Towards Computational Models of Cultural Dynamics Based on the Grounding Model of Cultural Transmission

Jens Pfau

Department of Computing and Information Systems
The University of Melbourne
Victoria, Australia

jpfau@unimelb.edu.au

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

August 2012

Produced on archival quality paper
Abstract

The subject of this thesis is the computational modeling of the transmission of cultural information and of the resulting emergent cultural dynamics—the formation, maintenance, and change of culture. The main premise is that this subject requires a two-component research program: (1) the detailed study of the transmission of cultural information during social interactions, and (2) an uptake of these results to analyze the emerging phenomena at the level of societies. We adopt the view that computational modeling is a particularly promising tool for addressing the second component, while it can also support fields such as psychology and cognitive science in addressing the first component. The reasons for this are that computational models offer (1) an appropriate language for precisely describing mechanisms at all relevant levels of analysis (i.e. intra-personal, inter-personal, and societal-level mechanisms) and (2) a promising opportunity to consider all of these levels simultaneously. The consequence is that computational models can contribute to a refinement of the models or theories they represent, and they can be used to rigorously explore the implications of these models or theories on varying scales.

We rely on the social-psychological grounding model of cultural transmission, which provides a detailed model of cultural transmission during social interactions, in constructing computational models of cultural transmission and dynamics. The research questions we address are (1) how can the grounding model of cultural transmission be translated into computational models of cultural transmission and dynamics, and (2) how can these computational models contribute to the refinement of the grounding model of cultural transmission and to the understanding of cultural dynamics. The contribution towards the first research question is a series of computational models of cultural transmission and dynamics based on the grounding model of cultural transmission. The contribution towards the second question is an analysis of each of these models with respect to its refinement of the grounding model of cultural transmission and its contribution to the understanding of cultural dynamics. The models we introduce and discuss are, in particular, (1) a model of the co-evolution of cultures, social networks, and geographical locations, (2) a model of the communication of stereotype-relevant information, considering that stereotypes are a particular type of cultural information, (3) a detailed semi-formal model of the grounding model of cultural transmission, and (4) an architecture of joint action which accounts for empirically observed phenomena relevant to cultural transmission.
Declaration

This is to certify that

(i) the thesis comprises only my original work towards the Ph.D. except where indicated in the Preface,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Jens Pfau

Jens Pfau
August 2012
To my wife, my father, and his father.
List of Publications

The major part of Chapter 4 was published as follows:


An earlier version of the results described in Chapter 4 was published as follows:


A part of Chapter 5 has appeared in the following publication:


Preliminary and condensed versions of Chapter 6 appear in the following publications:


Acknowledgements

This thesis has greatly benefited from a number of people that deserve special mention.

First and foremost, I would like to thank my supervisors Michael Kirley, Liz Sonenberg, and Yoshi Kashima for their continuous mentoring, support, encouragement, and trust in the last years. All of them have spent an extraordinary amount of time and effort on the interactions with me. I am especially grateful for the insightful discussions we have had and for the opportunities that they have provided. It has been an utmost joy to work with them.

I am grateful to Adrian Pearce, the chair of my advisory committee, for general advice on my project. Two anonymous reviewers have provided insightful comments that have helped shape the final version of this thesis. I appreciate their effort.

In addition, I would like to thank all those colleagues and friends that have spent time with me discussing the topic of my work or that have otherwise offered their comments. In particular, I would like to thank Frank Dignum, Gert Jan Hofstede, Gil Tidhar, Ian Wood, Kathