that are not crucial for the fiction or story and may include characters that are not actors from the gameplay perspective.

A game actor may be projected into multiple layers if it takes part at the different social dimensions of the game (see Figure 27.1). For example, players often have an avatar in the game that represents them. The avatar is the agent that performs the actions in the game world for the player. Avatars may have some autonomy and make some decisions about the actions to perform or may strictly follow what is requested by players. Players may have different avatars in different phases of the game (e.g., Thomas was Alone (Bithell Games, 2012)) or even have a direct choice on
which avatar to take in each situation (e.g., Trine [Frozenbyte, 2009]). The avatar is often the vehicle to define the limits of players’ actions and perceptions of the game world, as is also a mechanism to help players project themselves into the game world, supporting immersion. The avatar binds the way players see themselves in the game world and how others see them in the game as well. Nevertheless, games do not necessarily provide avatars to players, and they may directly manipulate the objects in the game world and may have social interactions with the gameplay agents without having an explicit embodiment in the game world.

Many agents in the game world are not controlled by players (e.g., are not avatars), these are frequently referred to as Non-player Characters (NPC). NPC, typically, have both gameplay functions and narrative functions. The gameplay layer captures the mechanics of the interactions that the agents support (e.g., which actions they perform and how they respond to the actions of others). This defines the affordances of the social interaction space that players can explore in terms of gameplay. These interactions have, additionally, some interpretation in the fictional world and may convey narrative meaning. These are captured in the narrative layer, where the agent is perceived as a character in a narrative setting. For example, an agent may be the means by which players get a resource needed for gameplay (e.g., ammunition), and may represent a character that players can relate to and that shares stories about the game world.

In the case of a social robot playing a tabletop card game like Sueca [Correia et al. 2016], the SIA takes the role of an artificial player in the player layer. The game world is supported by physical items (e.g., the table, the cards) and there is no need for an avatar as the player directly manipulates the cards. Additionally, the game
does not define a fiction or narrative that projects the SIAs in the narrative layer. But a social robot can also control an avatar or play a character in a narrative if the game supports such a setting, for example, in tabletop Role-play Games [Fischbach et al. 2018].

All layers are important to building a good game experience, but games define them deeper according to the experience they try to provide. For example, the feelings of fellowship are more strongly conveyed by the social interactions in the player layer, the feelings of mastery are conveyed by the gameplay layer, and the feelings of fantasy are conveyed by the narrative layer. Each layer presents different challenges for SIA. For example, for a SIA to perform well in a game and support the games’ target experience, it will need to be able to be part of the players’ community; to be able to offer challenging social interactions to players (e.g., that need coordination or the use of persuasion); or to be able to play roles that enrich the fiction and the narrative, respectively. In fact, it is important to align the roles in the perspectives of the three layers to provide high-quality game experiences and promote players’ immersion. For example, it is desirable that storytelling and gameplay are well integrated in a game, and for that there should be a strong relation between the actions that agents perform in the game world and their narrative meaning in the story setting.

We will center our discussion on the social roles that SIAs can take in terms of gameplay. We discuss how these are related to the other layers, but we will not focus on narrative functions of SIAs and types of players in games. The reasons are, on one hand, because, with a few exceptions [Warpefelt and Verhagen 2016], studies about the gameplay functions of SIAs are not common, but also because the use of SIAs for narrative scenarios are discussed in Chapter 26 on “Interactive Narratives and Story-telling” [Aylett 2022] of this volume of this handbook.

The definition of the three layers helps us study and define the concrete elements SIAs need for each dimension of a game’s social experience. These promote distinct types of social exchanges framed in nested contexts. Many social interactions occur inside the game and are framed by the contexts that the game creates, but the social interactions extend outside the game world as well. There, the social context is wider and includes awareness of the contexts that the game creates.

- **In-game interactions**: all social interactions that occur within the game world. These can be of two types:
  - **In-character interactions**: are social interactions framed by the fiction and narrative of the game. These are character to character social interactions that are coherent with the fiction and the narrative established by the game. Character actions and drives should be consistent
with the fantasy portrayed by the game and the character’s narrative goals.

- **Out-of-character interactions**: are social interactions that have the gameplay mechanics and dynamics as the core frame of reference. These are agent to agent interactions that are coherent with the gameplay rules. Agents are driven by the gameplay goals and their actions are not necessarily aligned with the game’s narrative. For example, players may use their avatar to steal a powerful item from another agent to get a gameplay advantage (e.g., deal more damage), while the motivations of the character to steal are not supported in the narrative. The main concern in this example is to maximize performance and points and not keep the agent in-character. Another example is the case of dialogue interactions whose content is about gameplay mechanics rather than characters’ speech.

- **Out-game interactions**: all social interactions that occur outside the game world. These can also be of two types:
  
  - **Interpersonal interactions**: are social interactions that take place between the players outside the game world. These are not conducted through the game world (e.g., are not performed through agents or objects in the game world and do not use in-game communication tools.). These are out-of-character interactions as well and are not subject to the fictional narrative nor the gameplay rules. These interactions often extend beyond the gameplay session. For example, players may keep discussing the game results for a while after finishing playing or start discussing the strategy for an incoming match.
  
  - **Cross-layer interactions**: are social interactions that engage game actors across different layers, mixing and bridging the context of interaction. These can be, for example, player to agent social interactions, when players do not have avatars and directly request actions from agents (e.g., their soldiers in a squad). Can also be character to player social interactions, when a character talks directly to the player, for example, to express frustrations about their decisions. In fact, all “out of character” social interactions involving game characters are inherently cross-layer interactions and take the character out of the narrative context. In this case, the meaning of its actions is no longer only bounded by the narrative and fiction but is also based on the other layers.
There are several social roles SIAs can take in terms of the gameplay layer. The first distinction is referent to the collaborative nature of the social interaction toward the player. SIAs may play roles as teammate, opponent, or be independent.

- **Teammate**—by taking this role, SIAs are committed to work together with the player. They perform tasks that are required by the challenges provided by the game. Typically, these tasks cannot be solved alone. Teammates share goals and all succeed if the goals are achieved. They can take team-specific sub-roles depending on the structure and task of the team. To play as a teammate, SIAs need to be able to understand the team's social context, shared goals, and plans and be able to execute the actions of the plan that typically require some kind of coordination.

- **Opponent**—by taking this role, SIAs are committed to obstruct players and try to avoid their success. Players need to overcome these SIAs to be able to progress in the game. The SIA in this role may also be racing with players to achieve victory for themselves. To perform this role, SIAs need to be able to understand the challenge they impersonate and be able to execute, and define, strategies to beat the players.

- **Independent or Neutral** are SIAs that have individual goals and purpose in the game world that are not strictly aligned to helping the player nor committed to obstructing their victory. The goals SIAs have in this role are not related to the goals players try to achieve to win the game, but these SIAs may help the players and join a coalition and team up with players for a while, if that fits their purpose. They may also oppose the players if their own goals are threatened or if they team up with the opponents in a coalition.

In the case where SIAs work together with players they can have different roles depending on the relation they have or build with players. They can play roles as companions, subordinates, advisers, or helpers.

- **Companions** accompany the journey of the player. Sometimes referred to as “sidekicks,” they are “on screen” with the players most of the time of the gameplay experience. They pay special attention to the goals and actions players try to perform but may have their own goals as well. Companions often have a presence in the narrative layer, as well, to support building a deep interpersonal relation with the player.

- **Subordinate** is an agent that perform tasks for the players. Also referred to as henchmen or minions. They mostly perform tasks and goals delegated by players. This involves a power relationship as players have control over
the goals that subordinates commit to. These agents have autonomy but only to fulfill the designated goals. They may be proactive, nevertheless, and autonomously take goals they believe are relevant to the player.

- *Adviser or helper* is a SIA that indirectly contributes to the efforts players make toward achieving their goals in the game. Advisers convey information to players about the game state and provide advice about gameplay actions and strategies. They can be specialized in certain areas of gameplay (e.g., economic, military, research, as in the Civilization game series). The information may be proactively suggested or only given when explicitly requested by players. They often introduce players to the game mechanics and support their learning about the game.

There are several SIAs in games that are independent and neutral regarding the players’ goals. They have their own goals in the game, and it cannot be assumed that they will have a benevolent attitude toward players’ requests. They are, nevertheless, important to conveying the gameplay experience and it is expected that players need to interact socially with these SIAs to progress in the game.

- *Provider* is a SIA that provides resources, information, services, and tools that players need to progress in the gameplay and fulfill their goals. The players get what they need after a successful interaction with providers. This means that they, typically, need to make an effort to succeed in the interaction. This may be a simple commercial exchange (e.g., if the provider is a shopkeeper) or may require some kind of negotiation. But the option to freely provide the resources after a simple contact is open as well. The effort in this case is the time spent to go to the provider. Providers may also unlock new gameplay options (e.g., avatar abilities).

- *Challenger* is a SIA that provides challenges to players (e.g., a quest giver). These are somehow similar to providers as they may provide rewards as well. But their main role is to explicitly define goals for players to follow. They may serve as “gatekeepers” that lock and unlock the game progression as they may have strong control over the goals that are open to the players. The challenger role may be taken by an opponent as it may raise a special confrontation goal for the player that is triggered once the player meets it for the first time (e.g., a boss battle).

- *Commentator* is a SIA that describes the gameplay action and may present an assessment of the gameplay results as well. Commentators are not at the service of the player as the advisers. Although, the information they provide can be useful to help and guide players. They present a shared view of the game
state to all the agents in the game world that can influence the gameplay decisions. However, they often serve the audience of the game (e.g., commenting on a football match), which includes players. In this sense, they have cross-layer agent–player interactions or may even be placed outside the game world (at the player layer).

- **Background** are SIAs that are used to bring social life to the game world. They do not influence the game progression but may be affected by it. These SIAs react to players and engage in social interactions if requested. Background SIAs provide context to the interactions with other SIAs and may depict and support understanding of the game’s social world. They often have a strong representation in the narrative layer to help enhance the social dimensions of the fictional world. Nevertheless, they are actors in the gameplay layer as they may constrain the gameplay actions players take (e.g., a player may not be allowed to kill an opponent in a public space).

The social roles SIAs assume can vary depending on the situation. A SIA may take more than one different social role at the same time. For example, a SIA may be a companion and adviser at the same time and may act as a subordinate or provider in other situations. Therefore, context is important. The same player can interact with the same agent in different situations, and each is driven by different social roles. Both the agents and the players need to understand the relation between context and roles.

We would like to highlight that, despite not being the focus of the discussion here, the narrative layer is crucial for the experience, in particular to support building social relations and emotional attachment with the SIA apart from the practical perspective of the gameplay layer. But we want to stress, as well, that to have good integration between the gameplay and the narrative experience of a game, it is important to reflect carefully and define the social roles the SIA take in terms of gameplay together with the narrative roles they will play. The gameplay is the unique aspect that differentiates games from other storytelling media.

### 27.3 History/Overview

To study the evolution of SIAs applied in games, it is important to acknowledge that the embodiment they assume is a critical factor to distinguishing them. As such, we divide our review of the past two decades of SIAs in games into two categories: **virtual SIA** (agents in digital environment) and **robotic SIA** (agents with a physical embodiment). Also, it is important to note that in this chapter we only refer to SIAs clearly used in games, both from academia and industry. And we do not include contributions in which SIAs are applied to Serious Games (as these are discussed
in Chapter 28 on “Serious Games with SIAs” [Gebhard et al. 2022] of this volume of this handbook) and we do not focus on the use of SIAs as narrative elements (as discussed in Chapter 26 on “Interactive Narratives and Story-telling” [Aylett 2022] of this volume of this handbook).

### 27.3.1 Virtual SIAs in Games

Since the birth of digital games, the game industry has presented players with increasingly more complex and believable environments. From simple 2D grid-based maps to large procedurally generated worlds, games present several opportunities for players to explore their landscape and interact with their entities. Among these game elements, SIAs are often used in commercial games to present its story. Still, there are cases where these social agents assume an important role in the gameplay layer, with a direct impact on the game systems.

In recent years, games evolved to support larger virtual worlds, more characters and, ultimately, larger and richer interactive spaces between players and SIAs. One of the most used approaches to supporting richer social interactions with SIAs is through verbal behavior. In games, SIAs' dialogue can be used to request actions from players, usually guiding them through the narrative or providing additional challenges to further explore the environment. However, they are also used as channels to share accessory information about the game experience and establish empathic and meaningful relationships between the player and the game elements.

SIAs' dialogue is often supported by the fictional story on the narrative layer, but the actual exchanges happen with agents in the gameplay layer. Some games use linear narratives that offer a near cinematic experience supported by well-written dialogue lines where SIAs engage in dialogue with other agents following a rigid script, as in Uncharted: Drake’s Fortune (Naughty Dog, 2007) or in The Last of Us (Naughty Dog, 2013). This approach reduces the player’s capability to influence the narrative flow in a discernible way. To create a false sense of agency, in Uncharted 4: Drake’s Fortune (Naughty Dog, 2016) certain conversations between the player and SIAs requested dialogue options with no impact on the storyline but still demanding the player’s input at the gameplay layer. But in linear narratives, SIAs are also used to mimic a living world by promoting social interactions between agents. For instance, in The Last of Us (Naughty Dog, 2013) and Far Cry 4 (Ubisoft, 2014), developers were able to create interesting conversations between SIAs by adapting the dialogue based on their context, the player’s actions, and other game events. On the latter, the player’s companion is also capable of assessing possible threats and channel that information to the player using dialogue. In BioShock Infinite (Irrational Games, 2013), a system that used the player’s companion to highlight...
relevant interesting locations was introduced. In addition to her dialogue lines, an intelligent system that positioned this SIA in the game world was used to shift players’ attention to unexplored parts of the scene.

Other games rely on richer dialogue interactions with SIAs that are supported by dialogue systems that offer multiple conversational choices. One of the most common effects of these conversational gameplay is the impact on the narrative’s ending, usually organized in branches. BioWare’s Star Wars: Knights of the Old Republic series, which started in 2003, introduced SIAs that spontaneously engaged or interrupted players and other agents’ conversations in a meaningful and impactful way to the experience. Several of Quantic Dream’s titles, such as Fahrenheit: Indigo Prophecy (2005), Heavy Rain (2010), or Detroit: Become Human (2018), rely on multiple dialogue choices to affect short-term interaction with other SIAs and the environment, as well as long-term relationships, leading to multiple game endings. Although the player is presented with a reduced set of options every time the game demands for a dialogue action, its limited range of affordances hides a rather complex branching narrative that is disclosed throughout the gameplay. Some open world games, such as Fallout 4 (Bethesda Game Studios, 2015) and Assassin’s Creed: Origins (Ubisoft, 2017), populate their worlds with unique agents that can engage the player in conversations that reveal information about game events (e.g., such as side-quests) that expand the gameplay space.

Using the player actions toward other SIA, including through dialogue systems, to affect game systems rather than the narrative is a different approach toward creating emerging experiences. Both Alpha Protocol (Obsidian Entertainment, 2010) and Fallout: New Vegas (Obsidian Entertainment, 2010) use a reputation system that models player’s action toward other game characters that lead to different social interactions with agents and their faction’s members. In The Elder Scrolls IV: Oblivion (Bethesda Game Studios, 2006), agents’ behavior is guided by a morality system that affects their obedience to social norms depending on the need to satisfy their personal goals. While playing Anthem (BioWare, 2019), the player is offered dialogue options during conversations that affect how other SIAs, and respective factions, perceive the player’s reputation, limiting the number of resources available. In Dying Light 2 (Techland, TBA, 2022), the game agents are managed by a morality system that controls the social interactions with the player, filtering the resources provided to the player, the combat opportunities, among other restrictions that directly affect the gameplay. These systemic approaches enable the player to have a more direct contribution in shaping the gameplay experience.

Other titles elevate the dialogue between the player and other SIAs to a core mechanic of the gameplay. One particular game that excels at using dialogue
systems to affect the gameplay is L.A. Noire (Team Bondi, 2011). The player’s character is a detective and a significant portion of the interaction with SIAs (witnesses and suspects) relies on questioning them to expose lies. On one hand, the player must gather information about the crimes by questioning them about the evidence collected. On the other hand, the player must challenge the SIA statements when they are lying by watching their verbal (the information) and nonverbal behavior (facial expression and posture). Besides using dialogue system, games also rely on parser-based solution, as is the case with Façade [Mateas and Stern 2003] and Event[0] (Ocelot Society, 2016). Façade is a well-known interactive drama that uses a parser-based solution to handle player’s written input during a conversation with two game characters [Mateas and Stern 2003]. The player’s goal is to save the marriage of the two SIAs by interacting with them through open dialogue. The game has several endings that can be reached based on the player’s interventions and the characters’ reaction to them (modeled by a drama manager). The interactive space affords several possibilities for the player as the social relationship between the two SIAs changes during the game. The commercial game Event[0] (Ocelot Society, 2016) also used a parser-based approach to consume the player’s input when interacting with an AI managing a spaceship. During the game, the player must communicate with the agent controlling the ship’s systems through written commands. These interactions not only modify the environment but also affect the relationship between the player and the agent.

Besides the conversational mechanisms of game agents, other social capabilities have been endowed to SIAs in games. For instance, in Black & White (Lionhead Studios, 2001) the creatures learn from the players actions and adjust their interactions with the environment and its characters accordingly. In Alien Isolation (Creative Assembly, 2014), throughout the game the main enemy seems to evolve by learning new approaches that are unlocked in the SIAs behavior trees.

In combat scenarios, agents can also coordinate their efforts by assuming specific combat roles and adhering actions from other SIA. In F.E.A.R. (Monolith Productions, 2005), enemies exhibited a squad behavior using Goal-Oriented Action Planning (GOAP) to coordinate attacks that show dynamic and coherent behavior among several SIA. In Halo 3 (Bungie, 2007), enemies and allies are recruited to join combat tasks as necessary and the collective behavior of each bundle of SIAs helps establishing a sense of coordinated action. The enemy SIAs in The Last of Us (Naughty Dog, 2013) can coordinate ambushes on the player where a coordinator agent assigns different roles to each agent in combat (e.g., flanker). Dishonored 2 (Arkane Studios, 2016) also has coordinated crews of SIAs that join their efforts to intelligently searching or attacking the player. Group interactions were also addressed by Prada and Paiva [2009] in the Perfect Circle game. In the game, one
player solves a series of puzzle-like challenges together with four SIA. They need to cooperate and coordinate decisions about the best course of action. This research stressed the importance of addressing the socioemotional dimensions of the group dynamics together with the instrumental actions.

Another approach to strengthen the social relationship between agents and players is to design SIAs that exhibit human-like capabilities, such as emotions and personalities. Each NPC in Far Cry 4 (Ubisoft, 2014) had a micropersonality that characterized the NPC’s motivations, needs, and desires and helped in creating an illusion of agency when these SIAs interacted with each other. In Event[0] (Ocelot Society, 2016), the AI character models its stress level and affection toward the player, adapting the social interactions between them accordingly. Similarly, the alien in Alien Isolation (Creative Assembly, 2014) models the player’s stress level to balance the gameplay between stressful and relaxed moments. Researchers have also explored game agents with complex decision-making mechanisms that promote interesting interactions. Using Fallout 3 (Bethesda Game Studios, 2008) modding capabilities, researchers have deployed a canine companion capable of identifying the intent behind the player’s actions and give advice [Doirado and Martinho 2010].

In other games, SIAs can have a long-term social relationship with the player while going on with their lives. The agents populating the Hitman (IO Interactive, 2016) world exhibit situational awareness by interacting with other SIAs based on their context when performing their routines. In Red Dead Redemption 2 (Rockstar, 2018), all SIAs have daily routines and, when encountering other characters, the interactions are influenced by past events and exchanges with the player. Both previous games also highlight the value of proxemics, namely when norms are violated, such as trespassing private property. The SIAs in these games will adjust their actions based on the social distancing between them and the player.

In game genres such as strategy, management, and simulation, the player does not directly interact with SIA. However, he/she can manipulate the environment, thus affecting the social dynamics of the population. To an extent, the NPCs are capable of socially interacting with each other and the world but are unaware of the presence of the player. In The Sims series (Maxis), the SIAs not only have a complex human needs model but also were able to interact with each other and allow for emergent social relations to unfold. Besides the daily routines, such as eating or sleeping, the characters could establish social relationships with other NPCs based on their mood and personality. Other games such as Sim City (Maxis, 1989) or Cities: Skylines (Colossal Order, 2015) invites the player to modify the game environment with no interaction over the NPCs in the world. Rather, the player must rely on these modifications to shape the physical elements of the game, thus
defining the social dynamics of the populations and, ultimately, influencing their interactions. In Prom Week, the player must satisfy social goals for each level in order to progress in the game [McCoy et al. 2011]. These goals are satisfied when the interactions between the NPCs, which are triggered by the player’s input, lead to a certain social state. The SIAs rely on the Comme il Faut (CiF) model to manage the social physics of the game [McCoy et al. 2010]. CiF was also employed in mods for commercial games such as The Elder Scrolls V: Skyrim (Bethesda Game Studios, 2011) and in Conan Exiles (Funcom, 2018) to improve the social interactions between NPCs [Guimaraes et al. 2017] as well as extending the original model with an emotional appraisal mechanism [Morais et al. 2019] toward creating more believable characters.

Other games use social simulations techniques to generate worlds before gameplay. Then, throughout the game, the player can interact with the game characters and experience the resulting social environment. Dwarf Fortress (Bay 12 Games, 2006) uses several techniques to procedurally generate populations for the player to experience. The underlying complexity of the generation includes models for the relationships of the NPCs that can be contemplated during gameplay and a knowledge propagation system that generated plots and events to be explored by the player while interacting with the world and its social beings. Another example of a gameplay experience that revolves around knowledge propagation is Talk of the Town [Ryan et al. 2015]. The communication between NPCs are subject to several mechanisms that shape the exchanges between characters, in particular social phenomena such as lying or eavesdropping.

Besides the systems that manage the opponent civilizations, on the Sid Meier’s Civilization series NPCs are used as advisers in different game-related subjects (e.g., military strategies, financial policies). Additionally, the player can interact with SIAs that represent enemy civilizations through in-game combat or diplomacy.

### 27.3.2 Robotic SIAs in Games

Social robots have rarely been used in commercial games (note that we are excluding toys in this analysis, where the use of robots is more common). Yet, during the last two decades several researchers used games to explore the interactive space between human players and social robots. These robotic players are usually placed in games to support experimental studies and are capable of promoting a small set of interactions. However, this type of SIA tends to assume a place not fully explored in game environments with their virtual counterparts.

Understanding the impact of SIAs’ *embodiment* in games has been explored in several research works in social robotics. In 2003, Christoph Bartneck explored how
the placement of a robotic character on a “future intelligent home” would affect
the enjoyment of interactions in the household [Bartneck 2003]. A game setting
was used to conduct an experimental study where participants would have to trade
stamps with an eMuu SIA (robotic or digital embodiment). The results suggest that
participants were more willing to forgive the robotic player’s errors when compared
to the digital counterpart.

Also, in Komatsu and Abe [2008] researchers conducted an experimental study
to investigate if the presence of a robotic SIA would persuade participants. All par-
ticipants would play Nintendo’s Picross game for a brief period of time alongside
a social agent (either an embodied or a digital version of PaPeRo). Then, particip-
ants with the robotic partner would be interrupted and the agent was replaced by
the digital version. After 10 minutes, the agent would ask the participants if they
wanted to play another game. The results suggest that players were more willing to
change games when the robotic partner was physically present and the backstory
was read. Using a chess game, researchers studied how the opponent’s embed-
ment (robotic and digital version of iCat) would affect the participants enjoyment
[Pereira et al. 2008]. The results indicate that players’ enjoyment is higher when
playing against a robotic partner. Using the Keepon robot, researchers studied if
the embodiment of a conversational robotic tutor (physical, digital, none) had an
effect on the performance of humans while solving puzzles [Leyzberg et al. 2012].
The results suggest that using a robot increased the participants’ learning gains
over the course of the experiment. In Wrobel et al. [2013], the authors conducted a
user study to assess elder users’ enjoyment while playing a card game with a social
agent. Wrobel et al. used the game of trivia StimCards to explore how elderly peo-
ple engaged with a laptop (no embodiment), a virtual agent (digital embodiment),
and a robot (physical embodiment) while playing the card game and receiving feed-
back. The results show that the robotic version was preferred to the virtual agent
due to its physical presence.

To investigate not only the impact of the embodiment but also the physical
presence SIA’s had on player’s perception of the robot, researchers extended the
possible modalities to include real-time videos of robotic players (through tele-
presence). Using the classic Tower of Hanoi puzzle, Wainer et al. [2006] studied
the human’s performance and perception of social interactions while playing a
game with a robot (ActivMedia Pioneer). The findings indicate that physical
embodiment can affect the perception of a social agent’s capabilities and the
player’s enjoyment of a task. To further explore the role of “material embodi-
ment,” the authors conducted another experimental study using the same scenario
[Wainer et al. 2007]. The results demonstrated that the physical robot was inter-
preted as more helpful, watchful, and enjoyable than the tele-present and digital
versions. Also, participants identified the embodied versions of the robots as more perceptive and appealing. In Komatsu and Kuki [2009], the authors concluded that players who interacted with a robotic agent followed by its digital version attributed the same character and personality to the robotic and digital versions.

To deploy social board game opponents, Pereira et al. [2012] identified several design guidelines for developing socially present artificial players. The guiding factors identified were used to develop a scenario using the popular game Risk where three participants play against a robotic player using the EMYS robot. This scenario allowed researchers to conduct an experimental study that verified that when the guidelines proposed were applied to the design of the SIA, participants’ perceived social presence improved [Pereira et al. 2014]. Similarly, in Fischbach et al. [2018] researchers design an interactive tabletop game with a robotic game master that can assume multiple roles during the game and promote collaborations between players.

The effect of emotional expressions by social robots has also been researched in the context of game applications. Bartneck [2003] conducted an experimental study to understand if robots’ emotional expressiveness was relevant during a negotiation game. The results of the study suggest that the enjoyment of the player increased when the robotic character expressed emotional responses. In 2008, Leite et al. designed and developed a game scenario that placed a social robot capable of exhibiting emotional behavior (iCat) as the player of a chess game. Using this scenario, a preliminary study that brought children and robots together to play a game of chess revealed that the human player’s perception of the game increased when the robot’s affective state was aligned with the game state [Leite et al. 2008a]. The authors used the same scenario to study how the agent’s emotional expressiveness [Leite et al. 2008b] influenced a human players’ performance. The results show that players performed better when the social robot displayed emotional behavior. Using an iCat robot to play a card guessing game, researchers studied how children in different age groups would express emotions based on the game’s outcome [Shahid et al. 2010]. The results suggest that young kids are more expressive than older ones when collaborating with a robot.

Using the Coin and Strings board game, researchers studied how a human player would perceive its social performance when playing against an emotional social robot (EMYS) [Petisca et al. 2015]. The authors concluded that a robotic opponent capable of sharing emotion, might hinder the human–robot interaction and the social perception of the robot. To explore the effect of emotional sharing and competence while executing a task, the authors conducted another study that focused on the effect of each one of the previous factors on the perception of a robotic opponent in a game [Petisca et al. 2016]. The findings suggest that when
Chapter 27  Socially Interactive Agents in Games

placing robotic players in a very competitive scenario, such as this board game, the robot’s emotional responses might be ignored by the player. Still, when the competence of the robot was lower, the participants perceived it as more concerned about the player when compared to the high competent variant.

The use of verbal and nonverbal behavior was also studied in the context of social robots playing games. Using AIBO (a pet robot), researchers created a scenario for the robot to learn how to play several games based on the feedback provided by a human participant [Austermann and Yamada 2008]. The scenario was used to study how different modalities of feedback (speech, gesture, and touch) affect the perception of the robot learning capabilities (measured with questionnaires). The results show that speech and touch are the most natural ways of providing feedback. Komatsu et al. [2010] used a treasure hunting game to explore if subtle expressions, such as fast sounds, when expressed by a robot companion (MindStorms) would affect a human player’s actions. The authors conducted an experimental study that shows subtle expressions as a successful mechanism to convey a robot internal state and also suggest that players adapted their decisions based on the type of the subtle expression used. Similarly, in Leite et al. [2013], the authors studied different types of empathic behaviors, and their findings suggest that facial expressions and verbal utterances positively affected the perception of friendship toward the robot.

In Lehman and Al Moubayed [2015], the authors elaborate on the 2D platform game Mole Madness used to study the interaction between a robotic player (using FurHat) that can interact with a human player while both control the game character through speech commands. Researchers used this game scenario to conduct an exploratory study on children’s verbal and acoustic synchrony while exploring the engagement of speech-controlled gameplay [Chaspari et al. 2015]. Additional studies to explore how prosodic patterns would influence children engagement during a game were conducted [Chaspari and Lehman 2016, Sadouohi et al. 2017]. Both experimental studies indicate that when a robot synchronize its speech prosody with the children from the start of the game, the engagement is higher when compared to the robots that did not exhibit any synchrony from the start.

Using the humanoid robot NAO, researchers studied how robots’ different social categories would affect human behavior during a card game [Häring et al. 2014]. In the scenario, all players had to interact with two robots: one belonging to participants’ social in-group and the other one being a social out-group member. They conducted an experimental study in which participants were either asked to collaborate with the in-group and compete with the out-group (congruent condition) or vice-versa (incongruent condition). Their findings indicate
that players evaluated their actions more positively in the congruent condition and more challenging in the incongruent.

In Shahid et al. [2014], the authors explored the effects of playing alone, with a robot, or with a friend in the interaction with social partners. The subjects were kids from different cultural backgrounds (Pakistani and Dutch). On the one hand, the results show that Pakistani kids enjoyed the collaboration more than the Dutch ones. On the other hand, participants had an increased enjoyment and expressiveness while playing with robots than when playing alone, but less than when playing with a friend.

Toward studying how players socially interact with robots with different roles (partners or opponents) and goals (relationship-driven or competitive), researchers used the Sueca card game with multiple robotic players. In Oliveira et al. [2018], an experimental study was conducted, and its results suggest that players gazed more toward the competitive robot when it was their partner, while players gazed more toward the relationship-driven robot when it was their opponent. The authors also identified a higher frequency of socioemotional support toward human (when compared to robots) and partners (when compared to opponents). Using the same scenario, researchers explored group-based emotions and membership preferences in teams of humans and robots [Correia et al. 2018]. The authors concluded that group-based (that reflect the group actions) were able to emphasize trust and group identification more than the individual robot (with emotions that focus on individual behaviors).

Toward promoting collaborations in teams of humans through the presence of a robot, researchers placed a social robot (using Keepon) alongside a team of two playing a game called build-a-rocket [Strohkorb et al. 2016]. In this scenario, players had to collaborate to conclude the task while the robot periodically participates with interventions about the task or about the interpersonal relationship of the players. The results of an experimental study with children suggest that a robot with task-oriented interventions help to increase the participant’s scores while a more relationship-focused robot improves participants’ perception of their performance.

In Correia et al. [2019], an experimental study using the competitive and relationship-driven robots were used to determine the human player’s preferences toward the robots. The findings indicate that when no previous interaction with the robot occurred, participants preferred the relationship-driven robot. However, when participants had already played with the robot, their preference is influenced by other factors (e.g., competitiveness of the human player or the game outcome).
Placing iCat as a bystander that interacted with two humans playing a chess game, researchers concluded that, when the robot displayed empathic behaviors, players rated their companionship higher when compared to the ones that interacted with the neutral robot [Leite et al. 2010]. In a real-world environment with children, another experimental study was conducted with a fully autonomous version of iCat capable of recognizing facial expressions and exhibiting appropriate empathic behavior [Leite et al. 2012]. The results support previous findings, indicating that the empathic cues of the robot had a positive effect on the perception of the robot.

Using a trick-taking card game called Sueca [Correia et al. 2016], researchers conducted two experimental studies (one with elderly participants [Correia et al. 2016] and another with younger players [Correia et al. 2017]) to investigate the humans trust levels toward robots. Both studies suggest that humans are capable of trusting a robot as a game partner in a card game. However, trust levels vary based on previous interactions with the artificial partner: players that had already interacted with the robot increase their trust levels more than participants that had already interacted with human players.

To understand what social interactions between children and robots were beneficial to establish a long-term engagement, researchers designed game setting that placed a NAO robot playing Snakes and Ladders with children [Ahmad et al. 2017]. Using this game, an experiment was conducted, and the results indicate that adaptations based on the game state had no effect on the long-term interaction. However, emotion-based and memory-based adaptions had an effect on the sustainability of the relationship between child and robot.

Using a NAO robot and a railroad route construction game, researchers explored how a robot's vulnerability would affect players' trust [Strohkorb Sebo et al. 2018]. Using this scenario, an experimental study was conducted with groups of four players (three humans and the robot) and evidence suggest that when the robot said vulnerable utterances, the participants showed more trust-related behaviors toward the human partners. To study how robot's trust violation and subsequent repair would affect the players’ relation toward the robots [Sebo et al. 2019], the authors conducted another study based on the previous game that indicates that denying culpability might yield benefits while repairing human–robot trust, but not when deception is involved.

### 27.4 Similarities and Differences between Virtual and Robotic SIA

Both fields of virtual and robotic SIAs acknowledge the added value of SIAs for the players' experience in games. But the literature on social robots in games has not yet matched the number of commercial and research work that deploy virtual
agents in games. The application of SIAs in virtual worlds has been addressed for longer, but in recent years there has been increased interest in research on the use of robots in games as well. Both areas have been through significant changes in the past 20 years due to technological advances. Still, their research directions do not often align with each other, but both can benefit from wider awareness of each other’s contributions and achievements.

In Social Robotics, researchers focused their efforts toward studying the effects of using an embodied player on the players’ experience. But, with a few exceptions, the focus of study is often the relationship between players and the robots and not the perceptions and attitudes toward the gameplay itself. The interaction with social robots that play the game through an avatar or that play other roles rather than of an artificial player are not much explored. Robots sometimes take other roles together with the player role (e.g., the role of companion), but the social interactions are framed at the player layer, in the real world, and do not enter the gameplay and narrative layers of the interaction experience, for example, do not cross the magic circle into the game world.

Contrarily, Virtual Agents in games take more roles in the gameplay and narrative layers. However, they are not used as artificial players with a representation in the player layer that players can relate to outside the game world. There is extensive work on AI to play games, but the focus is on the creation of mechanisms to solve the game without any regard for the embodiment of the artificial player nor the social relations it can build with players, hence cannot be considered as SIA.

There is a natural bias of the research target by each field given the nature of the embodiment of the SIA, which is either strongly grounded in the real world or in the game’s virtual world. It is hard to decouple the use of SIAs from the embodiment they assume.

Both fields address the SIAs embodiment, virtual or physical, and take the expressivity of the body as an important factor. Both found results on the impact of the embodiment in the game experience. When it concerns artificial players, a physical embodiment seems to be preferable. But comparisons for other gameplay roles have yet to be performed. This will demand a different view of robot roles in games and the nature of games implementation will need to change as well. It needs to more easily integrate robots and other physical objects as gameplay elements. In turn, the degrees of freedom of the embodiment in virtual agents are higher and the expressivity of the body has been explored from an artistic point of view in addition to the strict social functions.

The complexity of the games used in the two fields is different. The games used in social robotics are much simpler in terms of gameplay. And the competences the SIAs demonstrate are also simpler, for example, the dialogue interactions are more
restricted. This is, on one hand, justified by the higher focus on the embodiment in the social robotics field that address less often the gameplay mechanics, but it is also because of the higher difficulty in the realization of more complex behavior and context awareness for social robots compared to virtual agents. The creation of rich interactive environments for games is also easier in a virtual domain compared to the real world.

Nevertheless, both fields give importance to the social intelligence of the SIAs and the emotional dimensions of social interaction. In fact, all matters discussed in the previous chapters of volume 1 of this handbook [Lugrin et al. 2021], in particular, in Parts I, II and III, and this volume of this handbook, in particular Part IV, are quite important for SIAs in games, regardless of the type of embodiment. For example, social cognition, expressive and nonverbal behavior, and models of interactivity discuss work that can be used to empower the capabilities of SIAs in games.

There are still many open opportunities for the research of SIAs in games. The whole interactive space that covers the range of possible affordances that SIA, both virtual and robotic, can bring to games while performing the roles discussed in Section 27.2 has not yet been fully explored in game research, although commercial games make use of such SIA. Part is covered in domains that we excluded in our analysis, for example, serious games and pedagogical agents (discussed in Part V of this volume) conduct research that fits well in social roles, such as advisers and companions. But, for some other social roles, the research on SIAs for games is not extensive. It can take inspiration from other fields, for example, multiagent systems, that deal with teamwork, decision making, and coordination, or social psychology that define models for social behavior, and game design and user experience research in general. In fact, the research efforts need to bring together AI with experience design and focus on the contributions of SIAs for the gameplay dynamics.

27.5 Current Challenges
The roles discussed in Section 27.2 have been used in games but not all have been extensively explored and addressed by research. Therefore, the interaction capabilities of SIAs in such social roles are typically simple. This limits the interaction space available to players, which reduces their sense of freedom and autonomy while interacting with such SIA. This affects players’ perceptions about the SIA, their abilities as social actors, and the quality of the social immersion, affecting the quality of the overall game experience.

There are many open challenges regarding the capabilities needed to perform well, autonomously, the different gameplay social roles. For example, more
research is needed to make SIAs play well the role of teammates and to be perceived as autonomous and trustworthy partners. Research is needed for opponents as well, for example, to make them able to present to the players gameplay challenges of a social kind, involving lying, persuasion, and taking into account social dynamics such as in-group and out-group bias, for example. The neutral SIAs can gain relevance in the game experience if they display stronger autonomous social behavior as well. They can show social motivations and goals, for example, to gain status in a group or to pursue a relationship with another SIA, and they can be more selective about the collations they enter and the support they give to players. This latter case requires memory and moral judgement mechanisms to avoid, for example, having SIAs blindly helping players that have been mistreating them and their friends.

In general, that are big challenges for the creation of SIAs with good social behavior in games, in particular, regarding the ability to understand social context, be aware of and understand social reality, and be able to adapt behavior to the context and other game actors (other SIAs and players). Adaptation to the player’s goals is one of the current research trends. But adaptation to the social context based on all the SIAs present in a gameplay situation also raised some attention. This may allow SIAs to change attitude toward players when new members join the team and provide means for SIAs to change social roles if they can take more than one.

Another major challenge SIAs face in games is the ability to show coherence in their behavior, and, at the same, time avoid repetitions that break their social believability and the players’ social immersion. Players often interact with SIAs for many hours during a game playthrough. In general, the creation of SIAs that deal well with long-term interactions is still a big and open challenge. Again, this will require good memory and situation assessment mechanisms. The communication with players and among SIAs, in a way that players can understand, is another open research challenge. This inherently implies dealing with Natural Language Processing (NLP) (for example, as discussed in Chapter 5 on “Natural Language Understanding in Socially Interactive Agents” [Pieraccini 2021] of volume 1 of this handbook [Lugrin et al. 2021]). Current NLP advances are opening new opportunities to SIAs. SIAs can have more flexible conversations with players, which have been typically quite restricted. For example, chat and voice interaction are being explored as novel gameplay interaction modes in games. This also opens the possibility for SIAs to participate in communication channels outside the game world (in the player layer). But this raises the additional challenge of supporting expression and understanding of informal conversations with lots of symbols, emoji, and non-normative vocabulary and spelling. However, if SIAs
take the role of artificial players, and take part in the player layer, participation in these communication channels will be important. Research in robotic SIAs that, typically, uses robots as artificial players in board game scenarios, where the verbal social interactions are essential for gameplay and to convey game experience, must address this challenge. But even if not pursuing NLP, addressing communication with SIAs in a way that is flexible and controllable at the same time is an open and crucial design and research challenge. It is also a challenge to extend this communication beyond the game session, addressing pre- and post-game interactions and dealing with repeated game sessions (e.g., playing the same game more than once).

There are specific additional challenges concerning the use of robots in games. Robots still need better capabilities to manipulate the physical components of a board game. Getting full understanding of the board game state is still challenging as well for more complex games with several types of components. But even in the case where the game is fully digital, which is typically realized using digital touch surfaces, the robots need to deal with social embodiment. For example, they need to be able to understand the social proxemics, keeping in mind the embodiment of other players and having appropriate gaze behavior.

One general challenge is to achieve a stronger collaboration between research and game developers. Researchers will gain richer environments and social interactions to explore, access to wide variety of players, and much interesting data to study. Game companies will gain knowledge and tools to understand better what autonomous SIAs can bring to games and become more equipped to create stronger SIAs in their games.

The use of autonomous SIAs comes as a game design challenge as well. Designers and developers need to make sure that the autonomy added to the SIAs does not go against the game design goals and in fact strengthens the design principles. Autonomy and complex AI behavior is not always needed. There is increasing interest in automated testing and validation tools to assess the complexity of the gameplay space and the way players play a game to make sure that the afforded interaction experience fits the design goals.

### Future Directions

Research of SIAs in games have yet many open challenges and opportunities for future directions. First of all, as stated before, there are social roles that are under-explored by research. A promising example is the use of virtual SIAs that take the role of embodied artificial players capable of engaging in social interactions in several activities related to playing games. These should include playing the game but
also taking part in the moments shared before and after playing the game. A further step would be turning such SIAs to true playmates that can play any game and, therefore, share heterogeneous gameplay experiences with players. This research would combine the research fields of SIAs and general gameplaying, such as Thuya the forever gameplayer [Gaina et al. 2019], for example.

In turn, there are several future directions for the use of robots in games. Researchers need to explore the use of robots in gameplay centered roles, placing robots as gameplay elements that are not artificial players. Robots could, for example, take the roles of subordinates (e.g., minions) that players control through delegation of actions or providers of resources or information engaging in negotiations with players. Note that these kind of roles have been explored in research but have not been taken into the domain of games, hence, have not been considered in terms of gameplay value. The use of robots this way, making them part of the game world, is promising for games like Live Action Role Playing Games (LARPG) and real life murder mystery games, which are becoming more popular.

The use of Augmented Reality (AR) is of great value to support the idea discussed above. A very interesting future direction is to combine research on AR with Social Robots in the context of games. AR can support the creation of gameplay experiences in physical environments that take advantage of the embodiment of robots and at the same time reduce the difficulties that robots still have with the manipulation of physical game elements. Players’ direct social interaction with the robot can be also enriched by AR, for example, by augmenting the views and perceptions over the state of the robots that could improve understanding (e.g., augmenting the robots’ expressivity) and by granting more contextualized options for the interactions, for example, parameterization that may be difficult using natural language and may be cumbersome using a classic graphical user interface.

Two related ideas are the use of social robots as avatars for players, exploring robots for the telepresence of players in physical games, and the use of social robots as players of fully digital games, by means of using a game controller as human players do. These ideas reflect the opportunities that SIAs with better capabilities can bring to game design. We believe that SIAs with enhanced social intelligence and autonomy, with both physical and virtual embodiment, will spark novel and completely new gameplay interactions that have not yet been explored by game designers. For example, there are opportunities to explore social interactions that are not built on top of explicit conversations, as SIAs become more competent in understanding player’s intentions. In a different perspective, SIAs can also be used in the process of game design by taking the role of co-creators to improve the creativity of the game designers. This would combine the research of SIAs with the research of procedural generation of games and computational creativity.
One promising research direction for the use of SIAs in games is sustained by a need expressed by players. Games are making more use of large open worlds, often procedurally generated (e.g., No Man’s Sky (Hello Games, 2018)), but these large worlds need to be populated by many SIAs to avoid conveying feelings of emptiness. The challenge is to procedurally generate large amounts of SIAs that show diverse and coherent social behavior and convey the feeling of organized social groups and populations that fit the generated worlds. Additionally, these SIAs should bring gameplay value as well as enhancing the social dimension of the game world.

Summary
SIAs are commonly used in games and their use and complexity has grown in the past years. SIAs typically take the role of virtual agents in virtual worlds, but more recently, as the field of social robotics emerged, interest in the use of robotic SIAs in games has increased.

SIAs have been used in games for several different reasons, taking several different social roles. They help to create and sustain the social dimensions of the game and improve the social immersion of players to convey better gameplay experience. SIAs can be artificial players, be part of the gameplay dynamics, and help to convey narrative meaning. In this chapter, we discussed the most common roles SIAs can take in games with focus on their contributions to gameplay.

We presented several examples in the use of SIAs in games both with virtual and physical embodiment. The research and use of SIAs has been different depending on the embodiment they take. In the case of Social Robotics, most research places the SIA as an artificial player, while that is not the case of Virtual Agents. Virtual Agents take more gameplay-related roles. Anyway, there are still many open opportunities for research, for both type of SIA, related to the agents’ capability of social performance. For example, games provide rich social worlds with several game actors and demand long-term interactions with diverse situations, which are still a big challenge. In fact, many SIAs in games still show simple behaviors that can be improved by using more advanced AI. With stronger SIAs, the gameplay offered to players can be enhanced in ways that are not yet explored. With better capabilities, SIAs can augment the interaction affordances and provide an enhanced sense of agency when players interact with SIA. Integrating such SIAs in the experience comes as a game design challenge as well.

It is important for the research of SIAs in games to mutually share the different perspectives that have been explored in different communities, in particular, virtual agents and social robotics, but also combining knowledge from other domains related to game AI in general, game design and interaction technology relevant for games.
References


I. Leite, C. Martinho, A. Pereira, and A. Paiva. 2008a. iCat: An affective game buddy based on anticipatory mechanisms. In Proceedings of the 7th International Joint Conference on


Serious Games with SIAs

Patrick Gebhard, Dimitra Tsovaltzi, Tanja Schneeberger, and Fabrizio Nunnari

Aiutami a fare da solo—Help me to do it myself
(Maria Montessori, [Montessori 2016, p. 69])

28.1 Motivation

Playing is exploring [Power 1999]. Both interconnect to the human drive of curiosity [Berlyne 1954, Kidd and Hayden 2015] and the essential emotional experience of security [Moser and von Zeppelin 1996]. Explorative playing is one way to learn about strategies, concepts, and knowledge. A game can be described by “the need to find or continue at once a response which is free within the limits set by the rules” [Caillois 2001, p. 8]. Such rules can be implicit or explicit. Implicit rules address social aspects of playing together, for example, being fair. Explicit rules might support games to be more available and understandable [Bente and Breuer 2009]. Usually, humans play games for having fun [Caillois 2001]. In general, the benefits a game could provide for a human player can be manifold. They depend on individual competencies, abilities, and the current situation [Garris et al. 2002] (See Chapter 27 on “Socially Interactive Agents in Games” [Prada and Rato 2022] of this volume of this handbook for further details).

Computer games are games that rely on technology for input, output, and processing logic. In such games, the logic realizes the type of game (e.g., action, puzzle, and strategy). While the game type defines the general gameplay, the genre describes the narrative concept. Typical game genres are fantasy, mystery, or war [Grace 2005]. An observation might be that the genre concept in games is related to the genre concept in theater, movie, and TV. Computer games support the human drive of exploring.

Serious games are computer games that are explicitly designed to examine and train specific skills or knowledge. The concept of gamification is different from that for serious games. Gamification “refers to the use [...] of design [...] elements
Serious games are explicitly designed to serve a serious purpose. Such games can be categorized by application areas (e.g., education [Prensky 2003], mental or physical health care [Macedonia 2008, Sawyer 2008], and training [Blackman 2005]). This categorization is related to the genre concept but misses a clear distinction of application area boundaries [Susi et al. 2007, Ritterfeld et al. 2009, p. 10, Laamarti et al. 2014] (See also Chapter 27 on “Socially Interactive Agents in Games” [Prada and Rato 2022] of this volume of this handbook for further details). A serious game’s application area defines its general narrative and content. Usually, serious games are evaluated according to how well users can improve skills by using them (Section 28.4). This chapter discusses serious games in the application area of training social skills employing Socially Interactive Agents (SIAs) with a focus on education and learning.

Why serious games with SIAs? We live in a social world. Social life and social interaction come with specific rules. Some are explicit, some are implicit. Both are defined by societies and their cohorts (See Chapter 13 on “Culture for Socially Interactive Agents” [Lugrin and Rehm 2021] of volume 1 of this handbook [Lugrin et al. 2021] for further details). For the training of social skills, serious games with SIAs can be a game-changer for many individuals and even our society as a whole. To bring social training in serious games closer to a human natural interaction experience, the employment of SIAs is mandatory. SIAs implicitly convey social values and norms interactively. Technology-wise, such agents are (partially) autonomous computer programs consisting of various software modules and hardware (sensor and actuator) components [Vinayagamoorthy et al. 2006]. SIAs can be seen as a representative of the used software and hardware. They denote the main interface of the interaction of and simulate human-like interlocutors (Part II).

Pedagogical role play with such agents offers great promise for social skill training. It provides learners with a realistic but safe environment that enables them to train specific verbal and nonverbal behaviors. As an example, centered on empowerment and inclusion, such games can help individuals with special needs or parts of the population gain skills in a playful and fun way that can make the difference in their everyday lives. As a result, learners benefit from the game-like environment, which increases not only their enjoyment and motivation but also enables them to take a step back and reflect on their behavior if necessary.

While current serious games with SIAs analyze the user’s verbal and nonverbal behaviors for the purpose of the interaction, their primary objective is to expose users with socially challenging situations. Implicitly, they do aim at teaching users appropriate socio-emotional communication skills and norms directly (Section 28.3).
From a psychological point of view, SIAs are a powerful technology that can even trigger social emotions, for example, shame [Schneeberger et al. 2019]. This aspect can be explained by the advanced reciprocal and reactive interaction abilities of such agents that come close to human interaction abilities [e.g., Gebhard et al. 2019b]. In general, their current state of the art might be sufficient that such agents might be able to serve as a psychological transfer or projection object [Grant and Crawley 2002]. With respect to ethical, legal, and social implications (See Chapter 3 on “Social Reactions to Socially Interactive Agents and Their Ethical Implications” [Krämer and Manzeschke 2021] of volume 1 of this handbook [Lugrin et al. 2021] for further details), further interdisciplinary studies must be conducted to explore the phenomena and possibilities of game mechanics employed in educational settings, for example, employment for psychological wellbeing and treatment (Section 28.4).

The next sections cover a general overview of the most prominent learning concepts, the historical development of serious games with SIAs, as well as current challenges and future directions. Finally, the chapter ends with a summarization.

28.2 Concepts of Learning and User Experience

This section discusses concepts that are related to technology-supported learning in game-like environments. All of the addressed concepts apply to serious games with SIAs. The section starts with the general paradigms of motivation and learning by doing. Further, the concepts of collaboration, socialization, embodiment, immersion, storytelling, and interactive narrative (cf. storytelling) are considered. For each concept, the focus is on the individual learning experience. This personal experience is tightly connected to each subjective situational experience—hence, her or his own (internal) emotions.

As games have been perceived as very promising for learning for some time (Section 28.4), a variety of learning concepts have been applied to games [Nebel et al. 2016, Qian and Clark 2016]. This section mentions some of the most prominent ones to motivate and concentrate on new directions that currently seem promising, especially in the context of SIAs. We also consider systems that implement elements of games even if they do not call themselves “games.” They are relevant as predecessors of SIAs, especially in the context of learning, with emphasis either on the concepts of social, interactive, or software agents.

Games have appreciated the role of motivation [Deci and Ryan 1985] that technology can play early on [Garris et al. 2002, Arroyo et al. 2003, 2014], and continue to invest in this direction through concepts of creating meaningful socioemotional experiences, which require cognitive effort [Kazimoglu et al. 2012]. Among the educational technologies, games come under the category of systems
that concentrate on affective aspects, together with communities of learning and computer-supported collaborative systems [Näykki et al. 2019]. This notion derived to a large extent from children's play and their innate motivation that this is characterized by. This is connected to the need to explore and apprehend the world in which children emerge in their “pretend mode” [Fonagy and Target 1996]. Learning to walk, talk, understand physical properties, and social relations through role play and playing with dolls, to name a few. Similar emotional and cognitive modes can be observed in grown-ups in states of flow [Csikszentmihalyi 1990], or immersion [Barab et al. 2009], and learning through authentic experiences [Csikszentmihalyi 1991, Kolb et al. 2014]. These result in learning. Standard behavioristic elements of reward (and punishment) also aiming to increase motivation [Skinner 1948] were another prominent concept in games right from the beginning, as it was easily associated with the notion of “gaming” [Peirce et al. 2008]. Symbolic awards such as stars, medals, and trophies are used to motivate participation as an equivalent of school grades with its advantages and disadvantages. Awards like this have found their way in novel approaches for immersive, interactive social training games [Damian et al. 2015a].

Motivation and behavioristic principles are not a standalone concept for learning. Elements of them are combined with the implementation of learning theoretical approaches, especially of the constructivist tradition, which has often been interpreted in the learning context as “learning by doing” [Schank 1995]. Here, the idea is that learning is knowledge construction, and hence emphasis should be put on supporting this construction. In this paradigm, inquiry or discovery learning [Squire and Jan 2007] are primarily based on the idea of self-directed learning and teaching problem-solving and domain skills [Squire and Jan 2007]. To address potential shortcomings of self-directed learning related to cognitive overload, systems provide scaffolds to guide learners through their learning experience as needed [Hmelo-Silver et al. 2007]. Minecraft probably exemplifies this theory as it provides possibilities to play as well as to create games. The aspect of creating games understands itself as drawing from the notion of constructionist learning [Kafai and Burke 2015]. Developers have also created authoring tools to include domain experts and teachers in the development loop, for example, in SimQuest [De Jong et al. 2022]. Simulation (games) [Lean et al. 2006] have been used especially for MINT subjects and to support problem-solving to this end [White 1984, Hmelo-Silver et al. 2007, Tsovaltzi et al. 2010] and also belong to the constructivist paradigm. Interactive inquiry systems and simulations implement game elements because of the interaction [Garris et al. 2002]. They provide tools, for example, graphs or labs, depending on the phenomenon under investigation, which can simulate the results of changing the values of the parameters in the
phenomenon. However, such systems do not themselves provide fully gamified experience. Still, they are often integrated into games, allowing, for instance, to solve mysteries in groups by combining information and using argumentation structures to learn together [Squire and Jan 2007]. Human-like agents (intelligent virtual agents or avatars) are also used in such systems to provide a social context to learning.

Such systems with collaborative learning affordances and agents implement in particular social constructivism, which is an extension of constructivism emphasizing the key role of the social context and social relations in constructing knowledge and goes back to Vygotsky [1978]. Social constructivism is then commonly realized in such games as (computer-supported) collaborative learning [van Joolingen et al. 2005, Stahl et al. 2006]. In this context, agents have been used to motivate learning through a human figure, for instance, by taking into account gender issues in avatars [Arroyo et al. 2009] or by implementing successful human teaching tactics [Graesser et al. 2001]. Agents may guide learners through the tasks of a lesson, for example, by giving tips, present learning material and concepts to be learned, pose questions and queries to prompt the learning process, and they may do all this by directly interacting with the learner in the form of an agent. An agent typically cohabits a virtual environment. It may support both actions and utterance input by students without natural language dialogue. It can also demonstrate actions in the environment, gaze, and gesture to capture and direct the student’s attention [Rickel et al. 2000]. To increase the human-like effect, a considerable amount of effort was put early on to create agents that could behave like teachers. In addition to applying successful human teaching strategies, these agents demonstrate emotion and hold a natural language dialogue with the student [Graesser et al. 1999, 2001].

Socialization is another concept that has been drawing attention to describe and support learning processes, and its most popular understanding in this context is vicarious learning [Bandura 1991]. In vicarious learning, learners observe and learn practices as they occur in their social context. A major interest is to explain implicit learning processes in informal settings. Learning environments have tried to apply such aspects of learning, especially, for example, with regard to social roles [Johnson et al. 1998]. Moreover, games are often being situated in communities of practice [Lave 1991], which builds on the notion of the identity development feature of games [Beatty 2014].

Despite the significance of social aspects of learning, the value of emotion and the way it may be exploited for learning is only beginning to be appreciated, especially with regard to motivation and group interaction [Mullins et al. 2013, Polo et al. 2016, Tsovaltzi et al. 2017]. Considerations of empathic agents are also being
integrated for learning, moving away from simple motivation constructs to better support emotions [Arroyo et al. 2009]. However, detailed modeling of the emotions as they occur in and through the interaction is still to be developed, and current emotional accounts are mostly based on observable measures of the interaction itself that are easier to track [Järvenoja et al. 2017]. Explicit models of individual intrapsychic emotional regulation, their role in interacting with SIAs, the possibility of leveraging such agents to trigger emotion regulation processes, and their learning potential have to our knowledge not yet been harnessed to support learning processes.

The social-psychological theory of embodiment [Niedenthal et al. 2005] is the direction that has probably inspired games based on SIA the most. In this context, “embodiment refers both to actual bodily states and to simulations of experience in the brain’s modality-specific systems for perception, action, and introspection” [Niedenthal et al. 2005, p. 184]. Especially combined with situational learning [Lave 1991], it supported the creation of games that represent the world in as much detail as possible to depict situational elements of a domain [Barab et al. 2007, Gee 2008, Pellás 2014]. Avatars are then used as a surrogate of the learner-player, which is supposed to carry over the embodied functions of learning [Riedl and Bulitko 2013]. As such, avatars may bridge the gap between emotional research in embodiment [Niedenthal et al. 2005] and learning. An application of embodied learning is perspective-taking, an aspect that has been developing based on neuropsychological findings supporting shared physical perspective-taking to enable socio-emotional perspective-taking [Kessler and Thomson 2010, Cole et al. 2016] and can have various applications in learning. Avatars are used to measure physical perspective-taking by having humans consider and report on the view of a human-like avatar [Samson et al. 2010]. This work also aligns with research on the role of self-awareness in the development of empathic concern [Hastings et al. 2000] that provides theoretical grounding. However, with respect to sharing physical perspective-taking with avatars specifically, findings are not conclusive [Cole et al. 2016]. This line of research has important implications on the use of avatars to increase immersion as well as for training purposes. Both assume that the avatar that represents the human player in the virtual world manages the transfer of perspective and emotions [Fischer and Demiris 2019]—hence the term empathic agents [Gebhard et al. 2018b]. Roleplay with virtual agents is a popular alternative approach based on socio-emotional perspective-taking whose function relies on this effect [Wu et al. 2013]. It has been used to educate users about cultural sensitivity. Employing role play with IVAs that represent different cultures, users are supposed to develop a better understanding of other cultures. Eventually, the users are expected to develop intercultural empathy and reduce their negative
attitudes toward different cultures. An example of such a system has been developed within the eCute project: The objective of MIXER (Moderating Interactions for Cross-Cultural Empathic Relationships) is to enable users to experience emotions that are usually elicited during interactions between members of a different group [Aylett et al. 2014].

Similarly, interactive narrative or storytelling [Rouse et al. 2018] applies the notion of immersion (and hence requires perspective-taking). However, a prominent characteristic is also the narrative itself because of theoretical accounts that narrative underlies our ability to construct reality [Bruner 1991]. Narrative, hence, carries the possibility of changing the way we experience it. This is especially relevant for interactive learning experiences. There, the idea is that users follow a narrative and empathize with the characters. This experience is supposed to occur especially through the possibility to interactively intervene and change aspects of the narrative, for example, the plot. This experience is situated in 3D environments inhabited by often real-time rendered virtual agents, which can engage in dialogue, show gestures, and actions that influence the plot, including other characters [Mateas and Stern 2003]. Hence, users should feel like being part of the narrative and learn from the simulation of a first-person experience. Recent interactive narrative theories give a central role to emotional experience and the body as an interface between emotions and the outside world to explain the manifold experience space (cf. social-psychological theory of embodiment). This experience space may be created through interactive narrative systems and the possibility to engage in multiple perspectives through them [Knoller 2019]. As a result, a new class of immersive learning experiences can be created (Section 28.4).

Emotion regulation is becoming famous with the rise of embodiment, as emotions are considered embodied responses that signal personally relevant events that can act motivationally [McQuiggan et al. 2008, Gyurak et al. 2011]. Beyond what has been known before, we are moving into the age where emotion recognition and regulation are related to notions of identity, self, and other-awareness [Harrell and Lim 2017]. This awareness is key to creating a relationship in the sense of attachment [Gergely et al. 2002]. It is also related to considering situational aspects to create socially aware agents [Barrett et al. 2019]. Emotional regulation, for example, in the form of (self-)compassion [Diedrich et al. 2014], can mediate motivation through situation appraisal [Goldenberg et al. 2017]. Narrative-based games that aim to support metacognitive learning processes are now considered for emotion regulation [Chytirogloiu et al. 2013]. In this particular game, users are able to explore, by multiple-choice interactions, the causal relationship between an emotional reaction and the situation, consequences of emotional responses, and effects of self-regulation of emotions. The game aims at learning new strategies for
regulating emotions. Similarly, games have been used to train socially constructed attitudes, like eating habits [Ptakauskaitė et al. 2016]. As such, new concepts of supporting motivation, especially through SIA, is considering emotion regulation and have recently started to look at the interplay of empathy with values of motivation [Charrier et al. 2018, Gebhard et al. 2018a].

28.3 History and Overview

This section presents with selected examples an overview of how serious games with SIAs have evolved in the past two decades. The discussion starts with a brief mention of how technology is used to support instructions and learning processes.

The field of technologically supported instructions emerged in the 1980 years [Rieber 1996]. The work back then centered on the fundamental concept of play and related aspects. Although playing is an essential concept for the development of abilities, playing was generally considered as unserious. Playing was not respected, working instead was. Despite this misconception, technologically supported games emerged. They focused on the different categories of play, for example, progress, power, fantasy, and self. One of the first examples was inspired by Piaget’s view on concepts of how children develop and learn [Piaget 1976]. Today’s serious games increasingly demand computational intelligence and perceptual skills to best grasp the player’s attention, behavior, engagement, and game-progress while modeling often complex and demanding game environments. The set of computational features also includes (computational) emotional and social intelligence to allow a deeper understanding of the player on different motivational levels, and then optimally adapt the learning pace. Besides, due to the rapid development in mobile and body-worn sensor technology, games can be taken into players’ real lives. (Part V).

Concerning the use of SIAs, first systems emerge around 2000. New methods and technology and interdisciplinary insights led to new concepts and fields of application. Serious games with SIAs provide users with a unique interaction experience. Some design aspects can be connected to a positive effect on learning and motivation. Wilson et al. [2008] as well as Bedwell et al. [2012] identified eight attribute categories designers should be especially aware of when developing gamified environments: action language, assessment, conflict and challenge, control, environment, game fiction, human interaction, immersion, and rules/goals. All of them are highly relevant for the design of serious games with SIAs. However, the design of believable and consistent interaction experience with a SIA is an important factor, if not the most critical factor (Part II).

The presented approaches are clustered in three groups that are interlinked and build upon each other but with a different focus: (1) content and story, (2)
immersiveness and interaction, and (3) individual needs and adaptation. The shown serious games with SIAs in these three categories are built upon well-researched learning and user experience concepts, which are discussed in the section before.

28.3.1 Content and Story
This group of serious games with SIAs represents approaches that focus on the tasks modeling the content and stories and related concepts such as conveying individual and cultural values.

Around the year 2000, inspired by educational concepts, IVAs were employed to support learning processes. One of the first systems is Adele, a pedagogical agent that augments different web-based learning environments [Shaw et al. 1999]. These agents react to user actions (e.g., text input, multiple choice) and have background knowledge of the learning material, which they rely on during a learning session. Steve is a well-known example of such an agent [Rickel and Johnson 1998]. STEVE is used in a game-style learning environment to teach the operation of certain submarine equipment. The combination of methods from different research areas of computer science, such as intelligent tutoring systems, computer graphics, and agent architectures, allows animated agents to take on different roles or present learning units from different perspectives in order to motivate trainees and students. Steve explains the exact procedures for maintaining specific devices step by step. He does this either in the tutor or the classmate’s role, who asks questions and, in this way, teaches the learning units. The focus in the design of such agents is on representing the to be conveyed knowledge, the development of methods for the analysis of user actions, and the modeling of the pedagogical and didactic skills. The latter two were realized mostly by plan-based cognitive architectures. To increase user motivation, such agents were extended by a model of emotions. Thus, agents are able to represent emotions (e.g., in facial expressions) and exhibit an extended nonverbal communication behavior [Elliott and Brzezinski 1998]. Studies by van Mulken et al. [1998] show that complex facts are subjectively easier to understand and more entertaining due to the explanations by IVAs. Multiple IVAs can be used simultaneously in different roles for different topics [André et al. 2000].

A bit later in the 2000 years, sophisticated approaches to interactive automatic storytelling with IVAs appeared (See Chapter 26 on “Interactive Narrative and Storytelling” [Aylett 2022] of this volume of this handbook for further details). A seminal work for that category is the Sam system. The IVA is designed for learning literacy skills [Ryokai et al. 2003]. Sam has the appearance of a child and is in the role of an advanced student. Sam collaboratively tells stories in a sophisticated way in order to let children learn literacy skills. The Sam system used fully functional voice
recognition and artificial intelligence methods to manage the interaction with children who had no particular disorder to learn reading and writing. Although not intentionally designed as a serious game, the interactive drama Façade created by Mateas and Stern [2003] sets new impulses for serious games using the concept of interactive storytelling. The drama game is designed around relationships between persons, probably the most important aspect of human life. It provides users with believable observations of social and emotional dynamics. The game’s action language is intentionally designed based on text-based multiple-choice actions, which naturally slow down the game progress and give users time to overthink their next steps.

Other systems in this area allow an advanced exploration of individual social values and cultural values (See Chapter 13 on “Culture for Socially Interactive Agents” [Lugrin and Rehm 2021] of volume 1 of this handbook [Lugrin et al. 2021] for further details). An example includes the anti-bullying Game FearNot! that has been developed within the European-funded project eCircus [Aylett et al. 2005]. The project investigates how social learning may be enhanced through interactive roleplay with IVAs that establish empathetic relationships with the learners. It creates interactive stories in a virtual school with IVAs in the role of bullies, helpers, victims, and so forth. The children run through various bullying episodes, interact with the virtual agents after each episode, and provide advice to them. The benefit of educational roleplays of this kind lies in the fact that they promote reflective thinking. Results of a conducted evaluation [Sapouna et al. 2009] showed that the system had a positive effect on the children's abilities to cope with bullying. A similar approach is MIXER (Moderating Interactions for Cross-Cultural Empathic Relationships). The aim is enabling users to experience emotions that are usually elicited during interactions of members of a different group [Aylett et al. 2014]. To this end, children are confronted with scenarios in which IVAs appear to violate previously introduced rules in a game scenario. Such a situation leads inevitably to frustration and negative attitudes toward members of the other group. By interacting with MIXER, children are expected to learn to reflect on other groups’ behaviors and reconsider potentially existing prejudices against them. The setting was inspired by the card-game BARNGA, which has been successfully used for the cultural training of adults [Thiagarajan and Steinwachs 1990]. Other than the authors’ expectations, the MIXER game did not foster cultural awareness in children in a pilot study. The authors assumed that the designed learning objectives in MIXER were not appropriate for the age group, which could not cope with the negative rule-clash-based conflict.
28.3.2 **Immersiveness and Interaction**

This group of serious games with SIAs represents approaches that focus on creating immersive experiences that support a compelling interaction with SIAs.

The beginning of the new millennium saw the use of 3D visualization techniques to create more immersive interactive training simulations with IVAs. A common application area, for example, is the medical simulation of virtual patients and their reactions [Cavazza and Simo 2003]. The underlying system does simulate an internal patient model that also includes emotional states. Around the year 2005, such systems were used for virtual training and education purposes [Ieronutti and Chittaro 2005]. More recent approaches provide a more realistic simulation of internal (patient) models and provide more realistic and immersive experiences exploring social aspects [Deladisma et al. 2007, Ochs et al. 2016]. Another application area is military simulation. One prominent example is the Mission Rehearsal Exercise that uses IVAs in a physical environment that projects the 3D scene simulation on a curved wall. This speech-enabled interactive training system is supposed to train soldiers in tactically and strategically demanding situations that also involve negotiations with civilians [Hill et al. 2003]. Later systems include more detailed modeling negotiation strategies including simulation of trust [Traum et al. 2005] or societal medical challenges [Lourdeaux et al. 2019]. Other approaches focus on creating a more immersive interactive experience in educating users about historical circumstances and combining a fully 3D projectable environment, such as the CAVE [Jacobson et al. 2005], and techniques of interactive storytelling [Cavazza et al. 2007]. Later research systems provide an extended analysis of the user’s social cues during interaction and their interpretation concerning the user’s internal states [Bee et al. 2010]. According to evaluations, the technology in general found to be used in games in terms of performance and acceptance [Lugrin et al. 2010]. The interpretation of recent study results suggests that for storytelling SR, the natural display of social cues (e.g., gaze, nodding, and smiling) or emotions is important for creating immersive transportation of users into the story [Striepe and Lugrin 2017].

Besides investigating aspects of virtual simulations to increase the experience of immersiveness, physical interfaces, such as RFID-based recognition of objects and user action, were investigated. A commercial example is the EyeToy game Kinetic by Sony (Sony Computer Entertainment, 2005). This game employs two 3D IVAs to motivate and animate fellow players to perform gymnastic exercises interactively. In content-oriented applications, users can associate a certain topic with a corresponding agent, which is beneficial for general understanding. One example is the IVA of the Autostadt in Wolfsburg. There, two IVAs, Jara and Taron, inform
visitors of the Autostadt about their function while constructing a model car from segments. The detection of user interaction uses RFID-technology to sense which car segment has been placed on a construction table. In this playful way, knowledge about the car segments (e.g., front, cockpit, rear) and about IVA technology itself is conveyed [Ndiaye et al. 2005]. Based on this, more sophisticated simulations of IVAs, including situational appraisal and believable emotional behavior, are created. An example is the virtual Poker Game with two IVAs as game opponents [Gebhard et al. 2008].

The design of the interaction and the dialog with an IVA always includes how much control can be given to the user. A mixed-initiative dialog provides more freedom to the user. However, it also requires more sophisticated language understanding capabilities than system-initiative dialog. In Endrass et al. [2014], we compared system-initiative dialog with mixed-initiative dialog in a soap opera-like game environment that included a text input interface to enable users to communicate with virtual agents. The users preferred the mixed-initiative dialog over the system-initiative dialogue even though the mixed-initiative dialog was less robust. Apparently, the experiential advantages of mixed-initiative dialog compensated for the lower amount of accuracy in natural language understanding.

### 28.3.3 Individual Needs and Adaptation

This group of serious games with SIAs presents approaches that focus on adapting to users’ individual needs to support empathic interaction.

There are different approaches in which SIAs assist or train users individually as they adapt to individual needs (Part III). Typical areas of application are health and wellbeing, social skills, education, and motivation. The area of social training with SIAs emerged around the year 2000 [Badler 1997]. Most of the approaches focus on the fact that communication conveys social information and individual emotions. Both impact relationships, social togetherness, in private and professional life. Such games have seen a rapid evolution in recent years due to advances in the areas of social signal processing as well as improvements in the audio-visual rendering of IVAs. The games complement or even substitute traditional training approaches. The concept of empathy plays an important role in such training environments with IVAs [e.g., Paiva et al. 2004]. There are similar approaches in SRs, such as systems that train children on the autism spectrum to learn in an explorative game environment how to create social relationships [Dautenhahn 1999] or for more general use [Breazeal 2004, Feil-Seifer and Mataric 2005].

As a use case for SIAs, the health context has been getting research attention for about 15 years (See Chapter 24 on “Health-Related Applications of Socially Interactive Agents” [Bickmore 2022] of this volume of this handbook for further details).
One of the first systems is Fit Track with the IVA Laura [Bickmore et al. 2005]. Laura has the role of an exercise advisor that interacts with patients for one month on a daily basis to motivate them to exercise more. Laura was equipped with different effective patient-provider communication skills (e.g., empathy, social dialogue, nonverbal immediacy behaviors) to build and maintain good working relationships over multiple interactions. From the perspective of serious games, a study showed that using those relational behaviors significantly increases the working alliance and desire to continue working with the system. This work suggests that computer systems that interact with patients, especially those that engage patients in dialogue or long-term, repeated interactions, can benefit by explicitly designing in emotional and relational communication behavior (Chapter 23 on “Socially Interactive Agents for Supporting Aging” [Ghafurian et al. 2022] of this volume of this handbook).

Techniques for the recognition of human socio-emotional behaviors and their synthesis using IVAs have been investigated in various cases: They cover the training of social skills in (1) group interactions in various application domains [Damian et al. 2015b, Lugrin et al. 2016, Chollet et al. 2018] and (2) difficult face-to-face interactions, for example, medical situations [Johnsen et al. 2005] that even train for the breaking of bad news [Ochs et al. 2016], job interview situations [Anderson et al. 2013, Hoque et al. 2013, Gebhard et al. 2018a], or for personal therapeutical assistance [DeVault et al. 2014, Gebhard et al. 2019a].

Within the project ASC-Inclusion [Schuller et al. 2015], techniques for the recognition of human socio-emotional behaviors have been employed to help children with autism improve their socio-emotional communication skills (See Chapter 25 on “Autism and Socially Interactive Agents” [Nadel et al. 2022] of this volume of this handbook for further details). IVAs help children to learn how emotions can be expressed and recognized via gestures and facial and vocal expressions in a virtual game world. A requirement analysis revealed the need to incorporate an appropriate incentive system to keep children engaged. Therefore, the authors implemented a monetary system that rewarded children with virtual money. Attention Deficit Hyperactivity Disorder (ADHD) can be observed mostly in children characterized by inattention, hyperactivity, and impulsivity and is caused by issues in the frontal/striatum brain areas [Cubillo et al. 2010]. Virtual reality training games provide such children with engaging and user-friendly environments that improve their motivation [Peijnenborgh et al. 2016]. The training for children affected by some pathology requires a more fine-grained control of the SIA, compared to, for example, the Sam system [Ryokai et al. 2003]. As reported by Parsons and Cobb [2011], children with autism are “often focusing on visual detail or parts rather than the whole.” This suggests that SIAs should be designed by minimizing details that
have no functional goals. The recent BRAVO research project [Barba et al. 2019, Nunnari et al. 2019] investigates an SIA system that is designed as a companion in the roles of stimulating therapy guide and a playmate. The key to this work is the design of the interaction experience and the SIA role based on the expert knowledge of psychologists and therapists. Also, the SIA’s facial expression is designed according to the results of the systematic evaluation of the recognition of such expressions (concerning human ones) by Tinwell et al. [2011].

Although not explicitly designed as being a serious game, the technological approach of Sidner et al. [2018] using and comparing an always-on always-reactive IVA (cartoon-style) and SR (Reeti robot) to fight loneliness and provide a happy experience for older adults at home is worth mentioning here (See Chapter 23 on “Socially Interactive Agents for Supporting Aging” [Ghafurian et al. 2022] of this volume of this handbook for further details). For some activities the system provides, it features gamification aspects to keep interaction light and fun. Users can interact by speech and by typical GUI elements with the system, which detects and analyzes social signals (face, gestures, posture) in real time. The technology is supposed to be present in people’s everyday life with the goal of being a companion to them. Which form of appearance (SR, IVA) is preferred by seniors and which kind is appropriate for specific tasks, such as entertainment, physical care with exercises, scheduling, casual small talk, and the news, was investigated in a rare one month-long pre-study. The study revealed that neither the IVA nor the SR compared to a control condition is significantly better in decreasing the degree of loneliness or increasing overall happiness. All participants reported a basic positive though not strongly positive attitude toward both agent forms. However, the SR was perceived as more trustworthy. This circumstance is explained with the less human appearance of the Reeti robot, which may have contributed to users’ sense of trust because of the uncanny valley effect in creating robots that seem more human-like.

28.4 Current Challenges and Future Directions
This section focuses on particular challenges that come with the employment of SIAs in serious games in the application domains of social training, social companion, and therapeutical support. This domain requires a more integrative and interdisciplinary approach than ever. Within that scope, two areas of challenges are discussed: (1) empathic understanding and adaption to users and (2) evaluating such serious games.
28.4 Current Challenges and Future Directions

28.4.1 Empathic Understanding

Researchers in the area of empathic agents are motivated by several reasons (see Chapters 10 on “Emotion” [Broekens 2021] and 11 on “Empathy and Prosociality in Social Agents” [Paiva et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021], for further details). A general motivation is that agents are more likely to be accepted if they are aware of the user as a social actor [Picard 1997, p. 247]. This includes the motivation that agents should act in a social (familiar) way. Empathic behavior is a part of this [e.g., Bickmore 2003, Tapus and Mataric 2007]. Connected are the research questions of is the system showing social (empathic) behavior and has the believability of such systems increased [e.g., Lester et al. 1997, van Mulken et al. 1998, Paiva et al. 2004, Dautenhahn 2007]. In order to do so, such agents must come with the ability to understand others at the level of intentions, motivations, and feelings, which includes perceiving and understanding others’ affective states and acting accordingly [e.g., Bickmore 2003, Conati and Maclaren 2009, Wilks 2010, p. 4, Marsella and Gratch 2014, Paiva et al. 2017]. These requirements are connected with the research area that remains to be investigated, namely how perspective-taking might influence collaborative and argumentative learning for identity development and attitude change [Tsovaltzi et al. 2014].

The next generation of SIAs, which are empathic cultural-aware agents, consider social values and norms that come with requirements and challenges that put those of current SIAs to the test in the areas of: (1) explainability, (2) observation, (3) theory of mind of users, and (4) adaptation. SIAs must be able to explain themselves on the behavioral and motivational levels. This requirement is mandatory since empathy is a collaborative process that requires both partners to disclose (private) information in order to establish a necessary level of trust. Trust is a concept related to feeling secure [Moser and von Zeppelin 1996], which is a base necessity to exploration and play. Hence, trust is mandatory for serious games with SIAs addressing social aspects.

SIAs observe the human-dialog partner on the level of social signals, including voice. Other technological sensors can be added (e.g., pulse or heartbeat). They detect essential patterns and sequences of social signals in interaction (e.g., smile, facial expression, gaze behavior, gestures, posture, or physiological values).

Using current and past multimodal information, approaches to SIAs try to understand the meaning of utterances and actions of a particular user. For the interpretation of behavior, different knowledge is needed that covers behavioral norms and values for the culture (e.g., Western, Eastern), possible group affiliation(s) (e.g., scouts, researchers, workers), and individual characteristics (e.g., personality). The social hierarchy (e.g., status), situational (e.g., home or work
environment), and relational context (e.g., family member, work colleague, or stranger) must be considered too. Moreover, knowledge about internal (subconscious) processes related to mental states and related observable behavior is mandatory. Mainly, a model of emotions that differentiates between external (observable) emotions/social signals and internal emotions is required. The first steps in this direction have been made with, for example, the MARSSI simulation of internal emotional states and emotion regulation [Gebhard et al. 2018a] and a novel architecture for the emotional interpretation of social signals [Aylett et al. 2019].

For the future, a more in-depth focus should be on the processes of intrapersonal emotion regulation, coping, and display rules. These processes are influenced by culture, group affiliation, and even family or individual values and norms. All this information must be considered for the simulation of possible mental states of the dialog partner, based on possible representations of goals, motivations, and wishes that can be put in relation to the interpreted behavior and internal processes and emotions.

Great potential lies in the employment of SIAs that (inter)act empathically by respecting cultural, behavioral values and norms and also by respecting the behavioral values and norms of groups and individuals. They show social-communicative abilities such as interpersonal emotion regulation, social mimicry, display rules, and emotional contagion. Therefore, the cultural-aware agent requires a representation of the dialog partner’s motivation and goals (relevant mental states) that use a model of cultural values and norms but also model values and norms of groups and individuals. As argued before, all these models are related to the concept of trust, which is mandatory for next-generation social serious games with SIAs.

The possibility of SIAs to emulate trust and human-like relationships would allow the creation of even more natural and immersive social serious games. Such SIAs would be in the role of a trustworthy companion as a learning partner in long-term social serious games, for example, in the application areas of social conflict training or long-term therapeutic assistance. Such agents must have the ability to develop and repair trust that is most relevant for the concept of relationship [Lewicki and Bunker 1995]. A major requirement for this is that SIAs are able to adapt to individuals on various levels, especially considering the agent’s role and status. This approach should reflect dynamic individual user aspects (e.g., physiological, such as the level of hearing or linguistic characteristics, such as the dialect or idiolect, and cognitive, such as the level of cognitive resources). Therefore, they should be able to learn individual characteristics, values, and norms and their relations (1) to internal representations such as motivation, goals, and wishes and (2) to behavioral aspects. The ability to adapt contains the ability to react and
address misinterpretations/simulations, e.g., making apologies. This approach has to include processes for reflecting and discussing interpretations and learning new values and norms. Social serious games with SIA at this level would allow a secure place for exploring serious issues at the most individual immersive level of joyful experience.

Concerning societies and common knowledge and understanding, the SIA research might be a driving force too. In implemented community ideas, (serious) games aim to enhance motivation at the group level but also to leverage knowledge co-construction. Hence, the realization of social aspects of learning goes beyond sharing results to get a trophy and merely competing against each other, like in typical “non-serious” games. It rather builds on basic social comparison processes and relates them to games’ cognitive or task structure. It thus embraces ideas of socio-cognitive conflict for knowledge co-construction [Mugny and Doise 1978]. Developers, therefore provide affordances such as chats with social media (e.g., wikis) and networking structures and design shared spaces for common projects, and allow the sharing of results not to compete but to build on each other’s contributions [Fields et al. 2013, Garrelts 2014, Du et al. 2016].

28.4.2 Evaluation
Across people and situations, serious games without SIA seem to have positive outcomes and even dominate traditional learning methods for cognitive gain outcomes [Vogel et al. 2006]. Serious games employing SIA can be categorized into two groups: the ones where SIA are in the role of a teacher (cf. pedagogical agents) and those where SIA represent interactants and enable difficult social situations to be experienced virtually (social training systems). These diversified applications require different evaluations, adapted to their individual purpose.

SIA in the role of pedagogical agents, for example, Steve [Johnson and Rickel 1997], Herman the Bug [Stone and Lester 1996], and SmartEGG [Mitrovic and Suraweera 2000], support the learning of complex problems. The evaluations of these systems can include the rating of the agent concerning its physical appearance and behavior and the learner’s assessment of the likability, helpfulness, or entertaining character of the whole system. Moreover, motivation, learning efficiency, and effectiveness can be either self or externally rated (objective or subjective). It seems that pedagogical agents can yield important educational benefits in the form of improved problem-solving, particularly for complex problems [Lester et al. 1997], and increase student motivation and perception of their learning [Mitrovic and Suraweera 2000]. However, there is no meta-analysis examining serious games with SIA in the role of pedagogical agents to support learning and no studies that examine the long-term effects of pedagogical agents (See
Serious Games with SIAs

Chapter 21 on “Pedagogical Agents” [Lane and Schroeder 2022] of this volume of this handbook for further details).

SIAs exploited in the application area of social training systems mostly do not overtake the role of a virtual teacher explaining a specific problem. In those serious games, the learner is often confronted with a difficult social situation, for example, a job interview [Anderson et al. 2013, Gebhard et al. 2018a], bullying [Aylett et al. 2005], challenging pupils [Lugrin et al. 2016], or aggressive behavior [Bosse et al. 2018]. The SIAs are in the role of interactants with which the user trains the specific situation. The social training system can be evaluated by the user applying standard questionnaires measuring, for example, the believability of the agents’ behavior [Niewiadomski et al. 2010] or the scenario or the user’s experienced co-presence [Bailenson et al. 2005].

The evaluation of the learning effect of social training systems is more challenging for several reasons. The best evaluation of the learning effect would be to observe the situation that is trained in the serious game in the real environment before and after the training including an experimental as well as a control group [Field and Hole 2002]. Those in the wild measurements of training effects are often infeasible due to, for example, the rareness of situations, privacy of other people involved in the situation, or ethical issues. An alternative could be role plays because of the possibility of capturing real and natural behavior [Freedman 1969]. As nonverbal behavior is crucial in social situations [e.g., Mehrabian 1972], role plays are particularly suitable [Sader 2013].

In both designs, in the wild as well as the laboratory setting, evaluation criteria have to be defined to assess the behavior of the user. However, in difficult social situations, the demands on those involved vary and it might be difficult to define what constitutes more or less desirable behavior due to different values, personalities, or societal rules. Therefore, the definition of externally assessed evaluation criteria for answering the question “if the learning situation was handled well by the user” is complex. Due to the different contents of social training systems with SIAs, it is difficult to develop a gold standard for their evaluation. Hence, the evaluation of such systems should depend on the respective context and motivational goals.

Apart from an external assessment, another source of data is the assessment of the user. It is tempting to ask users after completing the social training to imagine the situation in a real environment and to state how they would behave. As correlations between statements about predicted behavior in questionnaires and the real behavior are low, those measures may not be meaningful at all. Using self-assessment questionnaires might be a better choice in asking users for their performance. This approach, however, comes with various difficulties. Also, there
are several aspects that can hamper the validity of the results. General disadvantages for the self-assessment are (1) socially desirable responding [van de Mortel 2008]—tendency for people to present a favorable image of themselves on questionnaires self-affirmation—and (2) individuals are driven to protect their self-integrity and therefore might give embellished answers regarding the learning effect [Steele 1988]. However, as an additional information the self-assessed performance is a good indicator for the success of a serious game with SIAs, especially in the context of social training systems.

One aspect that is still understudied are the long-term effects of serious games with SIAs, especially in application areas that focus on social values, such as social training systems, and therapeutical assistance.

**28.5 Summary and Conclusion**

This chapter addresses the use of SIAs for serious games in various application areas. Research in this field comes with the challenge of how and in which roles such agents are used.

Serious games with SIAs are powerful tools for interactive learning. SIAs that behave plausibly and look realistic are needed to bring social training in serious games closer to a human natural interaction experience. Careful design of the environment and interactivity must convey societal and individual values to let users have a consistent and believable experience.

With over 20 years of research, many serious games with SIAs for a broad spectrum of application areas have been developed and evaluated. The chapter clusters the games in three interlinked groups of different approaches focusing on content and story, immersiveness and interaction, and individual experience. They all employ well-researched paradigms and learning concepts, such as collaboration, socialization, embodiment, immersion, narration, and interactive storytelling.

The different approaches’ richness reflects that there is no general method or recipe for creating a serious game with SIAs for a specific learning goal. Probably the most important reason behind this can be found in the individual differences in each of us. For every situation, we have different needs and requirements that influence how we manage situations.

Recently, serious games with SIAs that adapt empathically to individual users have been created and investigated. With the development of more precise technology-based detection methods of (non-)verbal cues and verbal understanding of the interaction, it has become more evident that the understanding of individual values remains difficult. A deep understanding of users, their actions, motivations, and wishes might not be possible since they might even not be obvious (consciously) for the users themselves. All this is connected with how we are raised,
which norms and standards (cf. cultural, cohort, and family) we got in touch with, and what personal experiences we made in our life journey. It seems necessary that approaches that allow an empathic adaptation to individuals need to explain and reflect on conveyed values and provide some ways to learn and adapt to new individual values.

The more an individual adaptation is required for processes of learning, training, and understanding, the more cultural and personal knowledge is required for the involved technological methods and techniques. This extension will lead to social serious games that employ SIAs to build and repair trust and long-term relationships. Both abilities are necessary for a joyful and secure long-term exploration of serious issues in various application domains. These challenges come along with the endeavor on how to evaluate and measure specific effects. In a sense, such kinds of serious games can be seen as a way to explore and reflect on the values and norms of societies, cohorts, and smaller social groups.

References


References


A.1 Introduction

This chapter contains a collection of interviews on current challenges and future directions that researchers are faced with when working with Socially Interactive Agents (SIAs), see Chapter 1 on “Introduction to Socially Interactive Agents” [Lugrin 2021] of volume 1 of this handbook [Lugrin et al. 2021] for a definition. The works reported in all the chapters of Volume 1 and Volume 2 of this book have highlighted the importance and necessity to take an interdisciplinary approach when conducting research on and developing SIAs. It requires dealing with many facets of multimodal behaviors that occur during an interaction between humans and other agents that can take place in a great variety of social domains. When working

*Birgit Lugrin and Catherine Pelachaud have a shared first-authorship on this chapter.
on finding key challenges that still need to be addressed, different clusters of questions were built and several experts in their respective fields were invited for the interviews. They thus discussed various aspects of the research with SIAs from different perspectives and laid ground for lots of future research directions, introduced thought-provoking ideas, and discussed the potential risks of this research area.

The interviews were conducted by two of the editors of this handbook:

Birgit Lugrin: Professor for Media Informatics at the Julius-Maximilians-University of Würzburg, Germany, and

Catherine Pelachaud: Director of Research at CNRS in the laboratory ISIR, Sorbonne University, France,

while the interviewees were authors of various chapters of both volumes of this handbook.

We prepared the interviews ahead of time by following a bottom-up methodology. When we outlined each of the chapters in the very beginning of this handbook, we asked every author to include a section on current challenges and future directions within their specific research domain of SIAs. We first ran through the chapters and gathered the challenges they addressed. Not surprisingly, there were several overlaps that faced similar issues or risks that were of importance for various implementations, but with a different focus. They covered very broad issues, addressing the need of novel computational approaches, evaluation protocols, but also societal and ethical issues. Then, as a next step, we defined a set of main topic areas and clustered them into four main topics:

- Social Interaction
- Computational Architecture
- Evaluation
- Ethics

For each of these topics, we defined a set of open questions to be addressed during the interviews. The questions addressed SIAs in both their potential embodiments, as Intelligent Virtual Agents (IVAs) or Social Robots (SRs). We planned for four interviews, one on each of the defined topic areas, and organized interviews with two or three experts, who were all authors of different chapters of this handbook. A fifth interview dedicated on ethics in the application of SIA for children with Autistic Spectrum Disorders (ASD) was also organized with a specialist in this area and who has also written a chapter.
We told all interviewees that the interviews will be conducted in a semi-structured manner, via video conferencing. They were informed that the interviews will be recorded and transcribed. The transcription was done semi-automatically, relying on automatic transcription tools, but manually going through the whole interviews and making corrections afterwards. Then, the draft of the transcription of each interview was sent to all the interviewees of the respective topic area for potential corrections and final approval.

This chapter is organized as follows: each identified topic area (and the subtopic on ethics in ASD research) is a section of this chapter, introducing the interviewees as well as the list of questions that were addressed, and then reporting the interviews (Sections A.2–A.6). At the end of the chapter, some concluding remarks are given (Section A.7).

### A.2 Interview 1: Social Interaction

For our first interview, taking place in October 2021, we have drafted the following five questions:

**Question 1:** How shall we integrate social functions to facilitate adaptation, rapport, or engagement into the interaction with SIAs?

**Question 2:** How will we consider individualization of the SIA to match different personalities, genders, or cultures?

**Question 3:** How much formality and natural language, for example, politeness, do we need? Should we have to say, for example, “Alexa play Netflix please” or simply give a command?

**Question 4:** And from the agent’s point of view? How much formality is needed here?

**Question 5:** Are robots the new IVAs? How do you foresee the potential of augmented reality?

### A.2.1 Participants

Cynthia Breazeal, Professor and Associate Director at the MIT Media Lab and the founding Director of the Personal Robots Group.

Jonathan (Jon) Gratch, Research Full Professor of Computer Science and Psychology at the University of Southern California and Director for Virtual Human Research at USC’s Institute for Creative Technologies.

Ana Paiva, Professor of Computer Science at INESC-ID, Instituto Superior Técnico, University of Lisbon.
A.2.2 Question 1

Catherine Pelachaud: And the first question we had was, how should we integrate social functions to facilitate adaptation, rapport, or engagement into the interaction with socially interactive agents?

Ana Paiva: I would start with one thing related with the development of social agents. Nowadays, we have a lot of discussions about whether to have a theoretical-driven approach to develop our agents’ social behaviors or a more “data-centered approach.” I think we need to consider that the two can go hand in hand, especially now with all the machine learning techniques that supports these very data-centered approaches. But those data-centered approaches are not enough, and I believe that without theories, like theories of emotion, theories about report, that support building models at a more abstract level, we may be re-inventing the wheel. So I believe that the one of the big challenges we have now is to combine these two approaches for development and embrace both the data driven in certain aspects, as well as the theoretically driven approaches, as more symbolic AI people used to do. But I am not sure whether Cynthia and Jon agree with me...

Jon Gratch: Historically, I have been a strong proponent of theoretical approaches. I’m somewhat changing my thinking on that in a sense. And I guess one of the issues, at least within affective computing, is in some ways the prominent theories have led the field astray in the sense that there’s been a predominant focus in affective computing on trying to recognize people’s feelings. And that comes from work by Paul Ekman who had a very strong sway over the field. Yet, it’s not clear that that approach is actually supported by the data. Ekman’s influence is much reduced in psychology, yet if you look at what companies are actually doing in the field they’re tied to his theory. And that can sometimes lead things astray. Sometimes the theory shapes the questions we ask, and sometimes those aren’t the right questions. So, I’m struggling with how to do that and I definitely believe we know that there’s a lot of problems with data driven, and you get these black boxes and they learn something stupid when you finally figure out what they’ve learned. So we need to tie the learned models back to constructs that we know are actually tied to the phenomena we’re trying to study. But somehow, some balance between the two is important and I guess that’s what you’re probably saying.

Ana Paiva: Yes.

Cynthia Breazeal: Yeah, and I agree. Both are important. Of course, it begs the question of what data is informing what theories. I suspect we probably feel that we don't have the comprehensive kind of datasets that we would like to see across geography, different groups, ages, cultures, etcetera. There is a broad diversity of people who we would like to have these systems to interact with, in richer and more competent ways. So, for my own work, we’re looking more and more at longer-term deployments, integrated into real human contexts, where people are going about their daily lives. We are trying to capture richer, more representative data that I think is going to challenge and force us to improve our theories. I see it as a virtuous cycle. A critical piece, I think, is challenging ourselves to get richer, more representative datasets. Of course, this gets us into the ethics discussion on how we design these systems responsibly? How to deal with these questions in light of design justice, and so forth. These are really important considerations for the field as we integrate our systems into society.

Jon Gratch: I just want to point out that I really appreciate some of the work Cynthia has been doing in terms of long-term interaction. Because when you talk about phenomena like rapport and social function, there is a fair amount of work in our community, but typically those are “one-shot” studies where participants never develop any kind of long-term relationship with the technology. So they don’t actually get to learn if there is really a function to those expressions. Designers make something look like it has a function and people treat it as so in a short-term experimental study, but then the designers don’t actually get around to actually implementing that function. In that companies are actually deploying this technology in the real world for persistent interactions, I think that makes that question of what are the functions of these behaviors much more important because it can look cute or look like it has a sophisticated cognition, but if it doesn’t follow through and actually do those things, then people just tune that kind of stuff out.

Ana Paiva: Yeah, I truly believe that it changes the way we look at the interaction and where the interaction is and so forth. When you look at one-shot studies is different from when you look at long-term interactions. In fact, we saw very clearly in an old study with the iCat that when we had just one short

interaction the novelty factor is there.³ People look closely at the robots and our agents as they are trying to figure out all the social signals and enjoying the novelty of the situation. But then when the interaction follows on the next day, or next week, and becomes repetitive, a lot of the signals become irrelevant, and the task becomes more important. The salience of the social signals decreases. So I remember with the iCat that after several interactions, the kids didn’t even look at the iCat anymore. They were just looking at the chessboard because that was the important they needed to look at, and the iCat, apart from some very specific moments, didn’t matter anymore. So there is a clear change, especially for the social signals, when they interact over long periods of time. Also, because they realized that the robots, or the agents, could not see them or cannot follow the same signals as we do, and don’t respond the same way. So, in fact, I remember at some point they would ignore what was going on because they knew that the system wasn’t able to respond in a natural way as we humans respond. So, yes, I totally agree.

Cynthia Breazeal: We are also finding that personalization is increasingly important as we delve into long-term interactions. People are changing and the system as they interact with it, and ideally our systems will be able to adapt and change to continue to be engaging and helpful to people. So we have been looking at, for instance, agents in a family context where of course you have different kinds of people—different ages, different ways they interact—so personalization is becoming a more important theme for our work, as being informed from this long-term interaction context.

A.2.3 Question 2

Birgit Lugrin: I agree, it is really important. And this really goes into our second question that is on personalization and individualization. So if you would like to elaborate a bit more on how we should consider individualization and how we should match different personalities, genders, cultures, and so on.

Cynthia Breazeal: I say YES to all of that because people are not the same. Of course, personalization also brings ethical questions around privacy, transparency, and accountability. It is a rich, multifaceted challenge beyond what algorithm we are going to apply to achieve longer-term personalization, but we need to consider the broader context of what that means. Again, for

---

people who are potentially living with these systems, I think it is also a really important consideration. I think cultural differences are important to capture, and we’ve already talked about different age groups, different applications. How the agent presents itself, and its role is going to vary depending on what the main kind of task or value proposition the agent is offering. In our work, we’re seeing things like differences in personality traits may influence the effectiveness of a given intervention for a given person. We found this in our emotional wellness work. We may discover this is another important aspect of personalization, which may be based on personality profiling—in addition to things like interaction style, and role, and all of these other things. It is a very rich research topic.

Ana Paiva: Yeah, I see that personalization is something that needs to be carefully considered because it can actually make the interaction worse. I’ve seen some situations where the adaptation changes make the interaction not so fluid. So, if the agent is changing as it interacts with one person, if that personalization is not well done, then we can have negative effects. And in fact, I wonder: How far do we want to go with personalization? Because given what’s going on with social media, I really think that exaggerating the personalization and all the information that’s captured about the individual user may have some ethical problems. We may need to think about personalization in different ways. I did my Ph.D. many years ago on personalization and user modeling, and now, many years later, I’m questioning some of the fundamentals and what drives such personalization. Because I don’t want to be sold something that the system already knows that I like. For example, I searched for trainers, bought some trainers, and now I keep getting trainers sold to me. I don’t want that kind of personalization. I want the systems to know about the relevant things about users that make the interaction more engaging. Or maybe engagement is perhaps not the right target but rather that users can learn more. But knowing about my training shoes? I think we need to be very careful about personalization. And it’s something that we as a community need to think about it. For example, with chatbots, do we want our chatbots to really get all kinds of information about us?

Jon Gratch: I think we need a better language to characterize different aspects. For a tutoring system, I think it’s somewhat uncontroversial you want the tutor to personalize the feedback based on the particular errors the student has. You might think that for speech recognition systems it’s uncontroversial to train and tune the recognition on that particular person. Although, you can even take that to the extreme: so when the person starts diverging from normal
English language, or starts doing racist things instead, do you want the agent to adapt and reinforce and support those kind of behaviors? I don't think so. But at a broad level, personalization is also an effective influence tactic. So companies, for example, want to allow people to customize and personalize, say their Alexa, or some other customer service app, so they feel it's their friend. But it's not their friend, right? It's the voice of the company. And it's there for a very specific purpose: to sell more stuff. More broadly, I'm somewhat conflicted about the idea, and Ana knows this because we were in a workshop together in Dagstuhl about whether we should make these things. When you're talking about personalization, you're using very anthropomorphic terms. So, should we anthropomorphize these things and make them seem like humans and seem like they care about us and have a relationship with us? Should human–human interaction be the gold standard, or should these somehow be different and maybe take advantage of the uniqueness of the technology to create other metaphors? There's a discussion at ACII, like is there such a thing as gender neutral speech? Because it seems to be that people want to personalize, or designers want to personalize their assistance to be women, right? Because that fits people's preconceptions, and then tends to reinforce those cultural stereotypes. So I don't, and I don't have an answer there, but we need a better language to talk about these things.

Catherine Pelachaud: Yes, I'd like to add something. So when we talk about personalization, I don't think it means solely that we could, for example, choose the colors of the agent. It could be also understanding when to personalize and which factors to personalize. Is it possible? It means people would have to personalize the agent from the beginning to the end. I think it is this aspect that we should look at.

Jon Gratch: Even if you're the consumer and you're creating this agent for yourself, in some ways, do you really want to create an agent that sort of fulfills your wishes in every way, or do you want something that confronts you and challenges you instead? There's a sense in which we can easily make this technology narcissistic, right? It just reinforces a person's current foibles.


That’s why I think you’re concerned, like Facebook is in a sense trying to do something like that.

Cynthia Breazeal: There is the question of how we approach the challenge of personalization and all of these different facets. And then there’s also making sure that it supports what’s important to people. Transparency, accountability, and explainability are going to be really important in addition to privacy and security. Privacy is often first and foremost in people’s minds when they hear something like personalization. But people need understand what it actually means for their system to personalize to them, and people should be able to further shape or change that. What if a system adapts to the person in a way that isn’t right? People should be able to adjust it.

Ana Paiva: And not only that, I think the user’s autonomy to decide whether they want to really get their data in the system, and the system to be adapting to themselves or not, is something important. The user needs to be in charge of that decision. And that’s to do with transparency, but also guaranteeing the user’s agency. That is something that I don’t like, when my social media (Facebook or others) tell me to buy something or recommend something that is assumed I would like—my agency is at stake. Plus, when I click “why am I seeing this advert,” it reports to me some general justification like I’m a woman, I speak English, I live in Portugal, or I’m between 25 and 60 years old. And I know that this is not the reason why I’m seeing a certain advert. So, personalization must be linked with transparency in a way that guarantees user’s agency.

A.2.4 Question 3

Catherine Pelachaud: We were thinking, and there’s some work especially in social robotics, on how much formality and type of natural language do we need? Would we say “Alexa, could you play Netflix, please,” or simply give a command “Netflix, open”?

Jon Gratch: You could imagine Alexa, if you are not polite, would have a model of politeness and complain. Dialog and social norms get constructed through the interaction. You may have seen, when Manuela Veloso gets on a soapbox about how technologists shouldn’t encourage people to talk to robots like people. They are not people, and so why should we reinforce that with language? And I guess it just touches on the idea that do we leverage and reinforce existing stereotypes or not? There is some research around Alexa,

specifically in children, where children learn certain interaction styles with Alexa and they learn that maybe it’s fine to be bossy or rude. And then the question is, does that transfer, because it’s human-like, to other humans? And I think it’s unclear. It’s the same problem with games, right? When people kill each other in games, does that make them more likely to kill each other in the real world? It is not clear. But I think it’s hard to imagine language, without making it have to know something about how humans use language. And so I think, at some level, we’re sort of stuck with the fact that people have emotions, and they use those emotions to communicate things. They have social goals, and those things probably have to be implemented and understood at some level by these machines, and then those machines need to reinforce some set of social interaction norms. What those choose to be, I am not sure. But I think eventually Alexa, it will try to enforce politeness norms.

Cynthia Breazeal: There’s the natural language understanding, and then there is how that builds on this notion of what’s the relationship? Which gets into roles, appropriateness, and so forth. For example, when we designed Jibo, we gave a lot of thought into the way the robot speaks about itself. Jibo always reinforces “I’m a robot.” And if you asked a particular question, like on religion, Jibo admits “I know nothing about that, you should ask another person.” So through the agent itself, and how it contributes to the conversation, Jibo continually reinforces “this is what I am, this is what I can talk about, there’s things I can’t talk about, that you really just need to talk to other people about.” Reminding and reinforcing Jibo’s differences to being human was a part of our design philosophy. We are discovering as we go into longer-term interactions, across age bands and application contexts, the one thing that people do want more from these agents—at least in the contexts we’ve explored like health, wellness, and education—people want more capable multiturn conversations and dialogue with the agent. They want the agent, at least within appropriate bounded ways, to remember their conversations and past contexts. People want to avoid this constant repeating of what was said yesterday, or the day before, like the agent has short-term memory loss. At that point the agent just seems stupid, right? So, there’re so many rich facets about how you do this well, in a way that respects people and their values, and what people want out of these systems. Drawing appropriate bounding boxes around these agents is really important. I think our field has been so caught up in the technical challenges, because of course

there're so many technical challenges. But as these technologies are getting out there, and people are starting to interact with them, and expectations for what they can do and can't do are being established—these broader ethical design frameworks are becoming more and more important for us. We need to really develop new ethical design methods and co-design methods. It's hard to say it's an all or nothing thing on any of these dimensions. It's about what's the right and responsibly bounded scope. And that just depends on a lot of things.

Ana Paiva: Yeah, I totally agree. Once you build your agent that is able to interact through natural language, it raises expectations from the point of view of the user, who expects the agent to learn and understand what is being said. So, natural language raises the expectations, and if the system does not meet those expectations, there's a problem as trust may decrease significantly. So, from a technical point of view it is hard, but on the other hand, if you don't do it the interaction becomes limited because people really want that type of interaction. So it's a balance that you have to juggle. Actually, in the Dagstuhl event in that we were a couple of weeks ago, Catherine was there too, Roger Moore was arguing that robots should speak with a robot voice. Robots, agents, or chatbots must disclose what they are, and talk with voices that are clearly mechanical, clearly not human so as not to raise too many expectations about the agent. So I think disclosing what it is, saying “I'm a robot,” “I am a chatbot,” and talking with the voice of a robot helps because then the natural language aspect becomes less demanding, and people don't feel cheated. I think managing expectations is important in natural language interaction between humans and robots.

Jon Gratch: But some other research shows that portraying yourself as a person is more effective for, for example, a mental health application. There's a number of studies where they actually manipulated robot backstory versus human backstory.

Ana Paiva: Are you saying that when the system pretends to be a human, it's more effective?

Jon Gratch: Yeah, so there are studies that show it is more persuasive and also elicits more honest disclosure. Somehow people find it more relatable. Even

---

8. Conversational Agent as Trustworthy Autonomous System (Trust-CA); http://www.dagstuhl.de/21381.
though in some sense of course they know it’s a machine. Tim Bickmore\textsuperscript{10} has done some of this work, and we have done some. So there’s a tension between what we might perceive as unethical and what is effective. Hopefully there’s not that tension, but sometimes there is.

A.2.5 Question 4

Birgit Lugrin: Thanks. That’s already going in the next question, and I would really like to know what you all think about it, because that’s the other way around: how much formality and politeness and things alike are needed from the agent’s side? Do you see a difference between what the user should be using for the interaction and the agent should be using?

Jon Gratch: Well, I think we are matching. If the machine doesn’t adapt, people adapt to the machine. I think there’s a natural tendency for synchrony, adaptation, and entrainment. I find I used a lot of these manners in terms of language with our virtual humans. People can’t help but train yourself to the technology. So I think that choice will influence how people talk, at least with this genre of entities. It’s a question of do you feel like it’s important to reinforce certain social politeness and norms. I think people will change their behavior based on those two different choices, but as to what is the right choice, I am less clear about that.

Cynthia Breazeal: From my personal perspective, given what Jon was saying, we do see people mirroring the behavior of robots, and the attitudes conveyed by them. It is because we’re people, and we do these things when we interact with others. I think it’s a skill, like any other skill. You reinforce the behavior you practice. So, if we want a society of people who are empathetic, compassionate, and polite, it behooves us to design systems that both convey and reinforce these behaviors, as well as to let people practice these behaviors. Because chances are, the behaviors that are reinforced over time become the default skills that people are going to apply to interactions across people, animals, and whatnot. So, let’s look at it from a very human-centered perspective: what do we want to encourage? I think it’s important to consider the influences of these technologies that are shaping us, both positive and negative. We have very transactional devices now, and we’ve seen examples of children being bossy or transactional with these devices.

\textsuperscript{10} Timothy Bickmore, Northeastern University, \url{https://scholar.google.de/citations?hl=de&user=x9kzObUAAAAJ}, author of Chapter 24 on “Health-Related Applications of Socially Interactive Agents” [Bickmore 2022] of this volume of this handbook.
Jon Gratch: It is less clear whether that generalizes to interactions with kids. We don’t know.

Cynthia Breazeal: We don’t know. Parenting plays a role, too. I’m just conjecturing that the more you behave in a certain way, the more that those behaviors are reinforced, the more it becomes default. Whether you intended to come across a certain way or not, I just see these as skills, practiced skills in general. I think these systems can play a role in helping us practice ways that we want to be as people, versus having us behave in ways that don’t serve those goals. For me, that’s more of a philosophical stance.

Ana Paiva: Yeah, I totally agree!

Cynthia Breazeal: We have data that shows that people do emulate the behaviors of these robots. I can imagine if people do that more and more, that will become more of a learned behavior. So, let us just be mindful of that when designing these systems.

Ana Paiva: I agree. My own chapter is about prosocial agents, promoting prosocial, and yeah, I totally agree.

Jon Gratch: My presumption would be, the more that thing is like a human, the more likely it is to get generalization. The more the thing clearly indicates something different, it ought not to generalize. But the data from games at least is very unclear as to whether violence in games generalize this to violence in the school.

Ana Paiva: There are other aspects, such as promoting altruistic behavior. And you actually can do that with games. There are systems that have shown that by interacting with a game they can help a lot to raise some more cooperative behavior. Of course, we don’t know if it lasts over a long period of time. We’ve done one game for teenagers to help address the bystander problem in bullying. And what we found was that by having this game and having the agents in the game, promoting these more altruistic and prosocial behaviors in bullying situations. So I think that agents can be a force for good, and that needs to be further explored. I think there are avenues to explore and we as a community should be building these systems to promote a kinder society.

Cynthia Breazeal: To acknowledge this, it is something that we need to understand more—the social influence of agents on us and these prosocial opportunities. It’s an area we need to understand more.

---

Jon Gratch: Yes, and it's interesting that there is a lot of work, and a lot of work here at ICT,12 where we are, and Ana does this as well, building agents to try to teach interpersonal skills that we believe will generalize to the outside world. But then when we think about building a video game you say, “that won't generalize the outside world.” There is a difference in the sense that people know they’re learning this skill, and they’re encouraged to try to have the mindset to apply it in the outside world. Whereas in a game they are just having fun. But it does seem that this is a very blurry boundary, and if it's going to work in one, it's got to work in the other, right?

A.2.6 Question 5

Catherine Pelachaud: I think we are ready for our last question, we had foreseen. Which is, are robots the new IVAs, and how do you foresee the potential of augmented reality?

Cynthia Breazeal: I guess it depends how you define an IVA. The word “virtual” would make me think robots are physical, or physically embodied. So I wouldn’t characterize any physical robot as a virtual agent. If we use the term “conversational agent,” or something that's more agnostic to the embodiment, then robots can be considered emerging conversational agents that fit in with the smart speakers, and the animated agents, and everything that we already have. I think in the field of human–robot interaction, there is a growing interest in the expansion of more advanced multimodal conversational abilities, in general. Virtual reality, augmented reality, other kinds of blended reality, and all different flavors of the amount of physical versus virtual permutations—we were interested in looking at things such as migratable AI. Right now, an AI persona might always be in a robot, or always in a particular device. But when you think about moving between contexts, from your home to your car, to your workplace, to wherever: what if you have an assistive agent that can migrate across these embodiments to always be with you? I think it is an interesting question. So, that rather than say it's a mixture, maybe it's also the transition between them. Migratable AI or migratable agents has been a research question for a number of years. And AI with all these conversational abilities is just bringing that to another stage in their development. But I think any of these embodiments have their distinct set of advantages. A physical robot has the advantages that a physical body affords. Similarly, a virtual agent has the advantages that a virtual body

12. Institute for Creative Technologies—https://ict.usc.edu/.
affords. So it just really depends on your application in terms of what mix makes the most sense. It will just vary depending on the design context.

All: Yeah.

Ana Paiva: I think both virtual agents and social robots are embodied agents, that is, they have bodies. The robots are physically embodied in your environment, in our physical world, and are able to act there. The other ones are in the virtual world. Of course, like Cynthia said, there may be some blended realities. I’ve seen situations where you have a head-mounted display and then your robot gets into the virtual world, so you can have all these combinations. But at the end of the day, embodied agents have properties that afford interaction through more senses because you can see them, you can touch them (in the case of the robot), and they act in our world. So there’s the body effect that I think is important in these embodied agents. And you can draw many conclusions from the virtual agents that can go into social robots, and the other way around. So the two communities should learn from each other, for sure, because there’re many things that have been studied in the IVA community that the HRI community should have heard, and the other way around. So I think they are both conversational and “embodied” social agents, and that’s it, and that’s our field. In fact, I started working in social robots when I was doing virtual agents and went to one of our agent conferences, and I saw the iCat, and realized that I could use the same models that I had in my virtual agents in the iCat. So then, almost automatically, you go from one to the other. And I think that’s the beauty of our field, the embodiment can be physical or can be virtual, and extend or combine these two.

Jon Gratch: My perspective is there’s a thread within both robotic and virtual human community, that we want to build things that are like people. And that research is very similar between HRI and IVA, if not completely overlapping probably. And it doesn’t take advantage of the specificities of the modalities, and the only shame is that there’s not more crosstalk. But as Cynthia was saying, there is uniqueness to the modalities, and I think in robotic systems we are actually forced to deal with the uniqueness of embodiment. It’s about extreme power requirements, it’s in a world that has to use actual sensors that we can cheat with the virtual human community. But the virtual human community seems more reluctant to explore what is unique or special or possible within the virtual and the augmented. And I think that’s because they’re too wedded to thinking of these things like real people. But you can transform the nature of the interaction, in augmented reality you can
you see things that are not real, and as well as with virtual. I would encourage the virtual community and the augmented community to think more outside the human box and take advantage better of what unreality affords in terms of interaction to create something unique and special about the community. Otherwise, we're just to go to HRI.

Cynthia Breazeal: I actually think we, in the robotics community, we want to build human synergistic systems. I mean, it is a huge endeavor to build an android that actually looks and moves like a human, and most of the field is not concerned with that line of research. We’re already in the design space of: robots don’t look like people, they don’t sense like people, but we want them to be compatible and able to interact with people, to be able to support people. So, for the robotics community, we don’t view the human as the gold standard for how we want robots to appear and behave. Rather, we want to design robotic systems that are synergistic and complimentary to people. There are fascinating advantages to the differences between robots and people. How can we leverage this complementarity to enhance human–robot collaboration?

Ana Paiva: I believe that social robots should be designed in a way that the sociality of the robot has to come together with the tasks that it is going to do in the physical world. It’s not enough just to be social, because then why do you need the robot? Why do you need it in the physical world? So I believe that exploring the physicality, changing things in your physical environment, and then on top of that adding the social interaction is one of the richness of the area of social robotics. Maybe that can be one of the differences from the virtual agents, I would say. But, at the end of the day, the two areas should go hand in hand.

### A.3 Interview 2: Computational Approaches

For our second interview, taking place in November 2021, we have drafted the following five questions:

Question 1: How shall we proceed to go beyond the SAIBA approach? Are incremental architectures the future?

Question 2: What is the next level of computational approaches? Is it machine learning only? Should we still consider symbolic AI and embed semantic information?

Question 3: How do we ensure transparency and explainable behavior?

Question 4: How shall we validate the computational architectures?
Question 5: Are robots the new IVAs? How do you foresee the potential of augmented reality?

A.3.1 Participants

Elisabeth André, Full Professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University.

Joost Broekens, Associate Professor of Affective Computing and Human Robot Interaction at the Leiden Institute of Advanced Computer Science (LIACS) of Leiden University.

Stefan Kopp, Professor of Computer Science and Head of the Social Cognitive Systems Group at Bielefeld University.

A.3.2 Question 1

Catherine Pelachaud: The first question is: “how shall we proceed to go beyond the SAIBA approach? Are incremental architecture the future?”

Stefan Kopp: This question presumes that we need to go beyond the SAIBA approach and that there are any shortcomings or limitations with it, which is true. There are several ways you could go beyond it, and there are important things that should be worked on. One is to overcome its very coarse modular structure with three large modules. People have been proposing more fine-grained components, and nowadays we increasingly see machine learning approaches that try to do end-to-end learning trying to get rid of a modular structure. At the same time, we still see more and more advanced approaches for highly specific modules. The other question is: are incremental architectures actually that important. I think absolutely yes, and we have been actually thinking on this right from the beginning. The SAIBA approach already includes the idea of incremental processing along its generation pipeline. It has BML chunking, and we have been considering co-articulation effects, or feedback signals being sent back at different time scales. And I still think that incremental architectures are really important, and it is something that we should embrace as a standard in the field. Especially if we want to build socially interactive agents that are responsive, that feel fluent and smooth to interact with, that are robust and also efficient when things get challenging, for instance in cases of communication problems.

Joost Broekens: I definitely agree with incremental and also modular approaches. I do feel that one element with this block-based processing process is that it seems to be still rather difficult to make the agents react fluidly on different time scales. So, fast reactive but meaningful reactions to stimuli and then some sort of a deliberative layer, I don’t want to call it a subsumption architecture, but I get the feeling that it’s still something that is needed. Let me put it this way. When I say this, it’s biased by me working on social robots lately. I feel that especially in social robots it seems to be usually a rather monolithic approach, trying to develop for a particular use case. Then there doesn’t seem to be much reactivity layers built in. I’m not sure if the SAIBA approach is very good at incorporating those kinds of fast reactive behaviors together with more high-level behavior.

Stefan Kopp: The SAIBA pipeline that was originally spelled out is basically only a generation pipeline. So when we talk about reactive as opposed to more deliberative behavior, it’s really about a full-blown architecture where we have to combine perception, processing, and generation via different routes, with different kinds of latencies, response times, but also depth of reasoning and planning ahead. Eventually, we need incremental processing at all of these different layers of the architecture, at the respective time scales and being integrated with each other.

Elisabeth André: Essential features of incremental architectures are missing in the SAIBA architecture. For example, incremental systems operate on a more fine-grained time scale. Perceptions and responses co-occur, reducing latencies through parallel sensing, planning, and action. An essential requirement is smooth error handling. When we communicate with each other, we interrupt and correct each other. Interruptible output generation is a tricky business. An agent that plans too much in advance and cannot be stopped at any point may appear awkward. People seamlessly correct each other when needed. And unless we can simulate such behavior in an agent, the agent will appear unnatural. Another point is the continuous prediction of the conversational context of what we do. When someone starts talking, we try to predict what is going on. We might even jump in. We are constantly re-evaluating what we think about what might be next. It is worth thinking about which features of incremental architectures to integrate into the SAIBA framework.

Stefan Kopp: I fully agree. This is also a reason why I think incremental processing is crucial for socially interactive agents. Like Elisabeth was just saying, in social interaction it’s hard to predict what is going to happen next beyond
A very short timeframe. It's also quite non-deterministic and thus complex to plan ahead. Classical planning approaches always hit the ceiling here in that regard. But incremental processing actually allows us to outsource this planning and processing to the interaction. By doing some limited, good-enough reasoning and then acting it out. Whether it works or not it's something that you will then see from how the interaction unfolds. Then you can respond to it. For example, we produce utterances incrementally and process the feedback signals of the interaction partner to adapt our behavior online.

Catherine Pelachaud: It looks like you have mentioned three main profits about incrementality. One is for planning, one for managing interruption, and another one is to handle behavior generation planning with machine learning that bypass behavior planner.

Elisabeth André: Also, the error handling on the fly.

Catherine Pelachaud: They are four main features that are crucial to be handled with incremental processing.

Stefan Kopp: Although for machine learning approaches it's also a challenge. These approaches often process more or less complete patterns and map them to some output based on correlation and features and so forth. If such patterns are not complete yet, as is often the case in incremental processing, this mapping is harder to learn.

Elisabeth André: Unless we use them for prediction, as is often done in dialogue generation. Machine learning approaches try to predict the next turn, the next word, or whatever. But of course, this approach comes with deficiencies on its own. It's not understanding. It's more predicting what to expect in a particular context. But do we want that for socially interactive agents? Maybe for some applications we do not need a deeper understanding? But for other applications we want to have agents with intentions and plans, agents who know what they are doing and know what the user is doing and not just show some seeming natural behavior.

Joost Broekens: At this last IVA conference there was a paper that I liked. They had an end-to-end trainable model. It was very clear that what was good in that approach is the fluidity of the movement. But if you would look at the iconic meaning of the gestures this was not that good. There is definitely a need, especially if you want to have full control over the communicative acts. You're going to go toward hybrid architectures with a symbolic as well as a machine learning approach. But the question is when do you do what and how do
you mix them? Maybe you get the interpretation of the machine learning on the meaning of the gesture. But there is a lot to be said for both actually, shown by the striking difference between the communicative function and the fluidity.

A.3.3 Question 2

Birgit Lugrin: This discussion on machine learning actually already led us to the second question: looking into the future, what will be the next level of computational approaches? Is it machine learning only? We already know that it’s probably not machine learning only. But what should be considered, for example, symbolic AI or embedded semantic information.

Elisabeth André: We will probably have a hybrid approach. Of course, it also depends on the application. It might be nice to have a chatbot just for fun and for your entertainment. But for some applications, the agent should have a deeper understanding of the conversation. I recently tested some chatbots. They gave the impression of a meaningful conversation, but only for a short time. After a few turns, they contradicted themselves and destroyed the illusion of an intelligent being. Also, most chatbots have tremendous problems with simple things like anaphora. The fundamental question is whether to focus on natural or intelligent behavior. We probably need both.

Stefan Kopp: I agree. I think it’s key for agents to be adaptive in a very efficient way during the interaction and to produce a behavior in a way that is responsive to the specific situation, to the specific interlocutor, to what was said before, to the unfolding discourse, and all these aspects. That actually requires an agent to be capable of really fast learning and fast adaptation. The hugely data intensive machine learning approaches struggle with this. I think the question here is not whether it’s machine learning or not. It’s whether it is statistical AI that is based on correlation patterns, generalized based on statistical significance, or it is model-based AI and machine learning where you extract structural knowledge regarding causes, effects, and situation-specific parameters. I think we need both. A good example is gestures synthesis. We have models that produce body motion frame by frame, from acoustic or textual features. But they have no model of what they’re talking or gesturing about, or what the bigger communicative context is. In result they can produce co-speech beat gestures very well, but not representational gestures. They lack fundamental levels of semantic and pragmatic aspects because they cannot be easily extracted from massive amounts of audio and video data.
Joost Broekens: You could imagine a system where, given that you have sufficiently rich labeled data, you could train the machine learning model to be able to cope with all kinds of variations of behavior. But you would be able to condition the model on the same sort of semantic information that model-based or agent-based approaches are able to get from their planning or from their reasoning engine. So then you can condition the machine learning model based on semantic information. I think that would be interesting. I would like to try it. It's a bit like speech, what Google and many other approaches are doing for speech generation. You can do it with Tacotron,\textsuperscript{14} for example. You can actually condition the speech generation on quite a lot of features, if you can detect the features in the original dataset. Then if you reverse the model, you can actually control those outcomes, and the speech that's produced by the model is using in essence symbolic conditioning information. Then you get emotionally varying speech. But it's not the model that figures out when to do it, not the machine learning model. It's an agent-based model that could figure out when to do it, to show empathy for example. So there are definitely interesting ways.

Stefan Kopp: It also leads us to the next question because model-based or symbolic AI techniques are needed to be interpretable or explainable.

Elisabeth André: The question, of course, is how much transparency we need in which situation. An agent that is supposed to induce a behavior change in someone may be more successful when communicating information subtly and indirectly. However, agents with implausible behavior cause problems in many situations. A few years ago, we developed an educational environment with virtual agents to help children deal with bullying at school. The children could advise the agents, and depending on the agents' emotional model, the agents either followed the children's advice or not. Some children got upset when the agents did not do what the children suggested. The children did not understand that the agents were afraid to carry out the recommended actions because the agents did not always convincingly demonstrate their fear. In this context, I would like to refer to an early IJCAI paper by Phoebe Sengers\textsuperscript{15} on “Designing Comprehensible Agents.” She argues that intelligibility should be an integral part of an agent's architecture rather than adding an explanation component post hoc.

\textsuperscript{14} https://google.github.io/tacotron/.

Stefan Kopp: But then, the agent has to portray it in a way that is understandable and accessible to the human. It’s obviously something different than explaining the 2 million parameters in a deep neural network. That’s a common problem of explainable AI and it also applies to some of the models that we use in socially interactive agents. There’s XAI\textsuperscript{16} technologies to analyze black-box models and build interpretable models. But I think we have to bear in mind that we have two kinds of addressees, one is the developer and one is the user. XAI is really toward the developer, being able to analyze what the system is doing and what not and why. But we also have the user who needs to understand what the agent is doing and for what reason. And they need other explanations. The technology that we’re developing actually may help to build better explainable systems because, in order to make themselves really understood, they must have abilities for dialogue or multimodal communication.

Joost Broekens: When I think about transparency and explainability, I wonder why we want that so badly for this type of technology. I think there should be a discussion about why we want explainability in the first place for advanced technology. If you buy a car, I bet most people don’t know how the car works, but it is predictable.

Stefan Kopp: No, I don’t agree. Laugh.

Joost Broekens: That’s right for you. But a lot of people will only know how it works from an input–output paradigm.

Stefan Kopp: Exactly.

Joost Broekens: If something goes wrong, if something unexpected happens, nobody knows why, then you go to a garage. As long your conversational agent behaves according to what you would expect, predict, in a particular setting, it is fine. But if at some point it says: “Well, maybe you should buy this medicine.” That’s weird. That’s probably the moment where you would want it to be able to explain. But how do you do that. Explainable AI is almost always conversational.

Stefan Kopp: You would think so but in fact the field is still working to achieve quite simple forms of conversational explanations, for instance for recommender systems. We wouldn’t even call it conversational explanation as the

\textsuperscript{16} Explainable AI.
user is often only able to ask: “why are you recommending this hotel to me now” and then is presented with some additional information.

Elisabeth André: But it’s not really a conversation. It is a follow-up question.

Stefan Kopp: Exactly, but it is called conversational explanation nevertheless. But what you’re saying, Joost, is exactly right. Users have to have a good enough understanding of the system that is cognitively manageable for them and good enough for them to predict the behavior and make sense of it, and then it works fine.

Joost Broekens: And adaptive also, at least at some level; adaptive to what the user, at that point, needs. If you don’t understand why someone says or does something, then you ask the question: “Why do you do that,” you get an answer at some level. Then you say: “well I still don’t get this and this, can you explain,” and you selectively go into what you need to know. That’s very challenging.

Stefan Kopp: Yes, explainability has to fit to the user’s information need.

Elisabeth André: There is an interesting paper by Chromik and colleagues who discuss dark patterns of explainability. They present several examples of explanations that do not benefit users but instead deceive or distract people, for example, by overloading them with technical details. When enhancing virtual agents with an explanatory component, we should exploit their potential to communicate information using multiple modalities. Virtual agents also enable us to provide information implicitly. For example, in explainable AI so-called saliency maps are employed to highlight parts of the input in focus of a neural network. Instead, virtual agents may gaze somewhere so that the users immediately understand where the agents are directing their attention. So far, research in explainable AI that focuses on the end-user is rare. Virtual agents offer the potential to provide explanations in a socially interactive multimodal dialogue.

Stefan Kopp: Another aspect is that to some extent we even capitalize on the non-transparency of our agents. If people would really know how Amazon Alexa works, they would still use it, but it wouldn’t be that much fun for too many of them.

Elisabeth André: Yes, indeed. Explanations per se do not increase user trust. The key is the calibration of trust. In one of our experiments, explanations helped users choose the smarter agent out of two. They lost confidence in the other agent because they found out how stupid it actually was.

All: Laugh

Catherine Pelachaud: At the same time, there is also this question of transparency, for example, of the mental models the agent has of the users. When the agent interacts the users, it builds a mental model of the users. Do you make this information available to the users?

Elisabeth André: The theory of mind, that’s a good point in that context.

Catherine Pelachaud: That’s not so easy; it is what the model computes.

Elisabeth André: But it might be interesting to verbalize that: “I believe you would like to do this and that, because if I was in your position....”

Stefan Kopp: In our domains like conversation and interaction, it’s often needed to do it in order to resolve misunderstandings, for instance. It is a repair mechanism to use metacommunication which refers to beliefs about beliefs.

Catherine Pelachaud: We should understand if that should be done during or after a conversation. It wouldn’t be the same mechanism.

Stefan Kopp: If you use state of the art language technology, that’s not how they approach dialog processing anymore, not in terms of what the user or the interlocutor thinks, believes, or wants. They have a record of high dimensional embedding of dialogue state. That’s not something you could explain easily.

Elisabeth André: The benefit of the typical explanatory AI approaches for the end-user is not always clear. Many approaches highlight essential parts of the input to explain how a neural network came to a particular decision, such as nasty words in hate speech detection. Such information might be valuable for researchers who have to tune the neural network. However, it does not suffice to highlight parts of the input on which virtual agents focus to explain the complexity of their behaviors. It’s as if I don’t understand Stefan's paper, and I ask him for help, and he sends me the document back, where he simply marks some words and says: “Okay, that’s my explanation of my fantastic approach.”

All: Laugh

---

A.3.4 Question 3

Birgit Lugrin: So about evaluation. How can we actually validate the computation and architectures in the future?

Joost Broekens: It depends. You mean the architectures or the agents that embody these architectures? I think it very much depends on what you’re looking for. For some of these architectures, or at least their instantiations, you might want to investigate a particular interaction phenomenon, for example; another may want to study the desired effect in particular use case scenarios. I’m not sure if that’s what you mean by validating.

Elisabeth André: We may look at evaluation metrics developed in software engineering as a first step. Here, general quality attributes, such as reusability or maintainability, have been defined. However, we also need to identify quality attributes of the architecture that are particularly relevant to embodied agents, such as how well the architecture supports the implementation of dynamic and reactive behaviors of an embodied agent.

Joost Broekens: But it’s a different interpretation of validation. There are many ways for validation.

Stefan Kopp: I think it’s both. It’s fair to say ultimately, we are going to evaluate agents in terms of the behavior that the architecture is able to produce and whether that actually fits, as Joost was saying, the demands of the application context, whether that behavior is understandable, acceptable, convincing, and you name it. At the same time, you can also validate the architecture in more technical terms like how efficient is it, how does it scale up, how does it allow generalization from one domain to the other, can we apply it to both virtual agent as well as to robots, or parts of it. We don’t have very good measures for the latter, at least.

Joost Broekens: Many of the architectures are quite modular in the sense that they, for example, have an appraisal module that’s connected to a particular BDI\textsuperscript{19} engine. But there’s also a personality module, and others that get all glued together. In one particular scenario, a module does or does not do the job. It’s quite hard to figure out what the validity for one of those modules really is, and if it is even relevant. I find it harder and harder to decide on this. This is a really difficult question to be honest. You would want these modules themselves to be software tested, interoperability tested, but also behavior tested.

\textsuperscript{19} Belief–Desire–Intention.
If you enter information into your appraisal module and happiness comes out, you want to be able to check whether that’s correct or not in many different cases or whether this was just a lucky case. That’s difficult because the modules are not that isolated in the way they’re used; they are usually developed more isolated than the real thing.

Stefan Kopp: I think this difficulty to answer this question might be one of the reasons why it’s so hard to get papers accepted that propose an architecture.

Elisabeth André: Such papers are accepted when people release an implementation of their architecture to the public. And the impact will be high if the community exploits it successfully. But it usually takes time for the developer’s work to catch on.

Birgit Lugrin: It’s very important though.

Joost Broekens, Stefan Kopp: Exactly.

Elisabeth André: Since embodied agents are supposed to simulate human-like behaviors, an architecture should be grounded in social and cognitive sciences theories. The implementation of the architecture should not only work. There should also be a social and psychological underpinning.

Joost Broekens: On the output side, I always like variations of scenario-based verification of the behavior. If you test out a couple of scenarios with your architecture, you show what works and what doesn’t. At least you get an idea of whether the complexity of architecture is really needed.

Elisabeth André: An architectural framework would be helpful if it enabled us to explore questions we might have on the nature of human-like behaviors. Even if an agent behaves strangely, it might still provide valuable insights for social and cognitive sciences researchers. I would like to have an architecture that enables these kinds of experiments.

Stefan Kopp: Exactly. I was going to say the same thing. Usually, we take the engineering approach. We want to have a good architecture that produces good behavior that meets the needs of the project and the users. But we also have to address a cognitive modeling challenge. We would like to have an architecture that is plausible, like a good model of a theory or understanding, like a cognitive or social science theory. It’s also a way to validate the architecture, whether it’s a good model and whether it’s theory adequate.

Catherine Pelachaud: Do you remember when we worked on SAIBA at the first meeting in Iceland. Stefan and Elisabeth, you were there. We defined four or five scenarios. We spent quite a lot of time on defining scenarios that could
serve as testbeds to find that if we needed a feature but could also be used for testing. It’s quite interesting that would come up with these ideas.

Joost Broekens: I don’t know of a set of accepted scenarios for testing novel emotion models. I’ve been thinking about this for some time. I know that Jon Gratch uses negotiation as a domain to test everything in. You can call it a scenario. But it’s not like a validation scenario. That is, if you do this and this and this, this should probably come out of your agent with these probabilities. It’s not like you can run your agent against some test scenarios. That would be nice; like a standardized interface you could use.

A.3.5 Question 4

Birgit Lugrin: Let’s go to the last question. Are robots the new IVAs and how do you foresee the potential of augmented reality?

Joost Broekens: No, and I’m not an expert on the latter. Laugh. I assume we’re talking about humanoid robots, so socially interactive robots. I think they won’t be a new IVA at all. That’s my personal view. We can disagree. We are still very much in search for the killer use cases for social robots in the first place. We’ve been focusing a lot in social robotics on perception studies with robots and showing that robots are motivating for children, engaging, etcetera. But I think that if you put them in a usage context, then it’s very hard to make a very concrete case. It’s actually much easier for IVAs to make many different use cases. They’re much more confined in terms of the input/output and of the control you have over them. Social robots, especially social robots that have a motivating and engaging effect for young children, for example, that I know most of are lacking quite a lot in terms of social awareness. This is really a difficult issue. IVA might also lack a lot, but there are many use cases you could think of for IVAs that won’t hold for social robots that are physically present. The other question, I don’t know.

Elisabeth André: I want to comment on the potential of virtual characters for augmented reality environments. Virtual characters that inhabit and thus augment our real environment hold great promise for various applications. Tourism is just one example. For example, users could choose between Cesar Maximilian or the Fugger merchant as a virtual city guide when visiting the old city of Augsburg.

Stefan Kopp: I would also agree with what Joost said in the beginning about social robots. People in robotics used to say that robots create a stronger presence, and therefore they are more effective in terms of their impact on the users. That’s true to some extent. They create a physical presence. But the question is what kind of social presence does an agent create. So what Joost was saying is totally true. Robots have yet to deliver on a lot of things that are needed, for example, expressivity, fluent interaction, responsiveness. There’re still many technical challenges with this. Right now, it looks like that they are not going to be the next IVAs. One could even say that the field is facing a risk of drying out. There’s this no real killer application or fully convincing use case of a social robot. My impression is that people are moving from purely social robots to manufacturing or service robots, to try to make them more socially intelligent. This is really promising and important because such robots are starting to enter collaborative settings or assistive scenarios. I think these different branches of robotics are going to merge. Social robotics will inform the development of more socially interactive classical robots for different everyday settings in different contexts.

Joost Broekens: That’s a possibility indeed. There’s also something else going on with what you said. Many of the social robotics platforms unfortunately are very much vendor locked, even the open ones. But the point is that there’s not really a very good content strategy for social robots. I like to approach social robots more as a new interactive medium rather than an agent nowadays. We’ve focused maybe too much on the fact that this is like an embodied agent that is supposed to have all kinds of human values. While in essence what we’re supposed to be investigating—I mean not us as we’re interested in something else—but there should be people investigating how to develop interactive content for this novel medium, just as if it’s like a laptop or a TV or a tablet. It has been lacking for quite a long time, and as a result, there is no content frame for social robots. And as long as there isn’t a lot of content to get on these devices, it’s hard to see what can you do with them. Even though you can do quite a lot with them just because they are actually physical humanoids.

Stefan Kopp: But even if they’re not very humanoid or anthropomorphic, what’s important about them is that they are forced to have minimal social signals, like minimal elements of social interaction. And that’s really super interesting. On the one hand, with IVAs you go for the whole enchilada, the full picture you would like to have like a very human-like appearance and expressiveness. This is also interesting but a very complicated and daunting task.
I’m quite excited about social VR, augmented reality, or extended reality technology that we have now. I think they can really be a big driving force for a lot of IVA approaches and technologies. There’s right now a lot of exciting work on machine learning for social behavior processing, animation, or graphics for the purpose of enabling social VR. It’s going to be a big playground for a lot of work that we have been doing, also in the IVA community.

Elisabeth André: One advantage of virtual agents is that you can take them easily with you. You don’t have to carry a heavy device, but instead they accompany you on your mobile device. Virtual agents may also serve as invisible friends who encourage you whenever you are confronted with a challenging situation.

Joost Broekens: I honestly meant I didn’t know that much about augmented reality. But I do hope that if there is so much potential that the same mistake will not be made as with social robots; namely, that you get again caught in that vendor lock-in. One of the biggest issues that people are trying to now commercially build is pillars. But it would be great if you could develop a cultural intelligent virtual agent that can give you information and tour guides in tennis, for example. But I would like to be able to run it and download it on any device. Otherwise, I will never buy one of those devices if I am stuck to the software. This is really an issue at the moment. You see that with social robots as well as you saw it with mobile phones well. It is that only when you get standardized platforms that are big enough to support a large community of users, and so friendly fellow person, that you will actually get sufficient content for those platforms. Android was like that. It was a silver bullet away.

Stefan Kopp: But then again there are different motivations. Will the technology make it to the mass market? This is one question. The other one is, will it be good technology in order to make scientific progress?

Joost Broekens: Absolutely, absolutely.

Stefan Kopp: In VR, there’re big technological advancements. You can have it without markers now. You can have it on a mobile, even with very detailed face tracking. There’re a lot of things that we can really envision nowadays there, and there could be great new developments also in our field.

Joost Broekens: I agree.

Birgit Lugrin: Anything you want to add or anything you want to state?

Elisabeth, Joost, Stefan: Thank you for the opportunity.
Interview 3: Evaluation

For our third interview, taking place in October 2021, we have drafted the following five questions:

Question 1: Do we need new methods for evaluation other than perception studies? What could they look like? Should we enforce in situ studies?

Question 2: How shall we define benchmarks that capture the quality of social interaction or to measure learning gain?

Question 3: How will we conduct and control long-term evaluation and integrate social functions to facilitate adaptation, rapport, or engagement?

Question 4: Shall we share the user models and data used for adaptation with the user to ensure transparency?

Question 5: Are robots the new IVAs? How do you foresee the potential of augmented reality?

A.4.1 Participants

Agnieszka Wykowska, Principal Investigator at the Italian Institute of Technology leading the unit “Social Cognition in Human–Robot Interaction,” and adjunct professor of Engineering Psychology at the Luleå University of Technology.

Timothy Bickmore, Professor in the Khoury College of Computer Sciences at Northeastern University.

A.4.2 Question 1

Catherine Pelachaud: Do we need new methods for evaluation other than perception studies? What could they look like, and should we enforce in situ studies?

Tim Bickmore: Well, I have strong opinions on this.

Catherine Pelachaud: Please go ahead.

Tim Bickmore: I have one foot in the social agents and HRI world and another foot in the medical world. And in medicine, they care a great deal about evaluation, and they don't take anyone seriously until they've done significant large scale, ideally longitudinal, properly powered, randomized, clinical trials. Most of my funding comes from the US National Institutes of Health, and in order to get funding, you spend half of your proposal writing about your evaluation plan. So it's a very big component of doing research in the healthcare world. These need to be actual in situ studies; you're putting artifacts out in the world. You have to have some kind of health outcome, on
ideally objective, non-subjective self-report measures, some kind of a blood draw, or accelerometry, or some other kind of objective measurement of outcomes. So that's sort of the standard in that world. To me, research is all smoke and mirrors until we get to that point.

Catherine Pelachaud: And you, Agnieszka?

Agnieszka Wykowska: I'm an experimental psychologist. So I would say that for us, it's very important to have well-controlled experimental studies. And it is very important to understand that you have to develop mechanistic explanations, which are not necessarily possible when you do things in the field. I'm not sure what you meant by methods other than perception studies? Whether perception like strictly speaking perception, so people being seated in front of a robot and perceiving the robots, or being just in an observational mode, or whether you meant perception as general lab studies? So I would say that we definitely should go away from just perception studies, and we should have way more of interaction involved. Especially when we talk about social cognition, because we know from a second-person neuroscience perspective that social cognition doesn't work in just observational mode. Social cognition is for interaction. So if perception is meant as opposed to interaction, then I would say we definitely need to try to have experimental protocols that are more interactive. If perception is meant as laboratory studies versus in the wild studies, here I think we are still in a phase where we don't understand mechanisms of social cognition, and we need to develop theories, mechanistic theories, to understand and be able to predict what happens. And only then would I say that we would be ready, from a science perspective, to bring robots into the wild. So if we want to have scientific explanations, we still need to have well-controlled experimental designs.

Birgit Lugrin: Yes, it was really meant that way, perception study versus interaction study, and as a second part of the question whether we should go into the wild more. What do you say about that, Tim? Because you go in the wild a lot, to the clinics.

Tim Bickmore: I always think of studies in terms of the causal chain from the thing that you're manipulating to the end effects, which may be the final health outcomes after a year of interaction. So, at the very beginning of the causal chain is “do people perceive what we want them to perceive regarding the social artifact,” “do they perceive our manipulations?”, so simple manipulation checks. “Do they actually interact with the artifact?” “Do they interact with it if it's a voluntary system?” And then, what's the next step in the causal
chain? So how does that impact their attitudes toward their health behavior? And then from attitudes, how does that impact their intention to change their behavior? And then from intentions to change, how does it change their actual observed behavior? And after some period of time, how does their actual behavior affect health outcomes, that you’re measuring maybe months later. And things can break anywhere along this chain, of course. So perception is important, but it’s only the first step, in my opinion. I remember a discussion I had with Stacy about this, that whether a character does two or three hand gestures in a given interaction might have very little to do with somebody improving their glycemic control, if they’re diabetic, a year after that interaction with the character, right? So some of these things, like focusing on the intricacies of hand gesture for a virtual diabetes coach agent, are just not very important, as far as I’m concerned. Even though I find them fascinating and am personally very interested in them. But if I’m looking at this from a long-term outcome perspective, some of these things are not very impactful, I would say.

Catherine Pelachaud: But, and I’m going to be on Stacy’s side on this, it’s true that over a long period of time whether one or three gestures, well I agree it may not have a strong impact. However, the behavior of an agent can be interpreted as a different attitude or is part of building relationships with others. So, it may have an impact on the interaction. It is this building up of behaviors perception that may have an importance. It is not so easy to control, in situ and in long-term studies.

Tim Bickmore: I agree that these things can be important. And also very important is getting initial acceptance. So for all of these longitudinal studies, we see declining use over time, for voluntary use systems, for the most part. And so some of these subtleties can impact continued use, which then impact long-term outcomes. So they can be important. It’s just that the impact of agent nonverbal behavior during one particular utterance during one particular interaction, like the number of gestures the agent uses, may not have much of an impact long term.

Catherine Pelachaud: So how could we merge these up? So Agnieszka’s approach is one type of approach because she’s really trying to understand the

mechanisms, while Tim is working on having an impact with a concrete application in mind. So how could we work on bridging the gap? I mean, how could we use some mechanisms that you can understand, for example, to adapt the agent behavior for your given application? How could we do that?

Agnieszka Wykowska: Well, I think that once there are mechanistic explanations and theories that are based on understanding the mechanisms, then the applications will follow. So zooming out a little bit, first fundamental science and then applied science, right? So of course the problem here is about generalizability of certain mechanisms. So we often have knowledge about the very basic ways that things function and then we need to think about all possible application scenarios. And that's I think where the difficulty comes. So even if we understand how things work, especially when it comes to the human brain, when things work in the lab, that doesn't necessarily translate to real-life scenarios. Often in the lab we're lacking ecological validity. That's why I said earlier that I believe that interaction is important, even if it's in the lab. So at least we do have some ecological validity interaction, and at the same time keeping experimental control. So it's a long process, I think, to get from fundamental research to applied research. But ideally, if we really understood how the mechanisms work, we would be able to then translate them to applications.

Tim Bickmore: I agree. Of course, they're both important. You want to build your agent applications on a sound foundation of results from experimentation and theory building. On the on the flip side though, even within health behavior change, there're lots of theories that are out there. And even when someone says that they're designing an application following some theory, there's this huge chasm of interpretation of how exactly they implement the theory in their application. There is always a lot of interpretation and subjectivity in the application design process. So it's always very difficult to say whether somebody has done a high-fidelity instantiation of a theory in the particular application that they're building. It may or may not reflect the results from prior, more controlled experiments used in theory derivation. If that makes any sense?

Birgit Lugrin: Yes, absolutely. I also think both approaches are important if we want to push the boundaries in socially interactive agent research: gain fundamental knowledge in controlled settings in the lab and understand what actually works out there in the world in a given context.
A.4.3 Question 2

Birgit Lugrin: In case you don't want to add more, I would move on to the next question. So we were also very interested in how we should define the benchmarks to capture the quality of social interaction or, for example, measure learning gains or things alike.

Agnieszka Wykowska: I think it very much depends on what is it that you're interested in. So even when you're saying quality of social interaction, what does that mean? Whether you're interested in whether the interaction is truly social, then quality is how social it is, or whether it's a quality of social interaction in a sense of whether it's comfortable, it's easy, it's intuitive, right? So I think that would very much depend on what is meant by quality of social interaction. So if it's about social, whether you want to understand that the interaction is actually social, I guess the important comparison is always a comparison with another human, because that is the truly social interaction we have. I'm always a supporter of objective measures, not subjective questionnaires. So I always try to design experiments in such a way that eventually we manage to have some markers of indicators, objective indicators of interaction. So if I was to think about a benchmark for assessing whether an interaction is social enough, it would probably be a way to measure whether similar social areas of the brain are activated when interacting with a robot compared to a human. That would be like a measure of how social interaction is. When it comes to, let's say comfort, one can look at physiological measures, the degree of stress, or how easy it is to solve another task in terms of cognitive load, and things like that. So I would definitely be for objective measures. But then the way one would benchmark things would be dependent on what is it exactly that we want to measure?

Tim Bickmore: I agree. In the medical world, there are some measures that assess qualities of doctor–patient and nurse–patient interactions. For example, there's the RIAS, which is the Roter Interaction Analysis System, but that measures things like how many utterances each party makes, how much dominance there is in the conversation, how many questions are asked and answered, and what kind of empathic opportunities are presented and followed up on. So there are some objective measures in the medical world to try and capture some aspects of this. And then, of course, there's relational outcome measures, which are things like trust and working alliance. So at the end of an interaction, or at the end of a series of interactions, you know how well the social interaction leads to some sense of relationship. So I think those are also important. Also, any perceivable social quality of
an interaction can be observed and measured, as long as you can get multiple judges to establish inter-rater reliability. But to me, again, these are only principally important in terms of how well they lead to some outcomes of interest. There are outcomes proximal to a given interaction that are important, for example, do people want to continue interacting with this agent? Do they come back and again to continue the interaction? Do they comply or do they adhere to the recommendations that the agent is making of them? Those are some of the important outcomes of the social dimensions of the interaction that can lead to longer-term task outcomes.

Catherine Pelachaud: So does that mean that for you, the dimension you’re looking at and its different measures depend also on the application. For example, if it was more toward education, it would be other aspects you’d look at. So in the long run, it would be learning, self-regulation, this type of thing.

Tim Bickmore: Absolutely, yes.

Agnieszka Wykowska: Yes. I agree with that, too. That it very much depends on the context and application. I was now just having this thought about, again, quality of social interaction. We have studies that show that social is not necessarily always beneficial for human performance. So it can be very well the case that, and in fact our own data showed it too, that if there is a robot that displays a lot of social signals, it’s very distracting for a human. So if a task requires focus and efficiency of performing the task, then actually social might not be the best way to go. So that’s another example where context matters and actually social might not be the way one would want to go. Just to add on the diversity of contexts.

Tim Bickmore: Absolutely. We did a study that appeared in the last CHI conference, in which people could choose to interact with an agent or using a graphical user interface for different kinds of health-related tasks. We looked at when they chose one over the other. And basically, if the task was a brief transactional interaction, they did not want to talk to the agent. Personal preference also played a role. Some people just liked talking to the agent. And so then they would in general bias toward wanting to talk to the agent if it was a more narrative task. But again, if it was just transactional where they were reporting something, then they didn’t want to deal with the socialities because they’re inefficient, right? They take longer to engage in.

---

Birgit Lugrin: I am wondering, since you both agree strongly that benchmarks should be very dependent on the application and the context of the application, that on the other hand means that it will stay extremely difficult to compare across systems. If everybody is defining their own benchmarks, how will we deal with that?

Tim Bickmore: Test beds. Competitions where you have standard datasets that you’re working against. But I can’t imagine what that would be like in social interaction with agents, but we could invent something. But I think it’s important to fix the context and the application domain to have these conversations.

**A.4.4 Question 3**

Catherine Pelachaud: That’s for sure. So the third question is how we will conduct and control long-term evaluation and integrate social functions to facilitate adaptation, rapport, or engagement.

Agnieszka Wykowska: I would say that’s a tough one. So at least from my perspective, saying that most of our work is done in the lab, and in a very highly controlled environment, this is definitely something that is very far away from our approach. It needs to be done, certainly. But longitudinal studies, how does one do that, making sure that everything is properly controlled? As I mentioned at the very beginning, things get very different all the time. So that’s a really hard task to understand—how things will develop over the time, a longer period of time, and at home in a natural environment.

Tim Bickmore: Well, this is a lot of what I do. So I would say one key requirement that a lot of researchers have difficulty with is reliability. Your agents have to work, and they have to work in the wild, and they have to work for a long period of time. So that requires, unfortunately, that you often have to back off on some of the technology. You have to build simpler systems that you can thoroughly validate, and make sure they’re actually going to work for a long period of time. But then that does open up interesting possibilities of actually studying some of these long-term issues. How can you change things over time, by having your agents modify their behavior over a series of interactions? So it’s an interesting area of research in and out itself.

Catherine Pelachaud: Here we wanted to discuss about long-term evaluation, but also what you had discussed before, when to add social interaction in order to facilitate adaptation, rapport, and engagement in a study. So should it be to continue using the application or to ensure maximum behavior change?
The question is related to integrating some social function. As you mentioned before, sometimes it's good to have, sometimes it's not good to have. Social function could be used to enhance engagement, or rapport building, but could be against performance. So in which term can we measure a long-term interaction? In terms of performance, in terms of engagement, rapport, etcetera?

Tim Bickmore: Well, there're other measures of engagement, for example, the length of any given interaction. If someone is allowed to have an unbounded conversation or interaction with an agent in a given session, you can look at how long those interactions last as a more immediate measure of the quality of the social interaction. In addition to how long they continue using the system over time. You can also look at uptake rates. For example, the agent can present people with opportunities for doing social chat, or off task talk, and see how many of those bids for new topics that users engage in as another measure of social interaction. We can also look at the quality of the user utterances, for example, are they more task-oriented in nature? Or, do they demonstrate more of a social interaction? For example, are they using idiomatic terms of engagement, terms of closeness when addressing the agent, or using close or distant forms of farewell when they disengage from the interaction? Those are qualities of language that you can also look at to assess how well the social interaction is going.

A.4.5 Question 4

Birgit Lugrin: So last question is whether we shall share the user models and the data used for adaptation with the user to ensure transparency? Or shouldn’t we?

Tim Bickmore: Well, probably we should, especially if this is something that is put out in the real world. It gives people a warm, fuzzy feeling to know that they could look there if they wanted to. However, it is likely that the vast majority of people will never bother. There’re a few people who really care about these things. But to most people, it is just too much work, and they don’t have time to bother. But that’s just my opinion. So maybe you just have a button that says “look at the user model,” but it doesn’t do anything, that would probably work just fine.

All: Laughing

Catherine Pelachaud: That is somehow related to a question we have on ethics. Since the system is building a representation of the users, and the users may,
most probably, not know what this representation is. But those representations are being used by the application. So you may want to have access to those information.

Birgit Lugrin: True, but it might also break the impression of having a truly socially interactive agent when you share the user model, and you see that it is actually rather simple.

Tim Bickmore: But, you know, users don’t understand, for the most part, how those models are used. It can be very complex systems that are interpreting them in different ways. So I think it’s more about just giving them a feeling of trust or confidence in the system, rather than actual understanding.

Catherine Pelachaud: Yes, and how can you build the system of trust into the entire system? Not only in the agent, but the entire system? Because as part of a long-term study, if you don’t get trust, people just stop interacting.

Agnieszka Wykowska: I completely agree with Tim on this point. I think it’s probably a feature that is nice to have for the feeling of comfort or trust, but probably doesn’t do much in practical terms. And regarding trust, I think it’s one of those very big words that are being used often, and they’re not so well defined to really understand how to measure trust and how to improve trust. Because what is trust in the end? It has so many dimensions. Is it about trust in terms of trusting the reliability of the system? Or social trust? And probably we trust other humans in very different ways. So with persons, we would trust a person in some dimension, but not necessarily in another, right? So I think it’s one of those concepts that is being used in HRI, or HAI, a lot without very clear definitions. And I think it would be very useful to have clearly defined concepts, such as trust.

Tim Bickmore: I agree. It’s one of those, what Minsky calls “a suitcase word” that has lots of lots of different meanings, right? There are some really clear ones: I’ve had a patient who installed a system in her home and came back a month later, and she hadn’t turned it on because she was afraid it was beaming data back to the hospital about what she was doing. So there’s a clear example of not trusting the system. But there are much subtler versions of that.

Agnieszka Wykowska: But maybe she will trust in how reliable the system beams the information.

All: Laughing
A.4.6 After the Questions

Catherine Pelachaud: So do you want to add anything? These were the questions we had in mind.

Tim Bickmore: Well, this is about evaluation, so we could talk about attempts to develop standard measures for agent interactions for the virtual agents community.

Catherine Pelachaud: Siska Fitrianie.

Tim Bickmore: That’s right. So that kind of work is important to do. However, there is a vast amount of work by psychometricians, working in psychology and medicine and educational psychology, who have spent years developing validated measures for many of the things we care about. They do validation studies of self-report measures that involve hundreds, if not thousands of people for long periods of time—to ignore that work is, I think, to our detriment. I still see researchers coming up with their own measures for their own studies. One principle I often see violated is that you should never invent a new measure in a study where you’re also using that measure as your primary outcome. You should do a separate validation study for it first. I tell my students to first spend a lot of time looking for an existing validated measure that taps into what you want before you even think about inventing something new, because it’s not our area of expertise and we don’t have the resources to properly validate it. So I think there’s a lot we can do in practice to improve the quality of our evaluations, by looking more toward what other people have already done.

Catherine Pelachaud: In human–human studies you mean?

Tim Bickmore: Yes, I think just because it’s a social agent doesn’t mean you have to invent a new measure.

Catherine Pelachaud: Yes, true. Any other recommendation?

Agnieszka Wykowska: Well, from my side it will be something that I already mentioned earlier: we have to define concepts in a more clear way, that are more easy to be operationalized. There are many concepts in HRI and HAI, as we said trust, empathy, engagement, and these are very complex concepts. If I were to be asked how to operationalize those, I wouldn’t know how, unless

we first come up with clear definitions. So I think the field really needs better, more operational, usable definitions of these concepts.

Catherine Pelachaud: Yes, it's true that if you have a very clear definition, you understand better what you measure.

A.5 Interview 4: Ethics
For our fourth interview, taking place in November 2021, we have drafted the following six questions:

Question 1: What is an ideal SIA? Should they be the perfect assistant or a companion?

Question 2: Should we as researchers on SIAs go beyond stereotypes, regarding gender and other diversity factors? Or should we follow user preferences (maybe including stereotypes)?

Question 3: How should we draw the line between persuasion and manipulation and transparency, for example, in health-related applications?

Question 4: Should SIAs be better than humans? What does it involve anyway?

Question 5: How do we manage dependency and addiction that potentially occur in a relationship with a SIA?

Question 6: How shall we deal with the popular fear of robots overtaking the world?

A.5.1 Participants
Ruth Aylett, Professor of Computer Science at Heriot-Watt University in Edinburgh.
Kerstin Dautenhahn, Professor and Canada 150 Research Chair in Intelligent Robotics at the University of Waterloo in Ontario, Canada.

A.5.2 Question 1
Catherine Pelachaud: So the first question is what is an ideal socially interactive agent? Should it be the perfect assistant or perfect companion or something else?

Ruth Aylett: I don't think there is an ideal one. I think the answer to that question is bound to be *it depends*. What are you trying to do? Generalities are a bad idea in AI and robotics. General systems don't work. Specific systems tailored to particular niches can work and can do useful things, but then you must tailor the SIA to your niche. Which means there isn't an ideal SIA at all.

Kerstin Dautenhahn: Yeah, I would agree it depends very much on the tasks, on the application areas, and also on the people. Is it about children? Is it
about adults? Is it about people from specific groups? I think that this really depends. Also, I don’t see a contradiction between assistant and companion. Companions can be assistive, and assistants can have an element of a companion in terms of being a friendly presence. So as we said, I would agree. It depends.

Ruth Aylett: So an assistant is a functional term. It describes a set of functional capabilities. A companion is not, in that sense, quite the same sort of term because it suggests an affective relationship, and something that’s involved in your social life, and not just in particular tasks. It’s not necessarily a task-related term. So I think those are two very different terms. If an SIA has no useful functionality, then I don’t see that it will be accepted in whatever niche you’re trying to put it. Because as we’ve seen from the failure of so many companies, the purely social aspects are poor compared to humans, and they are going to continue to be poor compared to humans for a long time. So if there’s no functional aspect to an SIA, I don’t see that anyone is going to give it floor space in their lives, whether professional or personal. So I think there has to be some kind of functional capability, whether you call that assistant or something else. As for companion, we’re getting into the ethical issues further down about affective relationships and people becoming dependent. So let’s park that one for a minute, unless we mean something completely different by companion than we would in the human case. Which is quite probable, I would have said.

Kerstin Dautenhahn: Well, I still think assistants, in order to be successful, in order to carry out their tasks, they need to have some social abilities. Whether we go down the line of encouraging relationships, that’s a different topic. But if you talk about therapy, for example, therapy for children, you cannot have a robot that is purely functional and just tells children “you have to do this,” “you have to do that.” No, for that particular application, you would need social engagement. So this is what I meant when I said that sometimes these terms can be overlapping. But I do agree, bottom line is we need these robots or agents to be useful, to actually do something.

Ruth Aylett: I definitely agree with what you just said. My point was almost the opposite of yours, which is you can’t just have social capability.

Kerstin Dautenhahn: Oh, OK. Yeah, certainly I agree.

Ruth Aylett: I agree with you that just having functional capabilities doesn’t work either. Particularly in the sorts of niches we’re looking at. So, if we don’t use those terms, if we use functional and social instead of assistant and companion, I feel a bit happier with the discussion.
We use those terms as they are commonly used in the literature. But I agree, it’s better to view it as a functional task or social.

**A.5.3 Question 2**

Birgit Lugrin: Thanks. So I think we can directly move on to the next question, if that’s fine by you. So we were wondering, should we, as researchers on socially interactive agents, go beyond stereotypes regarding gender or other diversity factors? Or should we follow user preferences that might include existing stereotypes?

Kerstin Dautenhahn: These days there are a lot of discussions on stereotypes, and also about norms. Should robots follow certain norms? I was part of a workshop where I brought up the metaphor of the “echo chamber,” which we all know very well. And so the question is, do we want to develop these agents in a way that they just match what people expect? Or do we want to make them into some interesting artifacts that might challenge some of these stereotypes and expectations? I myself would certainly prefer not to have robots that are presented in a very human-like way, for example, android robots, as a very particular person with a very particular background and gender. I would prefer more neutral and more challenging robot designs, and to base designs more on artistic skills rather than just trying to faithfully emulate human-like shapes and behaviors. So I think it’s an open question. If people have certain expectations and these expectations are not met, then they might disengage from the interaction. But on the other hand, I think through the design we can also challenge people and go beyond the stereotypes and norms. Which could make the human–robot interaction experience more interesting.

Ruth Aylett: I think there’s a different answer for graphical characters and for robots. Very human looking robots are a disaster area normally because they raise expectations about actual behavior that we can’t meet. Even if we wanted to. So trying to produce a very human-like robot is going to produce not stereotypical behavior but bad behavior, which is going to annoy people. Unless in very, very short interactions, like some of these scripted interviews you see Sophia doing. But anything that’s truly interactive, where the niche isn’t very, very narrow, like one or two interactions, a robot is going to fail if it produces expectations of humanness. Because it won’t behave like that, it will glitch, it will fail, and people will just get irritated about it. So unless you are in a deceptive situation, which I would say these Sophia interviews...

---

24. [https://www.hansonrobotics.com/sophia/](https://www.hansonrobotics.com/sophia/)
are, I don’t think that in the robot case there’s any point whatsoever in trying to make it look really human. And I agree exactly with what Kerstin said: We should look at art, and in particular at animals, cartoon characters, stereotypical machines maybe—depending on people’s preferences, and not at human stereotypes at all. So stereotypes are a loaded word, and it’s not clear in context what exactly you mean by it. If we think of graphical characters, we can produce a very photorealistic face and so on. Although again, its behavior will not necessarily match that appearance, it won’t be as fallible as a robot would be in the same situation. The ideal of young, attractive women has passed its sell-by in that particular area. I would like to see a little less of the routine assumption that that is the way graphical characters should look. For example, how about women that are not “beautiful?” If you’re going to have something that’s gendered, then women that are older, or some male characters. In principle, I think Kerstin is right. We would be better not having to gender the characters, but it’s almost impossible to stop people doing it. We try. We have a robot called Alyx, which is the Emys robot. It doesn’t have any amazing human-likeness and it doesn’t have any real gender cues. But people still gender it, in spite of our best efforts. So I don’t think you can stop people from gendering SIAs. You have to be aware of that. I came up with the idea of social affordances when I was talking about this in my book. So there’s a chapter about appearance, which is one of the things we’re talking about here and the idea of social affordances, which is: you look at a social agent, what do you think it can do from its appearance? A very human-like appearance gives you one set of social affordances, a machine-like blocky tin person gives you a different set of affordances. And these are expectations. So you could say they’re stereotypes that the user is then going to apply to the interaction. But you’ve also got behavioral stereotypes, and I worry about those too. So do we make these characters submissive? Do we make them overhelpful? That’s where Amazon’s Alexa went wrong, I think. It’s got the persona, though not the body of an overly helpful young woman. And as a result, in conversation you get sexual abuse, and get some really horrible dialogue. And that’s partly because of the way it taps into the misogyny of the people who are interacting with it, because it’s an overhelpful young woman.

So, yes, we've got to be conscious of these things. What social affordances are we providing and to whom?

Catherine Pelachaud: People have worked on the question of the voice. How to make a genderless, non-gendered voice?

Ruth Aylett: Well, a low woman's voice is not very different from a higher man's voice. Certainly, I've been mistaken for a man many times on telephones. So we can make the voice a sort of alto which could be gendered either way. It's not difficult to do that. It's a little bit different when you're using unit selection voices because these are based on real people, and they tend to accentuate the particularity of the person on whom they're based. So if you look at a male unit selection voice, it does tend to be very male and the female voices do tend to be very female. But that's because they picked people who have rather identifiable voices. And that's a choice. You could get people whose voices were not quite so gendered if you wanted to.

Kerstin Dautenhahn: If I think back two or zero years ago when we tried to make robots speak, there were only very limited choices on the voices. And sometimes we just ended up with a horrible voice because it was just what we had. But these days I usually encourage my students, if they have to pick a voice, try to make it as gender neutral as you can. And ideally also with a little robotic aspect to it, to at least not obviously invite people to treat it as if they are talking to a human being. And similar to what has been mentioned before by Ruth, there is the tendency to anthropomorphize. Regardless of what you present to people in terms of how the agent looks like, or how it behaves, people will end up saying “he” or “she” in my experience. I always say to my students when we write a paper, it is “it,” the robot. It, not he or she. And that's also how we should talk about them. But it doesn't really work in my research group. I'm not very persuasive, it seems. Whenever I have a meeting with my students and they talk about the Fetch robot, for example, a very mechanical looking robot, they often say “he.” And when they talk about Pepper, they often refer to it as “she.” And then I have to remind them it's a robot. In society in general, we are not encouraged to think more about gender fluidity, but it's very difficult to get that into our daily, day-to-day spontaneous language when talking about robots.

Ruth Aylett: Very difficult. You raised a very interesting point there about voice. I think there are specific issues to do with voices. We as researchers have

thought about appearance over many years, quite a lot. And we have thought very little about voices. So Pepper has a voice, which in my experiences, if you interact with it over a long term, it begins to get on your nerves very badly. And now, even more so, those voices are not particularly gendered, or they are childlike, if anything, in pitch. But they're not very comfortable to interact with over extended periods. We can use unit selection voices instead, and they are much more comfortable to interact with over long periods, but they do carry human baggage. So we have to ask ourselves, do we then say we shouldn't use unit selection voices, even though these are more comfortable for people over longer-term interactions, because they carry baggage with them? Should we go back to horrible synthetic voices which people are not going to want to interact with over long periods because they're just too irritating after a while? They do work in a demo but there’s a difference between what works in a demo and what works over the long term. An itchy voice works well in a demo. It’s kind of, wow, it’s a robot, it’s got a voice. And then if you do it for three days, it’s like, oh God, that robot gets on my nerves with its voice.

Kerstin Dautenhahn: Just as a side remark: my team created a voice for Fetch and they decided, I was not involved in that decision, they wanted Fetch to speak with a British accent. They could choose between American, Canadian, or even French English accent. And now they sent me a video of it. I found it really, really interesting. You have this very mechanical looking robot, it’s a more industry type robot, very strong, too. It’s certainly not what you would expect for a social robot. And then it speaks with this perfect British English. And I found that interesting. But as Ruth said, of course we need to see how in more repeated interactions that might work. But in a short-term interaction, it certainly creates this level of surprise. I mean, I was surprised when I saw it. A Fetch robot and a British accent, where does that come from? It was kind of surprising and then questions people, it doesn’t look anything like a human and it doesn’t look very social, but it speaks in this way. But we need to see, because long-term interactions are really key, and it is also in long-term interactions that people might realize how useless a robot is.

Ruth Aylett: I mean, the voice carries more information than gender. So particularly in Britain, the way in which you speak gives you an indicator of class as well, typically. So never mind just using a British, I would say Received Pronunciation (RP) voice, which may be what you’re talking about here, you can also give a robot a regional UK voice. So we’ve used a Scottish female voice on our robot typically, not a hard-to-understand Scottish, it’s what I would
call middle-class Scottish, but it’s clearly Scottish. We could have given it a Glasgow male voice that would give it a different impression. In either case, the voice being better than the appearance or more human-like than their appearance makes people think better about it than if it were the other way round. So we’ve done experiments on voice, and there’s something to do with the relationship between the appearance and the voice. We haven’t quite got that. We did an experiment with the NAO on the Emys with children, where we swapped the voices round. We ran them with the voice they came with and we also ran with the opposite voice swapped round. Then we asked people to evaluate them, and it turned out that they liked them best the way round that they were originally. They liked the NAO more as a sort of friend because it’s more childlike. I think the Emys carried more authority, and its voice was an adult voice, so that also carried authority to children. And they didn’t think it was as friendly as a result. And they didn’t like it very much with the NAO voice and vice versa. They didn’t like the NAO very much with a Scottish female voice, either. So there are social affordances, again. You’re creating expectations and you’re creating an image in people’s minds of what this thing is and what its abilities might be. When you give it all of these things, these bits of behavior and voices are much more important than I think we’ve really investigated.

A.5.4 Question 3

Catherine Pelachaud: So shall we go to the next question, which will be how should we draw the line between persuasion and manipulation, and transparency, for example, in health-related applications?

Ruth Aylett: Well, that’s a good one, isn’t it? Which is where the ethical issue becomes rather pronounced. Normally you would draw the line, if you were a person, by telling people what you’re doing. So you would say, I’m going to try and persuade you that…. Rather than being like a sneaky advertiser and trying to get underneath your psychological defenses. Humans do this kind of thing all the time. I agree that we shouldn’t try to manipulate people with our artifacts. I don’t think that’s really very ethical. Even if we do it as people. I don’t think we should make a habit of doing it with our artifacts. In which case you have to tell people what the system is, and what its limitations are. They have to know that it’s a piece of software and hardware, not a person, and they have to know what its aim is in relation to the interaction. And even

if that breaks the interaction a bit, I think you have to tell people what you’re doing. I don’t think you should deceive them about this.

Kerstin Dautenhahn: Yeah, I think there’s also a very fine line on what do we mean by manipulation. If I give the robot a certain shape, or behavior, or voice, or facial expression, and therefore make people think the robot cares. For example, if the robot laughs when someone tells a joke, or has an empathic expression and gestures and body language in a healthcare context, after all, this is faking, right? I mean, some people might disagree with me, but I don’t think robots have genuine emotions. They can, of course, simulate them. You can use them in order to control their behavior. They can certainly express a lot of emotions. They can perceive emotions from people. We have more and more sensors now, not only vision, physiological sensors, for example. They can know a lot on what people are doing and what states they might be in, and maybe detect their facial expressions, and then they could respond in a social way. But as far as I’m concerned, there is also a level of deception here because this is modeled according to human-like behavior. But the robots are not human-like. Robots do not actually know what pain is, or what love is, or what disappointment is. They can pretend really well that they do. Of course, I would completely agree that no one should intentionally deceive people or manipulate people. So, in a way, robots should be “better” than humans or more toward the ideal on how we think we should be. We should be honest with people. We should clearly say what our intentions are and what our goals are. But sometimes you cannot avoid deceiving people, although you don’t want to. But for me, deception is already if you encourage people to create a mental model in their mind about what the robot is, what it can do, what it cannot do, that doesn’t really match their real capabilities. And that’s really difficult. But it’s something I’ve seen in the last few years. There’s a lot of literature now in the field of HRI in particular about transparency and also explainable AI, which is related to that. So we need to see how that develops and whether any guidelines will come up. Researchers have already been developing some guidelines about transparency.

Ruth Aylett: They have, yes. There are people in Britain who did this with the funding organizations. There’s a set of principles to do with robotics, which seem very sensible to me\(^3\). You can overestimate how successful these systems are. So this is an interesting question. In practice, the ability of any of these systems to be sophisticated enough to manipulate anybody is a bit limited.

---

So I would disagree, Kerstin, that our systems recognize emotions. They're not even good at recognizing social signals, never mind what the emotions are behind the social signals. That was our experience when we researched our empathic robot tutor. It was really quite poor at this, and we were using physiological sensors, and we did have a strong context, which was a teaching game for map reading and it was still very difficult. So I think you can exaggerate how successful these systems are going to be. The problem you've got is that people want these things to be human, just like they gender them. So if you look at the original video, I think it's called Mechanical Love, where a long time ago a Danish author in a wonderful documentary looks at the Paro robot and looks at the rather arguable things that Ishiguro does. Without much commentary, she sits in the background. She follows the Paro's testing in a German old people's home. They test it with a lady who actually doesn't have dementia, I think, but that she hates the home becomes very clear. She hates the home, and she hates the other people. She's really miserable. They give her the robot. They unzip its belly. They show her it's a robot inside. Thereafter she treats it exactly like a living thing. In spite of the fact that they've done the ethical stuff, they told her, it's not a living thing. And in spite of the fact that it doesn't do very much, it doesn't talk, it hasn't much in the way of social behavior, just a bit of face following and a bit of recognizing the sound of one's voice, and a bit of response when you stroke it. She treats it like a living thing, much to the irritation of the people in the home. It's quite funny, actually. So even if we tell people this stuff, they will still impute this behavior to our systems, like gender. And there's not a lot you can do about that: if you like, people are manipulating themselves, because we all have these strong theories of mind. I mean, come on, we impute intentionality to printers and photocopiers, never mind to something which talks. So it's going to be very hard to stop people doing that. Yes, I agree. You have to tell people what your limitations are and probably at regular intervals over a long period, but over a long period people are going to notice the limitations, believe me. OK? You might get away with this for half an hour, but you're not going to get away with it for a week in the current state of the art.

33. See https://www.imdb.com/title/tt1186021/.
Kerstin Dautenhahn: I’m just sometimes getting worried about this aspect when people try to encourage affective interactions, because this is in a way different. I do agree people also talk to their printer or swear at their laptop if it doesn’t work. But they would have no problem buying a new one. If it breaks, they would not shed a tear, other than about the costs of the new printer. They can make the distinction between “Oh yes, I sometimes treat it as if it were human-like, but actually I still know it’s just a printer,” but I think this is different. Maybe in this care home example that you mentioned, and from what I hear Paro is indeed really successful in many care home settings, but this is different. I think Paro is encouraging people to treat the robot like a living thing, a pet-like robot. So I find that very worrisome sometimes. Although I can clearly see that there could be really good benefits.

Ruth Aylett: It works well for what it was designed, which was interaction with people with dementia who are not good at recognizing individuals anyway. I think the problem with a system over the longer term is that you invest it with a memory of your interactions and with a perception of its pattern of behavior over time, which we would call personality if displayed by a human. And that would seem unique, and therefore you would be concerned about it being killed and vanishing from the world. Of course, there’s no reason why we can’t download all of that into the next one, and it becomes resuscitated. It becomes a reincarnation of the original one. We probably have to think about things like this. That’s not human. And I think people will adapt their perceptions if or when they get to do long-term interaction with these things. I think at the moment they tend to be knocked out by the short-term interaction. They don’t think about the long term. It would be interesting to study Paro over the long term in care homes, if anyone’s doing that, and to see what its impact is, not just on the patient with dementia who has little memory but with the care home staff. People are running the care home, people with longer memories who will remember its behavior and may attribute personality to it from observation. I think we need to know what people think of these things over the longer term, indeed.

A.5.5 Question 4

Catherine Pelachaud: Let’s continue with the next question, which is: Should SIA be better than humans? What does it involve anyway? This question relates to the notion of the agent being perfect, ideal, etcetera, but also it relates to the relationship the agent builds with the user. For example, we say the
agent should be engaged, it should adapt, show empathy toward the user. The question is to be taken in those terms.

Ruth Aylett: Well, they aren’t going to be better than humans. Let’s be right about this. But as a goal, that would be a pretty implausible one. They are going to be a lot worse at everything. I think it depends on what you mean by better than, doesn’t it? If you want to say, should they be more ethical and should they be nicer to people than humans sometimes are, then probably the answer should be “yes,” we should produce good behavior. After all, we’re engineering this. If it behaves, as we would say in the human case, badly, then it’s because we’ve engineered it to behave badly. And then there’s an ethical question. When we were doing our empathic robot tutor, which we designed to help learners with boredom and frustration, I said “well, the educational application is too easy. So we’re not seeing enough boredom and frustration. We need to make it harder. So learners get frustrated and then we can deal with the frustration.” And some of the teachers said: “well, that doesn’t sound very ethical.” My suggestion had a good reason behind it, but I could see their point. So in principle, depending on our niche, we probably do want the thing to behave more patiently, more ethically within its limits, which are going to be severe, and just not get very angry or very grumpy or any of those affective states without a very good reason which had better be ethically determined as well. However, it will fail. We know that. Of course, it will fail. It won’t behave deliberately badly. It just won’t behave properly, as intended. People may well interpret this as bad behavior. But it’s just the fact that these artifacts are not very good, currently. We know this.

Kerstin Dautenhahn: Now, when we previously talked about stereotypes, and I said that, maybe, robots can be better than people. What I was really referring to was the more ethical behavior, not better in the sense of physical capabilities or cognitive capabilities or even social capabilities. But more in the sense of we, as engineers, certainly don’t want to put in stereotypes or biases intentionally. I’m not sure if we can actually avoid that, when we use AI, for example. But better in the sense of making the robot more rational. I’m surprised that I’m using this word. But I think actually, in particular now, the pandemic is really affecting people on more than one level very, very deeply, and I sometimes see behavior that I wouldn’t expect in a more normal context. I think it’s because of all of these effects and because we are people, we are affected by everything that happens. I’ve now stayed mostly home for almost two years. For a significant percentage of my children’s lifetime, they have mostly stayed home. And so it is having an effect. People are affected
by everything that happens to them, the people they talk to, what they're reading, what they experience etcetera. And so robots, of course, since we programmed them, can be made in a way that is less irrational in the way how sometimes people behave. As Ruth said, I would not want a robot to suddenly shout at a person because they get really frustrated. The robot gets frustrated with the person because the robot may try to convince the person: “Oh, you should do this or you should do that.” The person doesn’t do that. A human would, maybe at some point if that continues over time, get frustrated and maybe even behave in a way that is not so nice, but robots definitely shouldn’t do that. They should be more balanced in how they react to things, and they should always put the welfare of the person that they are supposed to assist or be a companion of first. They should always have the person as their first priority. I’m not so much worried about how the robot feels and more concerned about how do the people feel that the robot is interacting with.

Ruth Aylett: The robot can’t feel, I agree with you. It has a model of emotion that’s all, and just like a model of rain isn’t wet, a model of emotion isn’t emotion. Nevertheless, people interpret it as having these internal states. Here’s an ethical question we haven’t posed and certainly haven’t answered as researchers. Why do we allow these systems to say “I” when we know they don’t have an “I?” “I feel,” “I say,” “I think.” We allow our systems to say these things. They don’t do any of those things. Those are deceptive statements. We do not have to do that. So, in the back story that Tim Bickmore and colleagues’ patient robot told people, which improved interactions incidentally, for post hospital treatment, they gave their character a back story and this made it more successful. It said “I like” various things and chatted about them. Of course, it didn’t. It doesn’t. It can’t. We could get around that by saying “In my memory” or “I’ve been given the ability to” and we wouldn’t want to do that, would we? Because we’d say that will break the interaction. So, we’re on the knife edge here between good interaction and deception. These systems do behave differently from humans and in some ways better. So the robot Paro is virtuous. It never gets upset with people with dementia at all. It will carry on doing what it does. That makes it of limited use to people without dementia, incidentally, as they will soon get the idea of what it does and lose interest. People with dementia don’t lose interest because of their

memory problems. They still find it new and interesting. It’s like what ex-US President Ronald Reagan said, “every six seconds, old people I know are new people.” Same thing here. It behaves better and certainly behaves better than a real animal. It will never bite people. It’s never sick on the carpet—like my cats all too often are. It’s better than a pet would be. But it also behaves worse than a pet would because it doesn’t have the same variability. It doesn’t have the same repertoire of behaviors. It doesn’t have the same selfhood. You can’t look into its eyes and know that there’s a personality in there, because there isn’t. So it depends what we mean by better, doesn’t it? More patient? I think robots should certainly be more patient than people, even if their model of frustration is such that it’s getting very high. The Action Selection System should never then come in and say “when frustration gets this high, shout.” That’s a programming choice which we should never make. So, there are some ways in which it should behave in ways that humans would not when it comes to negative social behavior.

A.5.6 Question 5

Birgit Lugrin: Talking about pets and about the companion robots, it actually neatly leads to the next question: How do we manage dependency and addiction that can potentially occur through the relationship with those socially interactive agents?

Kerstin Dautenhahn: I’m not sure about addiction. Do we have any examples of people being addicted?

Birgit Lugrin: We have smartphone addiction, for example. People see their smartphones as their companion and they get really addicted by them. That could probably potentially happen with interaction with your robot or with your virtual agent as well.

Catherine Pelachaud: There was this example, you recall, the Tamagotchi. That was not really a SIA.

Kerstin Dautenhahn: I wouldn’t call that an addiction. These days, people are also very fond of computer games and so on. But robots are still very, very limited in what they can do. For example, let’s consider the new version of the Aibo robot, which wasn’t produced for several years, but now there’s a new generation of Aibo robots. In Japan, there are some places where, if your Aibo robot breaks, you can go there and it will be repaired. And people actually prefer

to have their old robot be repaired rather than getting a new one. So whether you call that addiction or just an intense attachment to an object, I'm not sure. Addiction would mean that people then don't do what they would normally do. That's my interpretation. Many children these days, for example, play a lot of computer games. Most are not necessarily addicted to them, although they might try to play as much as they can. But it's probably OK as long as they are still doing all the other things that they should do as a child, doing homework, playing sports, and so on. So with dependency, I'm more worried about dependency for vulnerable people. I'm not so much worried about healthy adults. I'm more worried about, for example, people with dementia or children where the boundary between, is it a robot or is it a living thing, might very easily get confused. That's where we need to look in much more detail on how do we want to design the robot? Because dependency clearly is not healthy, let alone that these robots at some point will break. If a person thinks of the robot as a living thing, then they would be devastated if that robot breaks. So that's certainly something we need to avoid. But given the state of the art of social robots in particular, I don't think there is a danger of an addiction imminently coming up. But clearly, it's about the ethics of how people who design those robots and programmed these robots, what type of attachment or human–robot relationship they want to encourage. This might lead to dependency, and it might lead in future to addiction. At the moment, the field has more of the opposite problem, namely that as soon as people, in repeated interactions, interact with these systems, they often lose interest. So it's the opposite of addiction. But addiction to robots is certainly something we should avoid, like with any type of addiction.

Ruth Aylett: I think we have to be careful about overstressing this. I agree with your practical remark. The chances of this happening are low. Paro might produce dependency. You could always say its behavior is not so sophisticated that you could tell one Paro from another very, very easily. So I suspect that replacing one Paro with another Paro in its intended audience probably wouldn't be very noticeable. When it comes to children, we should be aware that children are already dependent on things, like favorite toys, because they impute personality to them. Many children will be very upset indeed if they lose a favorite toy, particularly when they're very young. You may have to buy them another one and hope that they don't realize you've done that. This happened to one of my grandchildren who lost his favorite soft toy; his parents bought another quietly when they couldn't find it. They also have pets in a lot of homes. Pets die on a regular basis. Children are devastated when
their pets die. But we don’t stop people having pets because of this. So I think we have to be realistic about some of this. The other thing is, I don’t think people are addicted to their phone, they’re addicted to some of the things behind the phone. The actual physical piece of kit gets replaced on a regular basis. You stick a cable in there and suck everything down into another phone. And there you are. You’ve got your phone back again, haven’t you? Because it has everything that previous phone had. Hence my remark about resurrecting robots. We could easily do that with robots too. Certainly graphical characters, they never die. You can just transport everything that was in one into another so they don’t have to die at all. You can replace them in the sense of their long-term interaction, any information they’ve acquired and all the rest of it. Just the hardware would need to be changed. There are ways around this problem. I don’t know that we want to encourage dependency necessarily. I do not think you can resist it. People will or will not become dependent. But I don’t think we necessarily want to encourage that, but we will have to deal with it. We will have to deal with this issue of what happens when it breaks down. We’re not going to stop people again. You can’t stop people. So let’s resurrect the robots so that when one of them goes, you can suck everything down into another one just like your phone.

A.5.7 Question 6

Catherine Pelachaud: Shall we go for the last question: How shall we deal with the popular fear of robots attacking the world? World meaning our jobs or not control politics, police.

Kerstin Dautenhahn: Robots that take over the world and wipe out humanity, that’s certainly something you could only laugh at. But what you mentioned robots taking jobs, this is a real fear and needs to be considered really, really seriously. Whenever new technology has been used on an industrial scale, it certainly led to changes, whether fewer jobs or different jobs, because then suddenly people were not doing what they have done for many, many years. They were replaced by machines, not necessarily by robots. First it was just mechanical machines powered by steam engines. They similarly also took over jobs that previously people did manually. Now, there’s a lot of discussion in the field: will robots also have a big impact? I’m actually more concerned about AI personally. I think AI takes more jobs than robots actually do. When it comes to actual agents, either virtual or physical, it’s certainly something we need investigate. We need to look at what the system should actually do, what we should promote. Is it about replacing what people are doing or is it
about providing tools that people could use? In my own work, for example, I emphasize very much when it comes to, let’s say, therapeutic applications that we are not trying to replace a therapist or a teacher or a parent. We are providing tools that in the hands of the people could be used in a therapeutic context, and I’m always very deliberate about that. Whenever one of my students has an idea: “maybe the robot could do what the therapist would do.” I say: “Well, we have to be careful here.” But, of course, that’s my personal choice. There are already lots of discussions also on robots and automation. We clearly need to keep this conversation alive and also involve and inform the public. I’m more concerned about media stunts with robots, like with Sophia. Someone just sent me a link to this interview that she gave. I looked at it and I thought, this is literally not possible. This must be fake. Then I looked into it. And yes, it was scripted. Of course [laugh] it was scripted. But they didn’t tell people. You had to specifically look for some information or talk to people who actually know that robot. I knew when I looked at the interview this cannot be real because I’ve seen the same thing happening with Pepper. I was on a panel at some point in the UK where Pepper was introducing the moderator of the panel; they started making jokes back and forth. But in the background, there was someone from Aldebaran sitting there, desperately typing, typing away so that Pepper could give the appearance of having a really meaningful, complex, real-time dialogue. But the coordinator of the panel was very responsible. He disclosed that to the audience. He didn't let the audience leave thinking that Pepper actually managed this super-interesting dialogue. No, he said: “Look, actually, there is a young man in the back basically puppeteering the robot.” I'm more concerned about people's opinion, the public's opinion on robots which is so much influenced by media stunts. It always makes me quite frustrated, actually. And I don't know why people are doing this. Well, I can guess why people are doing it, but I'm not very happy with that.

Birgit Lugrin: Movies are probably contributing to that as well, right?

Ruth Aylett: Yes, movies also contribute. If you ask people what their actual experience of robots has been, movies are one of the big components of that actual experience. There are also the Boston Dynamics videos, which are also largely tele-operated or scripted, and had a big influence too. People should know better, including the stunts that Kerstin mentions. So one of the 10 principles that the researchers in the UK came out with, that I mentioned earlier, was to have the moral duty to correct wrong statements in the press. I’m afraid we do have to do this. We don’t because we would rather research; we just
sit there and fume at the stuff that comes out, which we know to be absolutely wrong. But we do actually have to start responding and saying that it’s wrong. I’ve done this quite sharply in some cases, for example, with AIDA.\footnote{https://www.ai-darobot.com/.} This is supposed to be an artistic robot. It has an AI paint program attached to it with some graphics processing. Anyway, I won’t go on about it, but I wrote a piece on Medium about it. I also wrote a very sharp piece on a blog from someone in the poetry field. AIDA had a poetry generation program attached to it, so “she now writes poetry” (ironic voice). Hmm. I was fairly shocked about that. I wrote a book with a colleague specifically to counter this stuff, a popular science book we’ve just brought out.\footnote{R. Aylett and P. A. Vargas. 2021. Living with Robots: What Every Anxious Human Needs to Know. MIT Press.} But I will tell you that that is never going to get mass circulation because everyone prefers the alternative story. It’s much more fun. People don’t particularly want to have their illusions dispelled. They would prefer to believe the hype; it’s interesting. It’s fascinating. Maybe a bit scary, the same way that zombies are, that robots will take over the world. There are cultural reasons for this, at least in Western societies. So it’s going to be very tough. The reason you get the stunts is because people like them, and because it makes people money. No newspaper ever lost money by printing a story about robots taking over the world. People want to hear these stories. They do not want to hear that robots will not take over the world and are really quite clumsy and not very, very useful at the moment. That’s not much fun as a story. Why would we tell people that? That’s not going to sell newspapers. This is technology hype. Yet we have an active duty to pursue it rather than just sit there and view it. On the automation thing, I think Kirsten’s really made the important point here. This is part of an overall process of automation that started in the 1750s in Western Europe, which produced a severe trauma for the populations that underwent it as their old society was torn to ribbons and they were herded into insanitary cities and chained to factory machines. And that trauma echoes down the centuries since. Not for nothing did people try and break the machines at the time. Robots are just another element in this story as far as people are concerned, and that’s why they fear them. Should they fear them? Well, automation is a continuing process. I’m not even sure I agree with Kirsten that AI is more of a problem. There’s not an awful lot of AI in most of the systems at the moment, minimal amounts of AI. The problem is often, as someone in Edinburgh told me, that these systems are not very intelligent.
In fact, they're very stupid. But what you've got is hype, which tells everyone who uses this stuff it is really cool, it's really infallible, you can believe in it and you should use it. That is very dangerous indeed. Much of the stuff about robots, incidentally, is not about robots. I've been following some of these stories. They do not mean robot when they say robot, as we would understand the term. They mean the Internet, actually. They don't distinguish. What you're talking about is information automation, which is the current wave of automation. So in automating aspects of information processing on a geographically extended scale that we couldn't do before we have the Internet. But the hype is serious. Not primarily because they're knocking out jobs, but because they're overestimating the capabilities of what they can do, and they're excluding human judgment from processes which should not have human judgment excluded from them. So the computer says “no” written large everywhere, in very sensitive discussions. For example, the system in the US for sentencing, which helps judges decide whether a particular individual is likely to be a recidivist, break the law again. The sentencing system is trained on a database, which is completely skewed by the high proportion of black people in the US that have already been convicted of crimes because the justice system is racist, basically. So, of course, it's biased when it makes decisions. If it's a black guy, it will tell them that his likelihood of breaking the law again is high because it's using biased data, but “oh, the computer must be right. This is an intelligent system” (ironic voice). So the judge will take its advice. The problem you have here is not necessarily the systems. It's the illusions that people have about these systems, which are very deep indeed. For facial recognition, a UK police force used it in a football stadium of people and they got a 4% error rate in identifying people on their database of criminals. Well, they were lucky it wasn't higher than that is all I can say. Because if it's 95% accurate under good lighting conditions and you've got 10,000 people or 50,000 people, then you're going to get very high error rates. “But it's a technology that works,” they've been told. So I think our problems are there.

Kerstin Dautenhahn: When I said I'm more afraid of AI, I didn't mean AI in robots. I meant the AI that is used for example in law enforcement surveillance, the AI that's running on my phone that I might not even know about. But that AI knows a lot about me. Although I try proactively to switch off as many features that I can. But I'm sure I overlook many of those. This is more what I'm afraid of. This is more what is taking over the world. Surveillance cameras everywhere knowing our every single move; also initiatives to
build smart cities. Here in Ontario, before I moved to Waterloo three years ago, they were in the process of approving a smart city within Toronto with, basically, 24/7 continuous surveillance. I was very happy, when two years ago, they actually scrapped those plans. There was very strong opposition to
collecting everything from within your house, outside, on the street, where
you go, what you shop, who you are, etcetera. But in some other countries
and other places, people already have that. They have very, very high surveil-
lance. These are more the things that I’m concerned about and not so much
whether robots will take over the world.

Ruth Aylett: I agree.

A.6

Interview 5: Ethics in the Application of SIA for Children with
Autistic Spectrum Disorders

We have organized a specific interview to discuss ethical issues that arise when
modeling SIAs interacting with children with Autistic Spectrum Disorders. It
happened in December 2022. We discussed the following questions:

Question 1: What are the ethical issues related to technological development, com-
mitments announced to family members (that are not maintained due to the com-
plexity of computational development), and institutional use of socially interactive
agents?

Question 2: How should we draw the line between persuasion and manipulation
that we sometimes need to get certain effects that are desired, and transparency in
ASD-related applications?

Question 3: Should SIAs be better than humans?

Question 4: What does it involve for SIAs to be better than humans?

A.6.1 Participant

Jacqueline Nadel, emeritus CNRS Research Director at La Salpêtrière Hospital,
Psychiatry Department

A.6.2 Question 1

Catherine Pelachaud: The first question we would like to ask you is: What are
the ethical issues related to technological development, commitments
announced to family members (that are not maintained due to the com-
plexity of computational development), and institutional use of socially
interactive agents?
Jacqueline Nadel: I would like to say first that for an institutional use of socially interactive agents, you should obtain the permission of an ethics committee. The ethics committee will ask questions about physical security (for instance, for robots they cannot break into small pieces that can be swallowed or eaten; temperature should not be too hot or too cold; electric elements should be secured; the SIA should not fall easily, ...). I am thinking of NAO, for example. NAO falls easily. Also, the psychological aspects related to dependency and addiction should be considered. Questionnaires to families and experts as well as results of pilot experiments can be asked by an ethical committee to document these points. The ethics committee will ask for detailed explanations about the objectives and procedures that should be offered to the users or their families if they cannot decide by themselves so that the decision is taken with full awareness. The family should be informed of the benefits and shortcomings of SIAs. There should be no illusion. You should say to the parent that it is not magic. The use of social agents will not change totally the specificity of the person. It is important to immediately notice the regularity of use, so as to develop good routines in the use of SIAs. Specially for virtual reality, the ethics committee will ask questions about the feeling of presence. It is a very important factor to consider as it may break the line between perceiving an object as real or as virtual. A special problem is now about mixed reality when you have a virtual object projected in a real room. This is a very big problem for people with autism who have not developed false belief. In this case, it is very difficult for them to tell the difference between real object and virtual object in the room. They are afraid of the situation because they don’t know what the virtual object is doing in the room they know. So this is a very important and novel point for the ethics committee to take into account. For the family, the most important element for me to consider is for the parents to be free to meet the SIA and start interacting with it or to observe their child interacting with the agent. This is the best. When this is done, usually the parents feel confident with the situation, the objectives of the interaction with the SIA and with how their child may behave, with the way their child will behave. It is a very important element for the ethics committee that the person, be parents or experts for example, can see the design and be aware of the different elements of the design. I think that if this is done there are no more ethical issues that are worth to be developed.

Maybe I should speak about the fact that usually the ethics committee asks that experts in the domain of autism and also associations of parents of children with autism have already seen the design of the SIA.
If I look at what we did recently where we built a virtual platform for children with autism with a collaborative agent, we had the ethics committee ask us questions: did experts in the field see the material? Is the material secure? How did you use the material? And also, we started with neurotypical children to see if there was a problem that people can explain. This is part of the ethical issues but it is also part of the design of the research itself. So it is not easy to distinguish what is ethical and what is the research itself.

If you would like me to extend more, please tell me.

A.6.3 Question 2

Birgit Lugrin: Thank you, that was very informative. Maybe you can elaborate a bit more on how we should draw the line between persuasion and manipulation that we sometimes need to get certain effects that are desired, and transparency in ASD-related applications?

Jacqueline Nadel: For me, the best is of course transparency. The parents have to be aware of what is done during the experiment and when the children have a good cognitive level, they should also be made aware. This question regarding persuasion and manipulation has become so important that now if the participants are not aware of what is happening during the experiment there will be no agreement by the ethics committee. And then, you can also consider that, at least for high functioning people with ASD, they appreciate honesty; they appreciate people to be simple, to be honest, to be directly asking questions to them, and to be directly explaining what they will do during the experiment. Transparency is a need because people with autism don’t know how to lie. They are not liars at all. They don’t appreciate it if people are not directly honest with them. You can try to persuade them that something is good for them providing you really think it is good.

The difference between persuasion and manipulation can be a subtle one and remain implicit in the situation. It is especially true with nonverbal persons. But manipulation can also be a deliberate strategy. I am thinking of social psychology. You can sometimes have participants that do not know what the objective of the research is. They will have been informed about something that is not the real objective. As far as I know this is really not possible anymore. I know of a lot of research in the field that will not be possible to conduct now. They may have been done 10 or 15 years ago. Now you can no longer conduct this kind of research. But it remains that for people with low performance and people that are nonverbal, the difference between manipulation and persuasion is a very difficult question. What is often observed
is that people with autism have real difficulties in choosing, making a decision, taking an initiative. If you start persuading a person with autism that something is good for them, they can be encouraged to do so; and so, persuasion will have a good effect in the situation. But manipulation is something else. You manipulate when you propose a procedure without indicating that you can do something else. You propose how to behave and when to behave. Many instructed programs do manipulate insofar as there is nothing else to do than to follow the instructions. So, the person is not free to say “no” or to do something else but has to follow the instructions. But now manipulation will be prevented thanks to the investigation of the ethics committee. Preventing the person from being manipulated is really a rule that the ethics committee will have in mind.

A.6.4 Question 3

Catherine Pelachaud: You have already answered our third question which is should SIA be better than humans. But “better than human” is somehow a manipulation.

Jacqueline Nadel: Absolutely, your question, of course, is perfectly valid for human partners, and even maybe more. Sometimes we can measure better if a system is manipulating a child than if a human being is manipulating a child. For instance, if I take the example of turn-taking. Turn-taking is a very good parameter to measure reciprocity and to measure the involvement of the user with the system. You can reorganize the system if you see that the child does not take her/his turn, that the social agent is always taking the initiative in the situation, always initiating things that the user follows. I think there are many implicit aspects in play in a situation of dependency or addiction. With a social agent you can more easily find the solution than with humans. For instance, a program can be developed in order to stop positive feedbacks between the social agent and the person. Imagine, the person may be asking the same question over and over to the social agent, or the person is imitating repeatedly the behaviour of the agent. You can stop these positive feedbacks. You can consider the interactive mode between the social agent and the individual with autism. This is not true for human partners. It is very difficult to stop a stereotypical behavior when it appears during an interaction between a human partner and the child with autism because it will break not only the interaction but also the relationship.

I think the danger of dependency and addiction is potentially less important for SIA than for human partners. Regarding addiction, you can have an
a priori agreement with the person with autism concerning the duration of use of the SIA, for instance. If you have an agreement for a timer, then you will have an interaction with the social agent for 10min and no more. So, you use a timer and everything will stop after 10min. You will avoid dependency because you will have, at first, limited the time when the user will be in the presence of the social partner. This is something you can do with the social agent that is far more difficult to do with a human partner. Thus with a SIA you are in a better situation than with a human partner to manage the question of dependency.

Now for addiction it is something different because it depends largely on the objectives of the program. If the program is a short-term program, there is not much of a problem of addiction. If you have a six-week program where each week the child will meet the social agent two times, there is no real danger of addiction. But if it is a long-term program, for example running for one or two years, you would like the child to develop social skills, then in this situation, there could be a big problem of dependency and addiction. For long-term developmental programs, addiction may appear progressively. I would say that addiction is maybe a special problem with social agents because they are better than humans.

A.6.5 Question 4

Birgit Lugrin: This nicely leads us to our next question. You are saying SIAs can be better than humans. So, what does it involve for SIAs to be better than humans?

Jacqueline Nadel: To some extent they are better. They are always ready to welcome, never in a bad mood, never in a hurry to receive an answer, they don’t look in the eyes, they are less sophisticated, they are more predictable. These fit particularly well the specificities of people with autism. It is an enormous advantage of social agents compared to humans. So, I would say that these capacities of the animated social agents are a very good way to allow children with autism to accept social situations. A lot of times, social situations are very difficult for people with autism because we are so different from one moment to another. Our eyes are always moving. Our facial expressions are always changing. It is something that makes us unpredictable for children with autism.

All the basic affiliative behaviors will be easier to learn with a SIA rather than directly with a human partner. The virtual partner will have an enormous advantage compared to a human being. Of course, you should progressively
decrease these specificities of the social agent in order to make their behaviors more similar to human behaviors. Maybe that is really the problem. You start with a social agent that has these wonderful advantages to be more predictable, to be always in a good mood, to be always ready to welcome; that perfectly fits the specificities of the child with autism. But progressively the child with autism has to adapt to human specificities. So, they have to adapt to unpredictability. That may be something to think about: how do you make your social agent less and less virtual and more and more human-like in the way it interacts with the child? That is, the agent becomes less predictable, becomes less happy to meet the child, becomes sometimes joyful, sometimes neutral; and this according to what the child has to understand about the social agent. Maybe this is the problem. But a good social agent, happy to welcome the child, is the best we can offer to the child with autism at the beginning of a social training. I do not see any danger with that. It is a good way to start a social training program. Afterward, the big problem will be about dependency. It is especially an important matter if the social agent takes initiatives and directs to some extent the individual's choices. In such a case, an agreement should always be obtained where the SIA should first ask the person to decide what to do and how. For people with no verbal language, the SIA should understand their gestures and facial emotional expressions so it can adapt to their personal tempo, their personal rhythm. Some are speedy. The majority of people with autism, especially nonverbal persons, are slow in their answer. If you don't wait you will take the initiative in their place. Thus, the social agent may interact in a way that takes into account the rhythm, the tempo of the user. You see what I mean.

Birgit Lugrin: Yes, definitively.

**A.6.6 After the Questions: Free Comment**

Jacqueline Nadel: Your clever questions are all part of a solipsistic view of the individual. As soon as we consider the person with ASD as part of a dynamic system relating them to others, all the answers can be modulated by the nature of the dynamics: the real problem of ethics lies in the fact in recognizing the individual with ASD as equal to you in a no-hierarchy conception of human rights.

**A.7 Concluding Remarks**

In this challenge discussion chapter, we have seen thought-provoking and critical discussions on the current challenges in research and development of SIAs. These
challenges covered both technical challenges and societal or ethical challenges. Although extensive research and development in the fields of SIAs have drastically advanced the state of the art in the last two decades, there is still a long way to go before we will achieve agents that can truly socially interact whilst being of practical use for people in their intended social domain. This is particularly prominent when we are looking at interactions in the wild and over longer periods of time. But we have also seen that there is lots of room for theoretical research in the lab to completely understand the underlying mechanics of social interaction with artificial entities. With this handbook, we aimed to bridge the gap between the two communities of IVAs and SRs. We have seen in each chapter of this handbook that there are very common research directions, ideas, challenges, and approaches. The challenge discussions particularly highlighted the need for and great benefits of the two communities working together and looking at each other’s implementations and research findings.

However, all the works presented in this handbook have also shown that the research conducted by this community is of great interest for (and can largely benefit from input of) other domains such as, for example, virtual/augmented reality, affective computing, game design, computer animation, or Kansei. Studies have demonstrated how users attribute communicative and emotional intention to autonomous agents with abstract figures, not only with human-like appearance; and that those results can apply also to autonomous entities such as voice assistants, conversational agents, assistive robots, but, why not also to autonomous cars. Theoretical and computational models on emotion, cognition, but also behaviors, speech, social space to name some chapters, can be of use not only to model and build embodied agents but also other autonomous entities. The communicative and emotional functions may be common to many of these entities; while their instantiation into behaviors will depend on their embodiment (voice, text, object, etcetera) and context of use. These areas are often also located in socially interactive domains and thus address similar psychological questions as well as technological ones. Furthermore, they will also be out for interaction with humans in the wild in the future and can thus benefit from the research findings presented in this book.

We are thus positive about our endeavor to bring the communities of IVAs and SRs closer together and are inviting other communities to join our journey!

References

References


Authors’ Biographies

Part I. Editors
Birgit Lugrin

Birgit Lugrin (birth name Birgit Endrass) is a professor for media informatics at the University of Würzburg, Germany. Since her first contact with a socially interactive agent (the Greta agent) in 2003, she has been fascinated about the research area. Ten years later, she received the prestigious IFAAMAS Victor Lesser Distinguished Dissertation Award, and the research award of Augsburg University for her research on enculturated virtual agents. Another ten years later, she still works with socially interactive agents in different application areas such as education, and could not be happier about the chance to co-edit this handbook and work with all the great researchers who have contributed to make this happen.

Catherine Pelachaud

Catherine Pelachaud is Director of Research at CNRS in the laboratory ISIR, Sorbonne University. She received her Ph.D. in computer graphics at the University of Pennsylvania, Philadelphia, in 1991. She participated in the elaboration of the first embodied conversation agent system, GestureJack, with Justine Cassell, Norman Badler, and Mark Steedman when she was a post-doctorate at the University of Pennsylvania. With her research team, she has been developing an interactive virtual agent platform, Greta, that can display emotional and communicative behaviors.
David Traum

David Traum is Director for Natural Language Research at the Institute for Creative Technologies (ICT) and Research Professor in the Department of Computer Science at the University of Southern California. He leads the Natural Language Dialogue Group at ICT. Traum’s research focuses on dialogue communication between humans and artificial agents. He has engaged in theoretical, implementational, and empirical approaches to the problem, studying human–human natural language and multimodal dialogue, as well as building a number of dialogue systems to communicate with human users. Traum earned his Ph.D. in computer science at the University of Rochester in 1994.

Part II. Chapter Authors

Elisabeth André

Elisabeth André is a full professor of computer science and founding chair of Human-Centered Multimedia at Augsburg University in Germany. Already a quarter of a century ago, she pioneered the concept of embodied conversational agents as a new metaphor for human–machine interaction and organized one of the first workshops on animated interface agents in conjunction with IJCAI in Nagoya, Japan. She has undertaken very successful interdisciplinary collaborations with psychologists, pedagogues, and media artists that have resulted in several award-winning multimodal user interfaces, among them pedagogically well-grounded and empirically validated learning environments for children and young people.
Ruth Aylett

Ruth Aylett is professor of computer science at Heriot-Watt University in Edinburgh. She originated the idea of emergent narrative in 1999 and has written widely on interactive digital narrative, both from a theoretical and practical perspective. She has worked on affective agent architectures capable of supporting characters in narrative worlds and was involved in a series of EU projects from 2001 developing narratively driven systems.

Timothy Bickmore

Timothy Bickmore is a professor in the Khoury College of Computer Sciences at Northeastern University. The focus of his research is on the development and evaluation of computer agents that emulate face-to-face interactions between health providers and patients for use in health education and long-term health behavior change interventions, with a particular focus on the emotional and relational aspects of these interactions. Prior to Northeastern, he spent two years as an assistant professor of medicine at the Boston University School of Medicine. Dr. Bickmore received his Ph.D. from the MIT Media Lab.

Jennifer Boger

Jennifer Boger is an adjunct professor at the University of Waterloo. She is passionate about creating technologies that enable more inclusive, equitable, and personal support for health, wellbeing, and quality of life for older adults. Her research focuses on transdisciplinary collaboration and user-driven design practices that blend state-of-the-art knowledge from computer science, engineering, and health sciences to create internationally renowned intelligent assistive technologies for supporting graceful aging. Jennifer believes that incorporating
multiple viewpoints is key in achieving novel, holistic, and implementable solutions to the large, complex problems surrounding aging and technology.

**Cynthia Breazeal**

Cynthia Breazeal is a professor and an associate director at the MIT Media Lab and is the founding director of the Personal Robots Group. She also leads MIT’s K–12 AI Literacy project. She is recognized as a pioneer in the field of social robotics and human–robot interaction (HRI). Her group investigates long-term HRI in schools, homes, hospitals, and so on, where the robot personalizes and builds relationships with people to help them achieve personal goals. She is an AAAI Fellow and recipient of awards from the National Academy of Engineering and National Design Awards. She received her doctorate from MIT in electrical engineering and computer science in 2000.

**Joost Broekens**

Joost Broekens is president-elect of the Association for the Advancement of Affective Computing (AAAC). He is associate professor of Affective Computing and Human Robot Interaction at the Leiden Institute of Advanced Computer Science (LIACS) of Leiden University. He heads the [a]social creatures lab, which focuses on understanding social interaction with and between artificial creatures. He is co-founder and CTO of Interactive Robotics. His expertise is in human–agent and human–robot interaction and the computational modeling of emotion.
Justine Cassell

Justine Cassell is the Dean’s Professor of Language Technologies in the School of Computer Science at Carnegie Mellon University. She also holds a chair at the PRAIRIE AI Institute in Paris. Previously, Cassell founded the Technology and Social Behavior Program at Northwestern University and was a tenured professor at MIT. Cassell has received the MIT Edgerton Prize, Anita Borg Women of Vision Award, AAMAS Test of Time Award, and a National Academy of Sciences Prize for Behavioral Science applicable to policy. She is a fellow of the AAAS, Royal Academy of Scotland, and the ACM.

Kerstin Dautenhahn

Kerstin Dautenhahn, IEEE Fellow, is Canada 150 Research Chair in Intelligent Robotics at the University of Waterloo in Ontario, Canada. She has a joint appointment with the Departments of Electrical and Computer Engineering and Systems Design Engineering and is cross-appointed with the David R. Cheriton School of Computer Science at the University of Waterloo. She is director of the Social and Intelligent Robotics Research Laboratory (SIRRL). Her main research areas are human–robot interaction, social robotics, as well as developmental and cognitive robotics. She has published more than 400 peer-reviewed articles (H-index 85) and has several senior editorial roles in international journals.
Patrick Gebhard

Patrick Gebhard is the head of the Affective Computing Group at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken, Germany. He received his doctoral degree, Dr.-Ing, from Saarland University in 2007. With his team, he researches the adaptation of psychological theories for the simulation of human emotions and social behavior to enable a human-like interaction with computer systems and evaluate socio-emotional interactive human–agent systems. His group developed the internationally recognized research frameworks ALMA for computationally simulating affect and VisualSceneMaker to model and execute interactive behavior for socially interactive agents.

Moojan Ghafurian

Moojan Ghafurian is a research assistant professor at the Department of Electrical and Computer Engineering at the University of Waterloo. She got her Ph.D. from the Pennsylvania State University and was the Inaugural Wes Graham postdoctoral fellow from 2018 to 2020 at the David R. Cheriton School of Computer Science at the University of Waterloo. Her research areas are human–computer/robot interaction, artificial intelligence, social robotics, affective computing, and cognitive science. Her research explores computational models of how humans interact with computers to inform user-centered design of intelligent systems in multiple domains, especially in healthcare and for supporting older adults, persons with dementia, and caregivers.
Sarah Gillet

Sarah Gillet is a Ph.D. student at the Division of Robotics, Perception and Learning at KTH, Royal Institute of Technology. She received her B.Sc. and M.Sc. in computer science from TU Dortmund, Germany. Her research goal is to use social robots to support and improve human group interactions. In her recent works, she explored how a social robot in a multiparty setting could help the inclusion of outgroup children.

Jonathan Gratch

Jonathan Gratch is a research full professor of computer science and psychology at the University of Southern California (USC) and director for virtual human research at USC’s Institute for Creative Technologies. He completed his PhD in Computer Science at the University of Illinois in Urbana-Champaign in 1995. Dr. Gratch's research focuses on computational models of human cognitive and social processes, especially emotion, and explores these models' role in advancing psychological theory and in shaping human–machine interaction. He is the founding editor-in-chief (retired) of IEEE's Transactions on Affective Computing, founding associate editor of Affective Science, associate editor of Emotion Review, and the Journal of Autonomous Agents and Multiagent Systems.
Ishaan Grover

Ishaan Grover is a second year Ph.D. student in the Personal Robots Group at the MIT Media Lab. His research interests are in machine learning and its applications in the field of human–robot interaction and early literacy education. His recent work focused on building graphical models using word semantics for predicting children’s knowledge from partial information about their existing vocabulary. Prior to this work, he worked on algorithms to enable reinforcement learning agents to learn from human feedback. In industry, he has interned at Apple, Palantir, and Microsoft where he worked on applications ranging from natural language understanding to anomaly detection.

Ouriel Grynszpan

Ouriel Grynszpan is a professor at the University Paris-Saclay and is affiliated to LISN (lisn.upsaclay.fr). His research is at the crossroads of psychology, affective neuroscience, and human–computer communication. Grynszpan has led several projects on digital technologies for autism. His research is published in psychology and human–computer interaction journals.
Arno Hartholt

Arno Hartholt is the Director of R&D Integration at the University of Southern California Institute for Creative Technologies where he leads the virtual human integration and core technology group. He is responsible for much of the technology, art, and processes related to virtual humans and related systems, in particular at the intersection of research and industry. He has a leading role on a wide variety of research prototypes and applications, ranging from medical education to military training and treatment, including the Virtual Human Toolkit, VITA, SimSensei, and BRAVEMIND. He received his M.Sc. in Computer Science at the University of Twente.

Teena Hassan

Teena Hassan is an AI and robotics researcher who is currently doing her postdoctoral research at the University of Bremen in the Robotics Group. She received her doctoral degree (Dr. rer. nat.) in 2020 from the University of Bamberg for her dissertation in the area of automatic mental state analysis based on facial expressions. Previously, she served as a research associate at the Bielefeld University and as a senior scientist at the Fraunhofer Institute for Integrated Circuits IIS in Erlangen. Her research interests include affective computing, human-robot interaction, and human-centered reinforcement learning.
Jesse Hoey

Dr. Jesse Hoey is a professor in the David R. Cheriton School of Computer Science at the University of Waterloo, where he leads the Computational Health Informatics Laboratory (CHIL). He is a Faculty Affiliate at the Vector Institute, and an affiliate scientist at KITE/TRI, both in Toronto. Dr. Hoey holds a Ph.D degree (2004) in computer science from the University of British Columbia. He has published over one hundred peer-reviewed scientific papers. He is an associate editor for the IEEE Transactions on Affective Computing and an area chair for the International Joint Conferences on Artificial Intelligence (IJCAI).

Kathrin Janowski

Kathrin Janowski is a doctoral researcher at the chair of Human-Centered Multimedia of the University of Augsburg. Her work focuses on multimodal dialogue with social robots and virtual characters, in particular the generation of nonverbal behaviors based on the agent’s personality or emotional state.
Stefan Kopp

Stefan Kopp is professor of computer science and head of the Social Cognitive Systems Group at the Faculty of Technology and the Research Center on “Cognitive Interaction Technology” (CITEC) at Bielefeld University, Germany. He obtained his Ph.D. from Bielefeld University, was postdoctoral researcher at Northwestern U, and fellow at the Center for Interdisciplinary Research (ZiF). His research centers around understanding the cognitive mechanisms of social interaction, multimodal communication or cooperation, and leveraging AI methods to endow intelligent technical agents with similar abilities. Stefan has co-authored more than 300 scientific publications and is member of several editorial/steering boards of international journals and conferences.

Jacqueline M. Kory-Westlund

Jacqueline M. Kory-Westlund is an independent scholar with the Ronin Institute, writer, and artist. She earned her Ph.D. in Media Arts and Sciences from the MIT Media Lab in 2019. Her research has focused on how we can use social robots to support and engage young children in learning language skills, including how children construe social robots, how children’s relationships connect to their learning (especially over long-term interactions), and the ethics of using robots in children’s lives. She is the author of a forthcoming book from Columbia University Press on thriving in graduate school while keeping a healthy personal life.
**H. Chad Lane**

**H. Chad Lane**, Ph.D., is an associate professor of educational psychology and computer science at the University of Illinois, Urbana-Champaign. His research focuses on the design, use, and impacts of intelligent technologies for learning primarily in informal settings. This work involves blending techniques from the entertainment industry (that foster engagement) with those from artificial intelligence and intelligent tutoring systems (that promote learning).

**Iolanda Leite**

**Iolanda Leite** is an associate professor at the School of Electrical Engineering and Computer Science at KTH Royal Institute of Technology. She holds a Ph.D. in information systems and computer engineering from IST, University of Lisbon. Prior to joining KTH, she had postdoctoral appointments at Yale University and at Disney Research. Iolanda’s research interests are in the areas of artificial intelligence and human–robot interaction. She is the co-editor in chief of *AI Matters*, the newsletter of the ACM Special Interest Group in Artificial Intelligence (SIGAI).

**Jean-Claude Martin**

**Jean-Claude Martin** is a professor of computer science and human–computer interaction at Université Paris-Saclay. He conducts his research at LISN (Laboratoire Interdisciplinaire des Sciences du Numérique), France, where he heads a multidisciplinary group of psychology and computer science researchers. He is the editor-in-chief of Springer’s *Journal on Multimodal User Interfaces* (JMUI). He works on social skills training (autism, job interviews, virtual patients, leadership) and personalized coaching (sports and e-health, wearable interaction, stress management).
Sharon Mozgai

Sharon Mozgai is the Associate Director of MedicalVR at the University of Southern California's Institute for Creative Technologies (ICT). Sharon received her master's at Harvard University in human development and psychology. At ICT, she contributed to The Virtual Human Project to create highly realistic and compelling training environments under the direction of Jonathan Gratch. She also supported SimSensei, a clinical virtual human platform built to sense real-time audio-visual signals, led by Skip Rizzo. Currently, she leads the development of an interactive virtual human system, Battle Buddy, focusing on long-term engagement to motivate and sustain behavior change.

John Muñoz

John Muñoz is a game designer and interface technologist specializing in the use of physiological signals to optimize the user experience while using interactive systems. John is currently a postdoctoral fellow at the University of Waterloo carrying out research in the fields of assistive technology, human-robot interaction, and virtual reality. John has co-designed multiple games that can be used in healthcare scenarios such as exercise promotion in older adults, self-regulation training in children with special needs, and cognitive training in neurorehabilitation therapies.
Jacqueline Nadel

Jacqueline Nadel is an emeritus CNRS Research Director at La Salpêtrière Hospital, Psychiatry Department. She is the author of 170 peer-reviewed papers and of several books at Cambridge University Press and Oxford University Press. She is involved in interdisciplinary programs interfacing social neuroscience, cognitive psychology, epigenetic robotics, computer science, and clinical interventions for severe autism. She is the scientific director of two education centers for children with ASD. Additionally, she edits the French scientific journal *ENFANCE*.

Fabrizio Nunnari

Fabrizio Nunnari is a senior researcher at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken, Germany. He entered DFKI in 2012 as a member of the Sign Language Synthesis and Interaction (SLSI) group, where he worked on the use of natural user interfaces for the authoring of sign language animations. Also, he now works on the connection between the aesthetics of virtual characters and the perception of their personality and on developing the YALLAH framework for the generation of real-time interactive virtual humans. He has been part of the Affective Computing Group since 2020.

Ana Paiva

Ana Paiva is a professor of computer science at Instituto Superior Técnico, University of Lisbon, investigating the creation of intelligent interactive systems by designing “social agents” that can interact with humans in a natural and social manner. She is also a fellow at the Radcliffe Institute for Advanced Study at Harvard University. Over the years, she has addressed this problem by engineering social agents that exhibit specific capabilities, including
emotions, personality, culture, nonverbal behavior, empathy, and collaboration, among others. Her more recent research combines methods from artificial intelligence with social modeling to study hybrid societies of humans and machines. In particular, she is investigating how to engineer agents that lead to more prosocial and altruistic societies. She has published extensively and received best paper awards in several conferences, notably, she won the Blue Sky Awards at the AAAI in 2018. She has further advanced the area of artificial intelligence and social agents worldwide, having served for the Global Agenda Council in Artificial Intelligence and Robotics of the World Economic Forum and as a member of the Scientific Advisory Board of Science Europe. She is a EurAI fellow.

Hae Won Park

Hae Won Park is a research scientist at the MIT Media Lab and a Principal Investigator of the Social Robot Companions Program. Her research pioneers the field of socio-emotive AI, embodied conversational agents, and long-term personalization of social robots. Her work spans from education for young children to wellbeing support for older adults. Her research has been published and recognized as best paper at top robotics and AI venues. She received her Ph.D. from Georgia Tech in 2014, during which she co-founded Zyrobotics, an assistive education robotics startup that was recognized as the best 2015 US robotics startup by Robohub.

Christopher Peters

Christopher Peters is an associate professor of computer graphics and HCI at the KTH Royal Institute of Technology, Sweden, where he leads the Embodied Social Agents Lab (ESAL). He received his Ph.D. in computer science at Trinity College Dublin, Ireland, in 2004. His research interests surround the generation and human perception of artificial behaviors in embodied humanoid agents, groups, and crowds.
Rui Prada

**Rui Prada** is an associate professor at Instituto Superior Técnico (IST), Universidade de Lisboa, and a senior researcher at INESC-ID in the AI for People and Society research group. He received a Ph.D. in computer science in the domain of artificial intelligence at IST in 2005. He was co-responsible for the creation of courses on game design and development at IST. He conducts research on social intelligent agents, affective computing, human–agent interaction, computer games, applied gaming, and game AI. With his team, he has developed virtual agents with socio-emotional behavior for games and robots that play games with people.

Diogo Rato

**Diogo Rato** is a Ph.D. candidate at Instituto Superior Técnico, Universidade de Lisboa, and conducts research at INESC-ID in the AI for People and Society research group. His research work focuses on designing socio-cognitive models that apply the concepts of social identity to interactive agents. His research interests also include studying human–agent interaction using digital games and deploying groups of social non-player characters in virtual worlds.

Hannes Ritschel

**Hannes Ritschel** studied computer science. He is research scientist at Augsburg University (Germany). His interests include reinforcement learning in the context of social robots in order to autonomously adapt their behaviors to the individual user in real time.
Tanja Schneeberger

Tanja Schneeberger is a researcher in the Affective Computing Group at the German Research Center for Artificial Intelligence (DFKI), in Saarbrücken, Germany, since 2015. With her background in psychology, she applies psychological theories to create human-like interactions with socially interactive agents. In experimental studies in which she transfers aspects of human–human-interaction to human–agent-interaction, her goal is to find out which role socially interactive agents can play for humans and how to create a successful relationship between a socially interactive agent and a human.

Noah L. Schroeder

Noah L. Schroeder, Ph.D., is a professor of educational technology and instructional design at Wright State University in Dayton, Ohio, USA. His research is focused on the design of multimedia learning systems, with much of his work focused on the design of pedagogical agents.

Dimitra Tsovaltzi

Dimitra Tsovaltzi is a senior researcher in the Affective Computing Group at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken, Germany, and a habilitated member of the faculty of Human and Business Sciences at Saarland University. She has been working on designing technologies to support the interplay of socio-emotional and cognitive factors for learning and group interaction, focusing on awareness tools, and possibilities for dynamic scripting. Her recent research includes modeling and technologically supporting socio-emotional and cognitive interactions. She is
currently investigating the intertwining of emotion regulation and group interactions for implicit technology-based supports.

**Marynel Vázquez**

Marynel Vázquez is an Assistant Professor in Yale’s Computer Science Department, where she leads the Interactive Machines Group. Her research focuses on advancing multiparty human–robot interaction, both by studying social group phenomena and developing perception and decision-making algorithms that enable autonomous robot behavior. Marynel received her bachelor’s degree in computer engineering from Universidad Simón Bolívar in 2008 and obtained her M.S. and Ph.D. in robotics from Carnegie Mellon University in 2013 and 2017, respectively. Before joining Yale, she was a collaborator of Disney Research and a post-doctoral scholar at the Stanford AI Lab.

**Hannes Högni Vilhjálmsson**

Hannes Högni Vilhjálmsson is a professor of computer science at Reykjavik University, where he leads the Socially Expressive Computing group at the Center for Analysis and Design of Intelligent Agents. His passion in research is to create virtual experiences, built on solid theory, that affect people in meaningful ways. These are often social experiences with applications in training, education, therapy, and health. He has co-founded three virtual technology startups and helped coordinate several grassroot efforts to promote an interdisciplinary approach to technical innovation. He received his Ph.D. in media arts and sciences from MIT in 2003.
Agnieszka Wykowska

Agnieszka Wykowska is a principal investigator at the Italian Institute of Technology leading the unit “Social Cognition in Human–Robot Interaction” and is an adjunct professor of engineering psychology at the Luleå University of Technology. She graduated in neuro-cognitive psychology (2006, LMU Munich), obtained her PhD in psychology (2008) and her habilitation (2013) from LMU Munich. In 2016, she was awarded the ERC Starting grant “InStance.” She is editor-in-chief of the International Journal of Social Robotics, a board member of the European Society for Cognitive and Affective Neuroscience (ESCAN), and a member of the core faculty of European Lab for Learning and Intelligent Systems (ELLIS).

Fangkai Yang

Fangkai Yang is a Ph.D. student in computational science and technology at KTH Royal Institute of Technology. Working with the ESAL led by Prof. Christopher Peters, he has been working on human–robot/agent interaction on the group level. He received his master’s and bachelor’s degrees in applied mathematics, and he previously worked in game studios developing physics game engines.
Index

AAT. See Animal-assisted therapy (AAT)
AAVE. See African American English (AAVE)
Academic platforms, 265, 271. See also Commercial platforms
Greta, 271–272
honorable mentions, 274–275
VHToolkit, 273–274
Accountability, 569
ACL. See Association for Computational Linguistics (ACL)
ACT-R, 265
Action, 411
Action modalities, social robot architectures, 90–92
Active Analysis rehearsal technique (Stanislavsky), 480
Activity of daily living (ADL), 370–371
ADAPT, 275
Adaptability, 80
Adaptable social agents, 164
Adaptation, 165, 381
approaches, 156, 164–166
decision-theoretic reasoning, 166–167
neural learning of personality-based behavior styles, 170–171
reinforcement learning, 167–170
Adaptive artificial personalities
autonomous interactive systems, 179
evaluating adaptive systems, 171–173
finding right level of sensitivity, 177–178
fine-grained behavior timing, 177–179
implementing adaptive systems, 163–171
personality and interpersonal stance, 157–160
psychological background, 157
in situ studies, 176–177
theories about interpersonal compatibility, 160–163
Adaptive social agents, 164
Adele, 535
ADHD. See Attention Deficit Hyperactivity Disorder (ADHD)
ADL. See Activity of daily living (ADL)
Adobe Photoshop, 276
Adoption modeling, 384
Adviser, 503
AF. See Atrial fibrillation (AF)
Affective computing, 442
Affective Presentation Markup Language (APML), 263
African American English (AAVE), 345
Agency, 159
Agent Persona Instrument, 319
Agents, xxix, 372
academic platforms, 271–275
cognitive architectures, 265–266
commercial platforms, 266–270
platforms, 264
selection, 387
story, 224
Aging, 367
challenges, 384–389
considerations, 380–384
future directions, 389–391
history and overview, 370–373
models and approaches, 373–380
AIBO, 512
AIML. See Artificial Intelligence Markup Language (AIML)
Alexa, 353, 569–570
Alliance, 406
Alyx, 603
Amazon, 286
Amazon Sumerian, 270
AMI meeting corpus, 132
AML. See Avatar Markup Language (AML)
Android, 281
Android Speech, 281
Animal-assisted therapy (AAT), 202, 211
Animals, relationships with, 202
Animated characters, 439–440
Animation techniques, 278–279
APML. See Affective Presentation Markup Language (APML)
Apple MacOS, 281
Application-centered architectures, 99–100
AR. See Augmented Reality (AR)
Arbitration techniques, 34
ARIA project. See Artificial Retrieval of Information Agent (ARIA project)
Articulated Social Agents Platform (Asap), 275
Artificial agents, xix, 79
Artificial intelligence (AI), xxi, 262, 464
AI-based intelligent tutoring systems, 331
Artificial Intelligence Markup Language (AIML), 263
Artificial Retrieval of Information Agent (ARIA project), 272
Artificial Social Agent Platform (ASAP), 85
Asap. See Articulated Social Agents Platform (Asap)
ASA. See Artificial Social Agent Platform (ASAP)
AsapRealizer 2.0, 280
ASD. See Autism spectrum disorders (ASD)
Assistive agents, 369–370, 385–386
Assistive robots, 370
Assistive technologies, 369
Association for Computational Linguistics (ACL), 54
Associative route, 92, 94
AtomSpace, 266
Atrial fibrillation (AF), 417
Attachment theory, 202
Attention Deficit Hyperactivity Disorder (ADHD), 539
Attractive effect, 17
Audio, 90
Audio sensors, 90
Audio-visual sensing, 273, 280–283, 291 and speech recognition systems, 271
Augmented Reality (AR), 262, 519
Authorable virtual peers, 347
Authority, 410
Autism spectrum disorders (ASD), 336, 437, 562
advantages of socially interactive agents, 437
current challenges, 454
deficits in social communication and social interaction, 437–439
ethics in application of SIA for children with, 618–623
future directions, 454–455
history/overview, 441–453
models and approaches, 439–441
motor coordination deficit, 439
similarities and differences in IVAs and SRs, 453–454
Autobiographical memory, 465
Auton SR, 416
Autonomous agents and multiagent systems (AAMAS), xxix
Autonomous interactive systems, 173, 179
Autonomous vehicles, adding social signals to, 35
Autonomy, 384
AutoTutor intelligent tutoring system, 228
Avatar Markup Language (AML), 263
Avatars, xxviii, 498, 532
Background, 504
BARD dataset, 131
BARNGA, 536
Battle Buddy, 274
Bayesian Network, 166
model, 175
BEAT. See Behavior Expression Animation Toolkit (BEAT)
Beats, 473
Behavior Expression Animation Toolkit (BEAT), 175
Behavior generation
sub-architectures, 86–88
Behavior Markup Language (BML), 31, 85, 175, 263
Behavior Planner, 271, 279
Behavior Realizer, 280
Behavior selection, 92
Behavioral animation, 17–19
Behavioristic principles, 530
Believable characters, 235
Bertie language model, 351
Big five model, 157
Bird-oids model, 17
Black-box machine learning models, 102
Blender, 276
BlenderBot, 352
Block Challenge system, 452
Blueprints Visual Scripting, 267
BML. See Behavior Markup Language (BML)
Body in social space, 8–12
Body movement, 310
Body regions, 9, 15–16
BodyChat, 19
Boids, 17
Boston Dynamics videos, 615
Brittle rule-based systems, 225
Build-a-rocket game, 513
Bystanders, 5

C++ (programming language), 273
C# (programming language), 273
CADIA. See Center for Analysis and Design of Intelligent Agents (CADIA)
CAIO. See Cognitive and Affective Interaction-Oriented Architecture (CAIO)
Cambridge University Python Multi-domain Statistical Dialogue System Toolkit, 284
Care. Coach, 416
Carnegie Melon University (CMU), 284
Case studies, 378
CAVE, 537
Center for Analysis and Design of Intelligent Agents (CADIA), 24
CereVoice, 286
CERO, 207
Challenger, 503
Character generation, 291
Character Markup Language (CML), 263
Character-based approaches, 468
Character-based systems, xxix
ChatScript, 285
Choice, 388–389
Chronic disease, SIA to address, 417
CICERO system, 451
CiF model. See Comme il Faut model (CiF model)
Civil inattention model, 13, 19

Clinical care, SIA for patient education in, 419–420
Close salutations model, 14, 19
CML. See Character Markup Language (CML)
CMU. See Carnegie Melon University (CMU)
CMU Panoptic dataset, 131
CMUSphinx suite, 281
Co-design, 373–374
Co-embodiment, 123
Coach mike, 314
Cocktail Party dataset, 132
Coffee Break dataset, 132
Cognitive and Affective Interaction-Oriented Architecture (CAIO), 94
Cognitive architectures, 264–266
Cognitive Behavior(al) Therapy, 405, 414
Cognitive-psychological architectures, 100
Cognitively inspired architectures, 99–100
Coherent multimodal ensembles, 79
Cohesion, 118
Collaborative storytelling, 341
Comme il Faut model (CiF model), 509
Commentator, 503
Commercial platforms, 264, 266–270.
See also Academic platforms honorable mentions, 269–270
unity, 268–269
unreal engine, 267
Common focused gathering, 115
Common ground, 479
Communication styles, 317
Communicative intent, 86
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communion</td>
<td>159</td>
</tr>
<tr>
<td>Community-based development</td>
<td>264</td>
</tr>
<tr>
<td>Companions</td>
<td>502</td>
</tr>
<tr>
<td>Companionship</td>
<td>370</td>
</tr>
<tr>
<td>Compassion</td>
<td>403</td>
</tr>
<tr>
<td>Complementarity theory</td>
<td>160, 162</td>
</tr>
<tr>
<td>“Complete model” of multimodal interaction</td>
<td>81</td>
</tr>
<tr>
<td>Complex formation</td>
<td>6</td>
</tr>
<tr>
<td>Compliance</td>
<td>117, 413</td>
</tr>
<tr>
<td>Computational approaches</td>
<td>576</td>
</tr>
<tr>
<td>participants</td>
<td>577</td>
</tr>
<tr>
<td>questions</td>
<td>577–589</td>
</tr>
<tr>
<td>Computational modelers</td>
<td>467</td>
</tr>
<tr>
<td>Computer games</td>
<td>527</td>
</tr>
<tr>
<td>Computer literacy</td>
<td>407–408</td>
</tr>
<tr>
<td>Computer systems</td>
<td>173</td>
</tr>
<tr>
<td>Computer-based learning</td>
<td></td>
</tr>
<tr>
<td>environments</td>
<td>309</td>
</tr>
<tr>
<td>Conformity</td>
<td>117</td>
</tr>
<tr>
<td>Connection-agents</td>
<td>26</td>
</tr>
<tr>
<td>Consumer robots, living with</td>
<td>212–213</td>
</tr>
<tr>
<td>Consumers</td>
<td>217</td>
</tr>
<tr>
<td>Contemplation</td>
<td>411</td>
</tr>
<tr>
<td>Contemporary virtual peers</td>
<td>332</td>
</tr>
<tr>
<td>Control-theoretic approaches</td>
<td>381</td>
</tr>
<tr>
<td>Convenience</td>
<td>405–406</td>
</tr>
<tr>
<td>Conversational agent</td>
<td>213, 574</td>
</tr>
<tr>
<td>Conversational behaviours</td>
<td>127–129</td>
</tr>
<tr>
<td>Conversational interactions</td>
<td>314</td>
</tr>
<tr>
<td>Conversational management</td>
<td>114</td>
</tr>
<tr>
<td>Conversational strategies</td>
<td>352–353</td>
</tr>
<tr>
<td>Conversations</td>
<td>121</td>
</tr>
<tr>
<td>Core literacies</td>
<td>332</td>
</tr>
<tr>
<td>Cortana</td>
<td>353</td>
</tr>
<tr>
<td>Cost</td>
<td>405–406</td>
</tr>
<tr>
<td>Covered entity</td>
<td>423</td>
</tr>
<tr>
<td>Cross-cutting issues</td>
<td>1</td>
</tr>
<tr>
<td>Cross-layer interactions</td>
<td>501</td>
</tr>
<tr>
<td>Cross-modal behavior generation</td>
<td>87</td>
</tr>
<tr>
<td>Cross-modal mapping approach</td>
<td>86</td>
</tr>
<tr>
<td>Cross-neurological theory of mind</td>
<td>337</td>
</tr>
<tr>
<td>Crowd sourcing</td>
<td>479</td>
</tr>
<tr>
<td>CryEngine</td>
<td>270</td>
</tr>
<tr>
<td>Crystal Island</td>
<td>313</td>
</tr>
<tr>
<td>Cubic cubits</td>
<td>8–9</td>
</tr>
<tr>
<td>Cubits</td>
<td>8</td>
</tr>
<tr>
<td>Culture training system</td>
<td>21</td>
</tr>
<tr>
<td>Cybernetics</td>
<td>442</td>
</tr>
<tr>
<td>Dancing partners</td>
<td>115</td>
</tr>
<tr>
<td>Data-driven models</td>
<td>135</td>
</tr>
<tr>
<td>Data-Oriented Technology Stack (DOTS)</td>
<td>269</td>
</tr>
<tr>
<td>De-centering</td>
<td>337</td>
</tr>
<tr>
<td>Decision-theoretic reasoning</td>
<td>166–167, 178</td>
</tr>
<tr>
<td>Decision-theoretic systems</td>
<td>166</td>
</tr>
<tr>
<td>Deep integration with healthcare systems</td>
<td>426</td>
</tr>
<tr>
<td>Deep reinforcement learning</td>
<td>480</td>
</tr>
<tr>
<td>DeepPavlov</td>
<td>285</td>
</tr>
<tr>
<td>Degree of expressiveness</td>
<td>80</td>
</tr>
<tr>
<td>Deictic pointing</td>
<td>11</td>
</tr>
<tr>
<td>Deliberative route</td>
<td>83, 94</td>
</tr>
<tr>
<td>Dementia</td>
<td>381</td>
</tr>
<tr>
<td>Designer analysis</td>
<td>374</td>
</tr>
<tr>
<td>Desktop robots</td>
<td>475</td>
</tr>
<tr>
<td>DialogGPT</td>
<td>352</td>
</tr>
<tr>
<td>Dialogue Management (DM)</td>
<td>53</td>
</tr>
<tr>
<td>Dialogue systems</td>
<td>506</td>
</tr>
<tr>
<td>Diffusion of responsibility</td>
<td>117</td>
</tr>
<tr>
<td>Digital environments</td>
<td>310</td>
</tr>
<tr>
<td>Dimensional model</td>
<td>200–201</td>
</tr>
<tr>
<td>Discourse Plan Markup Language (DPML)</td>
<td>263</td>
</tr>
<tr>
<td>Dishonored</td>
<td>2, 507</td>
</tr>
</tbody>
</table>
Distance salutations model, 14, 19
“Distant peer” paradigm, 345
DM. See Dialogue Management (DM)
DOTS. See Data-Oriented Technology Stack (DOTS)
Double empathy problem, 337
DPML. See Discourse Plan Markup Language (DPML)
Dual-route architectures, 83–84, 93–95
Dual-route processing, 94
Dual/multi-route architectures, 97
Dyad formation, 5
Dyadic conversation archetypes, 25
Dyadic model, 200
Dynamic Bayesian Networks, 128
Dynamic Decision Network, 166
Dynamic interaction, xxiv
Dynamic social environments, 136
Dynamic social interaction, 199
Eavesdropping, 509
Education, 221
  curriculum, 196
  environments, 123
  and training, 384
Educational robotics, 443
Effective “reciprocal teaching” paradigm, 341
Elea dataset, 132
Element formation, 5
ELIZA system, 414
Embodied agents, 213
Embodied conversational agents (ECAs), xix, xxx, 19–20, 82, 331
platforms, 24
Embodied conversational behavior community, 31
Embodiment, 134, 382, 509, 532
Emergent Narrative, 467
Emoting signals, 314
Emotion assessment, 374
Emotion regulation, 533
Emotional alignment, 384
Emotional climate, 128
Emotional expressions, 511
Emotional social robot (EMYS), 511
Emotional stability, 158
Empathic understanding, 540–543
Empathy, 403, 408
EMYS. See Emotional social robot (EMYS)
Enablement, 384
Engagement, 370–371
Entitativity, 118
Entity-Component-System (ECS), 269
Ethical standards assessment, 384
Ethics, 600
  in application of SIA for children with autistic spectrum disorders, 618
  participants, 600, 618
  questions, 600–623
  of virtual peers, 353–354
Ethnographic field studies, 378
Evaluating adaptive systems, 171. See also Implementing adaptive systems
  autonomous interactive systems, 173
  perception studies, 172
  simulation, 171
  wizard-of-Oz studies, 172–173
Evaluation, 590
  participants, 590
  questions, 590–600
Evidence-Based Practice, 454
Expected utility, 166
Explainability, 569, 583, 585
Explainable AI, 607
Expressive speech, 274, 286
External properties of PA, 310–312
Eye movement, 449
EyeToy game Kinetic by Sony, 537

F-formation. See Face formation (F-formation)
F2FCrowds system, 29
Fabula and Sjuzet, 472
Façade, 507
Face, 159
  face-to-face conversation, 18
  face-to-face interaction, 3, 407–408
  threats, 160
Face formation (F-formation), 6, 116, 121
  in social conversations, 116–117
  theory, 126
Facial Action Coding System (FACS), 279
FACS. See Facial Action Coding System (FACS)
Fake user model, 171
Far Cry 4, 508
Fast sounds, 512
FAtiMA from FearNot!work, 476–477
FDA. See Food and Drug Administration (FDA)
Festival tool, 286
Fine-grained behavior timing, 177–179
Five factor model, 157–158
Flexible dialog, 225–226
Fluid social interaction, xxv
FML. See Function Markup Language (FML)
Focus groups, 374–375
Focused gatherings, 115
Focused interaction, 12, 14
Food and Drug Administration (FDA), 422
Formalizing continuous spatial behaviour, 31–32
Formations, 5
  changes and situated interactions, 135
Frame, 23
Full-fledged conversational social robots, 97
Function Markup Language (FML), 31, 85–86, 263
Furhat approach, 288
Furhat Robot, 91

Gameplay layer, 498–499
Games, 493
  current challenges, 516–518
  engines, 478
  future directions, 518–519
  history/overview, 504–513
  models and approaches, 494–503
  players, and experience, 494–497
  SIAs, 497–504
  similarities and differences between virtual and robotic
    SIA, 514–516
  systems, 506
Gamification, 527–528
Gatherings, 4, 13
  CADIA populous, 23
  social space gatherings and groups, 21–26
Gaze, 134
  behaviour, 164
  gaze-based interaction, 449
Generalized Intelligent Framework for Tutoring (GIFT), 275
Index

Generation components, 101
Gestures, 314
  synthesis, 580
GIFT. See Generalized Intelligent Framework for Tutoring (GIFT)
Goal-Oriented Action Planning (GOAP), 507
GOAP. See Goal-Oriented Action Planning (GOAP)
Godspeed Questionnaire Series (GQS), 230
GOLIAH, 438
Google, 286
Google Assistant, 353
GQS. See Godspeed Questionnaire Series (GQS)
Graphical user interfaces (GUIs), 211, 268
Greta platform, 271–272, 279
Greta system, 24
Group dynamics, 118
  automatic perception of, 126–127
Groups
  adaptation, 135
  approach behaviours, 129–130
  conversational behaviours, 127–129
  datasets, 130–132
  generating behavior in, 127–130
  learning, 336–337
  locomotion behaviours, 129
  polarization, 117
GUIs. See Graphical user interfaces (GUIs)
Gunslinger system, 60, 476
Hardware platforms, 262–263
Health, 370–371
  burden of chronic disease, 405
  burden of overweight and obesity, 404
  burden of substance abuse, 405
  communication and education, 409–410
  current challenges, 422–423
  example SIA systems in healthcare, 416–420
  future directions, 424–426
  health-related applications of socially interactive agents, 403
  importance of health behavior, 404
  virtual agents and health change, 214–215
  literacy factors, 407–408
  logistical factors, 405–406
  models and theoretical frameworks, 408–413
  SIAs in healthcare, 414–416
  similarities and differences between IVAs and SRs for health, 420–421
  social-emotional factors, 406–407
Health behavior
  change, 410–412
  sensing, 424–425
Health Belief Model, 410–411
Health literacy, 407–408
  PA support for, 320–321
Heating, Ventilation, and Air Condition (HVAC), 313
Helper, 503
Henchmen. See Subordinate
HiDAC system. See High-Density Autonomous Crowds system (HiDAC system)
High-Density Autonomous Crowds system (HiDAC system), 22
High-functioning autism, 336
HIPAA. See Hospital Health Insurance Portability and Accountability Act of 1996 (HIPAA)
Holodeck, 465, 469–470, 482
Hospital Health Insurance Portability and Accountability Act of 1996 (HIPAA), 423
HRI. See Human–Robot Interaction (HRI)
Human face-to-face interaction, xix
Human public behavior studies, 4
Human social behaviour, 17, 113
Human-to-human interactions, 233
Human–human interaction, xxii
databases, 131
Human–human relationships, 231
Human–human support, 321
Humanoid, 90
social robots, 91
Human–robot group, 124
partner interaction, 120
team, 120
Human–Robot Interaction (HRI), 197, 474
HVAC. See Heating, Ventilation, and Air Condition (HVAC)
I-STORY-TELLING, 470
iCat, 514, 565–566
ICT. See Institute for Creative Technologies (ICT)
iCub
felt, 90
robot, 89
Id Tech, 270
IDEA. See Individuals with Disabilities Education Act (IDEA)
Identification, 117
IDIAP Poster dataset, 131
IDtension, 476
Immediacy, 408
Immersive semi-automatic VR system, 452
Impaired hearing, 380
Impaired theory of mind, 337
Impaired vision, 380
Implementing adaptive systems, 163.
See also Evaluating adaptive systems
adaptation approaches, 164–171
expressing personality with SIAs, 163–164
Impulsion, 24
In situ studies, 176–177
In-character interactions, 500–501
In-game interactions, 500–501
Inclusive participatory design, 384
Incremental architectures, 84–86, 95–96
Incremental process, 84
Incremental systems, 578
Independence, 384
Independent or neutral SIAs, 502
Individual adaptation, 135
Individualization, 381
Individuals with Disabilities Education Act (IDEA), 369
Influencing group dynamics, 123–125
Information, 115
automation, 617
Infrared sensors, 89
Ingroup-favoritism, 118
Initial interoperability, 262
Institute for Creative Technologies (ICT), 273
Integrated authoring tools, 291
Integration problem, 77
Intelligent adaptive techniques, 381
Intelligent systems, 312
Intelligent tutoring systems (ITSs), 309, 331
Intelligent Virtual Agent (IVA), xix, xxi, xxviii, xxx, 1, 114, 213, 262, 305, 307, 562
of Autostadt in Wolfsburg, 537
relationship with, 318–319
Intent Planner, 271
Intent-based multimodal behavior generation, 86
Intent-based Partial Order Causal Link (IPOCL), 473
Interaction architecture, 78
Interaction memories and learning of SIA, 98–99
Interaction methods and modalities, 378
Interaction space, 20
Interaction Velocity Prediction, 29
Interaction-aware behavior generation, 100–101
Interactive narrative and story-telling, 465
current challenges, 476–480
future directions, 480–482
history/overview, 469–473
models and approaches, 466–469
similarities and differences between IVAs and SRs, 473–476
Interactive Voice Response systems (IVR systems), 415
Interactivity, 469–470
Internal models, 310
and properties of PA, 312–315
Internalization, 117
Internet connection, 280
Internet of Things (IoT), 217
long-term interaction with internet of things voice assistants, 216–218
Interoperability, 262
Interpersonal compatibility theories, 160–163
complementarity theory, 162
interpersonal goals, 162–163
similarity attraction theory, 160–161
Interpersonal goals, 162–163
Interpersonal interactions, 501
Interpretation model, 101
Interruptability, 80
Intersected formations, 7
Interviews, 378
Intimate distance, 10–11, 116
Introversion, 169
Inverse kinematics, 18
Involvement shield, 13
iOS, 281
IoT. See Internet of Things (IoT)
IPOCL. See Intent-based Partial Order Causal Link (IPOCL)
IrisTK, 288
ITSs. See Intelligent tutoring systems (ITSs)
IVR systems. See Interactive Voice Response systems (IVR systems)
Jack (human software framework), 18
Jack models, 18
JackMOO, 18–19
Java (programming language), 273
   Java-based system, 288
JESTiMuLe, 446
Jibo robot, 211, 288
Joint attention, 449
Jointly focused gathering, 115
Kaldi, 281
Keyframe animation, 278
Knowledge Query and Manipulation Language (KQML), 263
KQML. See Knowledge Query and Manipulation Language (KQML)
Laboratory environments, 136
Lanes, 13
   formation, 18
Language
   model, 280
   skills tests, 221
LARP. See Live Action Role Playing Games (LARP)
Laser scanners, 89
Leader–follower steering paradigm, 25
Learner, inhabiting same space as, 313
Learning, 529–534
   by doing, 530
   environments, 310, 315
   learning-based architectures, 98
   learning-by-teaching, 311
Leiden Institute of Advanced Computer Science (LIACS), 577
Leisure, 371
Leonardo robot, 89
LIACS. See Leiden Institute of Advanced Computer Science (LIACS)
Life-long personalization, 426
Lifelong learning, PA support for, 320
Liking, 410
Linguistic content, 163
Linux, 281
LiteBody, 280
Live Action Role Playing Games (LARP), 519
Locations, 5, 9, 11
Locomotion behaviours, 31, 33, 127, 129
Long-term adaptation and personalization, 227–229
Long-term adoption, 388
Long-term interaction, 195, 197–198
   attachment theory, 202
   designing relational SIAs, 202–207
   dimensional model, 200–201
   dyadic model, 200
   ethical design practices, 238–239
   ethical issues and design practices, 234
   with internet of things voice assistants, 216–218
   persuasion and social manipulation, 236–238
   privacy and security, 238
   provisional model, 201
   relational AI for SIAs, 233–234
   to relationship, 196–197
   relationship models and approaches, 199–200
   relationships with animals, 202
   with SIAs, 207
social bonds and authenticity, 235–236
with social robots, 207–213
social robots, virtual agents, voice assistants, and SIAs, 218–222
trends in long-term SIA research over past 20 years, 222–231
understanding novelty, 231–233
with virtual agents, 213–216
Long-term program, 622
Long-term relational SIAs, 204–207
Long-term SIA
achieving long-term beneficial outcomes from relational properties, 229–231
long-term adaptation and personalization, 227–229
long-term engagement, 223–224
research over past 20 years, 222–223
from voice interfaces to engaging conversation, 225–227
Low prestige dialects, 347, 350–351, 353
Lumberyard, 270
Lying, 509
M-PATH, 275
MACH system, 448
Machine learning (ML), 289, 381
approaches, 87, 177, 225, 579–580
machine learning-based models, 102
ML-agents, 269
Machine-generated voices, 316
MacOS, 281
MAE. See Mainstream American English (MAE)
Mainstream American English (MAE), 345
Maintenance, 411
Manipulation, 620–621
Mapping components, 101
Marketing agencies, 237
MaryTTS supports, 286
Massive multiplayer online role-play games (MMORPGs), 468, 478–479
MatchNMingle dataset, 131
Material embodiment, 510
Mecanim, 268
Mechanical teaching machines, 308
Medical literacy, PA support for, 320–321
Memory loss, 380
Memory model, 82, 98–99
Metrics, choice of, 136
MHHRI, 131
Micbot, 124
Microservices architecture, 290
Microsoft, 286
Microsoft platform, 283
Microsoft Windows, 281
Midas Touch Problem, 178
Minecraft, 478, 530
Minions. See Subordinate
Minstrel, 469
Misconceptions, 309
Mission Rehearsal Exercise, 471, 537
MIXER. See Moderating Interactions for Cross-Cultural Empathic Relationships (MIXER)
Mixing theoretical models, 33–34
ML. See Machine learning (ML)
MMORPGs. See Massive multiplayer online role-play games (MMORPGs)
Mobility, 134, 425
   changes, 380
MOCAP. See Motion Capture (MOCAP)
Mocap maps, 278
Model interactions tools, 280
   audio-visual sensing, 282–283
   expressive speech, 286–287
   NLP, 283–286
   speech recognition, 280–281
Moderating Interactions for Cross-Cultural Empathic Relationships (MIXER), 533, 536
Mole Madness, 512
Morph targets, 276
Motion Capture (MOCAP), 131, 278
Motivation, 529–530
Motivational Interviewing, 405, 412
Motor coordination deficit, 439
Movement across space, 32
MRS, 476
Mudbox software, 276
Multidirectional architectures, 84–86, 95–96
Multidirectional flow of information, 84
Multilayer dialogue model, 126
Multimodal interaction, 78
   requirements for, 79–80
Multimodal perception, 133
Multimodal signals, 82
Multimodal Utterance Representation Markup Language (MURML), 263
Multiparty interactions, 136
   advances in, 118
   automatic perception of group dynamics, 126–127
   current challenges and future directions, 134–137
   evaluating and understanding groups, 118
   face-formations in social conversations, 116–117
   generating behavior in groups, 127–130
   group datasets, 130–132
   group phenomena, 117–118
   groups and teams, 115–116
   between humans and SIA, 113–114
   influencing group dynamics, 123–125
   measuring attitudes, 119–120
   measuring believability, 118–119
   models and approaches, 115
   proxemics, 116
   similarities and differences in IVAs and SRs, 132–134
   social gatherings, 115
   spatial behavior understanding, 120–122
   understanding group dynamics, 122–123
Multiparty social settings, 126
Multiplatform support, 290
Multiple conversational functions, 79
Multiple hardware platform, 289
Multiple joints movement, 91
Multiple timescales, 80
MultiSense, 282
MURML. See Multimodal Utterance Representation Markup Language (MURML)
Mutual eye contact, 35
Nabaztag robot, 288
Narrative layer, 498–499
Narrative paradox, 467
Narrative structuring, 465
Narrative-based games, 533
National Science Foundation (NSF), 240
Natural conversational interactions, 97
Natural language
corpora, 480
models, 223
Natural Language Generation (NLG), 53
Natural language processing (NLP), 280, 283–285, 309, 379, 517
Natural Language Understanding (NLU), 53
Naturalistic supports, 313
Navigate space, 12
NCT. See Novelty Categorization Theory (NCT)
Negative politeness, 160
Neural network approaches, 170
Neurotypical peers, 336
NLG. See Natural Language Generation (NLG)
NLP. See Natural language processing (NLP)
NLU. See Natural Language Understanding (NLU)
Non-player Characters (NPC), 499
NonVerbal Behavior Generator (NVBG), 274
NonVerbal behaviour Analysis tool (NovA), 282
Nonverbal cues, 226–227
NovA. See NonVerbal behaviour Analysis tool (NovA)
Novelty Categorization Theory (NCT), 232
Novelty effect, 388
NPC. See Non-player Characters (NPC)
NPC Editor, 284
NSF. See National Science Foundation (NSF)
NVBG. See NonVerbal Behavior Generator (NVBG)
o-space, 117
o-zone, 6
OCC model. See Ortony Clore Collins model (OCC model)
Older adults, 369
Olympus, 284
One-on-one interactions, 113
“One-shot” studies, 565
One-to-one to one-to-many, 135
Online adaptation mechanisms, 171
Online rehearsal, 480
Open Source Computer Vision
Library (OpenCV), 282
OpenCog, 265
OpenCV. See Open Source Computer Vision Library (OpenCV)
OpenDial, 284
OpenFace 2.0 perform, 283
Openness to engagement model, 19
OpenPose perform, 283
OpenSense platform, 283
OpenSmile, 281–282
Operating System platform (OS platform), 281
Opponent, 502
Oppositional culture, 336
ORIENT, 471
Ortony Clore Collins model (OCC model), 174–175
OS platform. See Operating System platform (OS platform)
OTA. See Over-the-air (OTA)
Out-game interactions, 501
Out-of-character interactions, 501
Outlined research methods, xxv
Over-the-air (OTA), 217

p-zone, 7
PA. See Pedagogical Agents (PA)
PAD Temperament Model, 158
Pair programming, 334
Panda3D, 270
Paralinguistic behaviour, 163
Parasocial relationships, 349
PARO robot, 371, 382
Participatory design, 373–374
Patient group relationships, 196
PDDL. See Plan Domain Description Language (PDDL)
Pedagogical Agents (PA), 307–308, 331, 409. See also Socially Interactive Agent (SIA)
communication styles, 317
designing and implementing PA, 310
empirical status of, 315–316
external properties, 310–312
history of, 308–310
important elements of design space of, 311
internal models and properties, 312–315
new and emerging areas for PA research, 321
relationship with IVA and SR research, 318–319
research-based PA design, 316–317
summary of empirical status, 318
support for lifelong learning, 320
support for medical and health literacy, 321
Peer interaction, 333–334
debate and disagreement, 335
difference in peer social interaction, 336–337
ethics of virtual peers, 353–354
models and modeling, 350–352
peer-based learning, 338–339
roles and relationships in, 334
social interaction during task behavior, 337–338
speaking styles, 335–336
user modeling and conversational strategies, 352–353
virtual peers and SIAs, 339–350
Peer programming, 339
Peer-based learning, 338–339
Peer-to-peer interaction, 334
Pepper robot, 420
Perception, 133
Perception Markup Language (PML), 274
Perceptual evaluation, 25
Peripersonal space, 20
Personal distances, 10–11, 116
Personality, 155, 163–165, 178
five factor model, 157–158
interpersonal circumplex, 158–159
and interpersonal stance, 157
models, 156
neural learning of personality-based behavior styles, 170–171
PAD temperament model, 158
personality-based behavior variations, 172
politeness theory, 159–160
Personalization, 567–568
Personalized Plan-Based Presenter (PPP), 175
Persuasion, 620–621
Persuasive technology, 410
Physical activity, 416
Physical agents, 387
Physical robots, 155
Physically embodied agents, 371, 387
Ping-pong games, 115
Plan Domain Description Language (PDDL), 481
Plan Network, 479
Players, 494–497
layer, 498
Pleasure–Arousal–Dominance Model, 158
Plot-based systems, xxvii, 478
PML. See Perception Markup Language (PML)
Policies, 389
Politeness theory, 159–160
Populus, 22–24
Positive politeness, 160
PPP. See Personalized Plan-Based Presenter (PPP)
Precontemplation, 411
Predicted timing behaviour, 85
Predictive repulsive force, 24
Preferred velocity, 17
Preparation, 411
Privacy, 384, 388, 569
Procedural animation, 278
systems, 479–480
Procedural generation, 478
Procedurally generated animation, 278
Processing components, 101
PromWeek, 476
Prosocial agents, 573
Provider, 503
Provision models, 201
Provisional model, 201
Proxemics, 116
classification system, 10
Psyclone messaging system, 271
Public distance, 10–11, 116
PyDial, 284
Python, 284
Q-learning algorithm, 168, 229
QUD. See Questions Under Discussion (QUD)
Questionnaires, 378
Questions Under Discussion (QUD), 50
R-Motor coordination deficit, 439
r-zone, 7
Randomized control trial (RCT), 210
Rasa, 285
RavenClaw, 284
RCT. See Randomized control trial (RCT)
Re-embodiment, 123
Reactive route, 83
Reading literacy, 407–408
Real-world environments, 101
Real-world multiparty interactions, 137
Reasoning method, 225
Received Pronunciation voice (RP voice), 605
Reciprocal teaching approach, 311
Reciprocity, 410
Recognition model, 101
ReCVEB. See Research Center for Virtual Environments and Behavior (ReCVEB)

Regions, 6

Reinforcement learning (RL), 480
  algorithms, 167–170

Relational agents, 82, 213, 274
  approach, 280

Relational AI for SIAs, 233–234

Relational SIAs, 199, 238

Relationships
  with animals, 202
    model, 200
    people, 196

Replicability of results, 136

Research Center for Virtual Environments and Behavior (ReCVEB), 26

Research-based PA design, 316–317

Responsiveness, 80

Retention, 413

Reusability, 585

Reynolds’ animal-focused models, 18

RIAS. See Roter Interaction Analysis System (RIAS)

RL. See Reinforcement learning (RL)

Robot Operating System (ROS), 60, 288

Robot Reeti®, 91

Robotic dog, 235

Robotic SIAs in games, 509–513
  similarities and differences, 514–515

  efficacy, 123
  gaze, 127
  gestures, 164
  personality profile, 165

  robot-based drama, 475

Robust multiparty interactions, 114

Roomba vacuum cleaner, 198

ROS. See Robot Operating System (ROS)

Roter Interaction Analysis System (RIAS), 594

RP voice. See Received Pronunciation voice (RP voice)

Rule-based systems, 227, 279

Safety, 384

SAIBA. See Situation, Agent, Intention, Behavior, Animation (SAIBA)

Saliency maps, 583

SALSA¹ dataset, 132

Salutation sequence, 14

Sam system, 535

SAPO, 476

SASO, 471

Science, Technology, Engineering, and Math (STEM), 339

Seat, 7

Security, 388

Sedentary behavior, SIA to address, 416

Self-reporting tools, 374

SEMAINE project, 272

Semiotic system, 79

Semisynthetic dataset, 132

Sensing
  modalities, social robot architectures, 89–90
  model, 101

Sensitivity, finding right level of, 177–178

Sensor data, 96

Sensory data, 92
Sentencing, 617
Serious games, 372–373, 527
  content and story, 535–536
  current challenges and future directions, 540–545
  evaluation, 543–545
  history and overview, 534–540
  immersiveness and interaction, 537–538
  individual needs and adaptation, 538–540
  learning and user experience, 529–534
  with SIAs, 528
Shadowing, 374
Short-term program, 622
Sidekicks. See Companions
Sigma (Σ), 266
Similarity attraction
  hypothesis, 312
  theory, 160–161
SimQuest, 530
Simulated user model, 171
Simulation
  process, 171
  of theory of mind, 351
Single deliberative route, 83
Single-route architectures, 82–83, 92–93
Situated agent, 369
Situation, Agent, Intention, Behavior, Animation (SAIBA), 85, 577–578
Slot positions, 25
Small learning rate, 170
Smart speakers, 196
SmartBody, 31, 280
  character animation system, 28
Smartphones, 196, 263
SML. See Soar Markup Language (SML)
Soar, 265
Soar Markup Language (SML), 265
Social affiliation, 15
Social agents, 79
Social behavior
  creating accessible social behavior toolkits, 34
  generation, 133
Social bonds, 346
  and authenticity, 235–236
Social Cognitive Theory, 411
Social communication, deficits in, 437–439
Social connectedness, 384
Social constructivism, 531
Social conversations, 115
  face-formations in, 116–117
Social distance, 10–11
Social engagement, 15
Social exchange theory, 201
Social Force Model, 17, 22
Social function, 597
Social gathering, 115
Social influence, 11
Social intelligence in virtual peers, 348–350
Social interactions, xix, 25, 497, 563
  deficits in, 437–439
  during task behavior, 337–338
  participants, 563
  questions, 564–567
Social learning, 310
Social nonverbal behaviour, 227
Social occasion, 5
Social perception, 22
Social presence, 511
Social puppet, 21
Index

Social relationship, 385
Social robotics, xix, xxi, 262

similarities and differences in, 287–289
Social robots (SR), xx, 1, 79, 89–90, 114, 199, 209–211, 219–220, 222,
224, 236, 305, 307, 368–369,
373, 387, 442, 562, 575
action modalities, 90–92
architectures, 88–97
and children, 209–210
dual-route architectures, 93–95
for health and wellness, 210–212
learning, 236
living with consumer robots,
212–213
long-term interaction with,
207–213
long-term multimodal behaviour,
102
multidirectional, incremental
architectures, 95–96
relationship with, 318–319
sensing modalities, 89–90
similarities and differences of
social robot architectures,
97–98, 218–222
single-route architectures, 92–93
Social Signal Interpretation (SSI), 282
Social signals, 77, 165, 314
to autonomous vehicles, 35
Social situation, 4
Social spaces, 3–4

adding social signals to
autonomous vehicles, 35
animating movement in tight
spaces, 32–33
behavioral animation, 17–19
body in, 8–12
creating accessible social
behavior toolkits, 34
embodied conversation, 19–20
formalizing continuous spatial
behaviour, 31–32
gatherings and groups, 21–26
immersive studies, 26–29
increasing multisensory fidelity
in XR, 36
interaction functions in, 12–16
mixing theoretical models, 33–34
models and approaches, 4
motivation, 3–4
relatively large indoor social
space, 5
similarities and differences in
IVAs and SRs, 29–31
space occupied by and
surrounding face-formation, 6
structure of, 4–7
Social support theory, 201
Social Tutor system, 451
Social-aware methods, 129
Social-aware navigation systems, 129
Social-psychological theory of
embodiment, 532
Socialization, 531
Socially assistive agents, 368
Socially Assistive Robot, xxviii
Socially Intelligent Agent, xxviii
Socially intelligent assistants, 155
Socially intelligent manner, xx
Socially Interactive Agents (SIAs), xix,
xxix, xxx, 1, 77, 113, 176, 195,
261, 305, 332, 339, 493, 528,
561. See also Pedagogical
Agents (PA)
advantages of, 437
agent platforms, 264–275
animation, 278–279
challenges, 288
cognitively inspired vs.
application-centered architectures, 99–100
designing relational SIAs, 202
designing SIAs as relational partners, 203–204
dialogue, 45–65
difference in SIAs as peers, 347–348
in different interactive settings, 445–453
distribution methods, 264
embodied conversational or virtual agent architectures, 82–88
evaluation measures, 102
examples of, xx
expressing personality with, 163–164
games, 497–504
in healthcare, 414–416
history and trends, 262–264
interaction memories and learning, 98–99
interaction-aware behavior generation, 100–101
long-term interaction with, 207–218
long-term relational SIAs, 204–206
modeling, 276–277
models and approaches, 80–82
nonverbal behavior generation and realization, 279–280
as peer, 340
relational AI for, 233–234
requirements for multimodal interaction, 79–80
social robot architectures, 88–97
technology, 195
tools to create appearance and nonverbal behaviour, 276
tools to model interactions, 280–287
uncertainty-awareness in social interaction, 101–102
Socially Interactive Robot, xxix
Socio-cognitive conflict, 335
Socio-emotional skills, 332
SoftBank Robotics, 442, 444
Source engine, 270
Spatial behavior understanding, 120–122
Speech recognition, 280
system, 280–281
Speech Synthesis (TTS), 53
Speech Synthesis Markup Language (SSML), 263
Speech-accompanying behaviours, 88
Speech-based health counseling systems, 415
Speed-runs, 496
Spoken Dialog Systems, 83
SSI. See Social Signal Interpretation (SSI)
SSML. See Speech Synthesis Markup Language (SSML)
"Stages of change" model. See Transtheoretical Model (TTM)
Standard Patient Studio, 275
Stanislavsky. See Active Analysis rehearsal technique (Stanislavsky)

Star Trek, 465

STEM. See Science, Technology, Engineering, and Math (STEM)

Step-and-slide technique, 13

Steve, 535

Stigma, 406

StimCards, 510

Story director agent, 468

Story Listening Systems, 332

Story-telling robots, 475

STORYTEC, 477

Streams, 13

Structuralism, 466

Students, 332

Student–tutor relationships, 196

Subordinate, 502, 519

Substance abuse, SIA to address, 417–419

Sueca, 514

Tabletop role-play, 478

Tablets, 263

Tacotron, 581

Tactical language system, 21

Tactile sensors, 90

Tailored communication, 409

Tailored health intervention, 425

TALESPIN, 464–465

TARDIS project, 272

Task segment, 20

Task-related data, 168

Teachable agents, 311

Teammate, 502

Technology acceptance, 384–385

Territory, 10

Tex2Shape, 277

Text-based multiple-choice actions, 536

Text-to-speech generation, 286

Text-to-speech voices (TTS voices), 316

Theoretical model of social space, 4

Theory of Planned Behavior, 411

Theory of Reasoned Action, 411

Therapeutic Alliance, 412–413

Therapist–patient relationships, 196

Therapy, 370–371

Thesian, 476–477

Tight spaces, animating movement in, 32–33

Timed Petri Nets (TPN), 98

TinyCog, 266

Touch, 90

TPN. See Timed Petri Nets (TPN)

Traditional modeling software, 276

Transactional segments, 10, 20, 116

Transparency, 569, 607, 620

Transtheoretical Model (TTM), 409, 411

Triggering step model, 25

Trust, 386–387, 406, 412–413, 541

TTM. See Transtheoretical Model (TTM)

TTS voices. See Text-to-speech voices (TTS voices)

Turn-taking patterns, 164

Tutor effect, 339, 342

Tutoring agents, 332

UE-HRI dataset, 131

Ultrasound, 89

Uncertainty-awareness in social interaction, 101–102

Unconstrained user utterances, 412
Understanding group dynamics, 122–123
Unfocused gatherings, 115
Unfocused interaction, 12
Unity, 267–269
editor, 268
Unreal engine, 266–267
User experience, 529–534
User modeling, 352–353
approaches, 374
User privacy, 340
User-centered design methods, 373–374
UTEP-ICT dataset, 132
VEIIG dataset, 132
Vernissage Dataset, 131
VHML. See Virtual Human Markup Language (VHML)
VHMsg platform, 273
VHToolkit. See Virtual Human Toolkit (VHToolkit)
VIB. See Virtual Interactive Behavior System (VIB)
Video modeling, 447
VIGART system, 449
Virtual agents, 22, 26, 173, 213, 219, 371, 387, 444–445, 575
architectures, 82
behavior, 102
behavior generation
sub-architectures, 86–88
dual-route architectures, 83–84
and health change, 214–215
interaction between virtual agent
and user, 448–450
long-term interaction with, 213–216
multidirectional, incremental
architectures, 84–86
non-interactive presentation with
virtual agent and user, 446–448
similarities and differences of,
97–98, 218–222
single-route architectures, 82–83
systems using multiple virtual
agents, 450–453
traversing three-dimensional
space, 17
and wellness, 216
Virtual characters, xxix, 587
Virtual Human Markup Language (VHML), 263
Virtual Human Toolkit (VH Toolkit),
273–274, 476
Virtual Humans, xxix
Virtual interactive agents, 79
Virtual Interactive Behavior System (VIB), 24
Virtual partner (VP), 441
Virtual peers, 332, 337, 339–350
debate, teasing, and
disagreement in, 343–344
ethics of, 353–354
roles for, 341–343
social intelligence in, 348–350
varying speaking styles in,
344–347
Virtual People Factory, 275
Virtual Reality (VR), xxviii, 262, 382, 439
Virtual Reality Job Interview Training system (VR-JIT system), 448
Virtual Reality Social Cognition Training system (VR-SCT system), 450
Virtual SIAs in games, 505–509
similarities and differences, 514–516
Virtual Storyteller, 472
Virtual user interface (VUI), 212
Vision, 90
Visual SceneMaker, 275
Voice assistants, 371
similarities and differences in, 218–222
Voice interfaces to engaging conversation, 225–227
flexible dialog, 225–226
nonverbal cues, 226–227
Voice-controlled devices. See Voice assistants
Voices robots, 155
VP. See Virtual partner (VP)
VR-JIT system. See Virtual Reality Job Interview Training system (VR-JIT system)
VR-SCT system. See Virtual Reality Social Cognition Training system (VR-SCT system)
VUI. See Virtual user interface (VUI)
WASABI, 275
Wellness, virtual agents and, 216
Windows, 281
Wizard-of-Oz mode, 344
studies, 172–173
WOLF dataset, 132
Working Alliance, 412
World of Warcraft, 478
XAI, 582
XML-based specification languages, 86
XR technologies, 30
increasing multisensory fidelity in, 36
ZBrush software, 276
Zero-effort technology (ZET), 383
ZET. See Zero-effort technology (ZET)
Zones, 6
The Handbook on Socially Interactive Agents
20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics
Volume 2: Interactivity, Platforms, Application

Birgit Lugrin, Catherine Pelachaud, David Traum, (Editors)

The Handbook on Socially Interactive Agents provides a comprehensive overview of the research fields of Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics. Socially Interactive Agents (SIAs), whether virtually or physically embodied, are autonomous agents that are able to perceive an environment including people or other agents, reason, and decide how to interact, and express attitudes such as emotions, engagement, or empathy. They are capable of interacting with people and each other in a socially intelligent manner using multimodal communicative behaviors with the goal to support humans in various domains.

Written by international experts in their respective fields, the book summarizes research in the many important research communities pertinent for SIAs, while discussing current challenges and future directions. The handbook provides easy access to modeling and studying SIAs for researchers and students and aims at further bridging the gap between the research communities involved.

In two volumes, the book clearly structures the vast body of research. The first volume starts by introducing what is involved in SIAs research, in particular research methodologies and ethical implications of developing SIAs. It further examines research on appearance and behavior, focusing on multimodality. Finally, social cognition for SIAs is investigated by different theoretical models and phenomena such as theory of mind or pro-sociality. The second volume starts with perspectives on interaction, examined from different angles such as interaction in social space, group interaction, or long-term interaction. It also includes an extensive overview summarizing research and systems of human-agent platforms and of some of the major application areas of SIAs such as education, aging support, autism or games.

ABOUT ACM BOOKS
ACM Books is a series of high-quality books published by ACM for the computer science community. ACM Books publications are widely distributed in print and digital formats by major booksellers and are available to libraries and library consortia. Individual ACM members may access ACM Books publications via separate annual subscription.

BOOKS.ACM.ORG • WWW.MORGANCLAYPOOLPUBLISHERS.COM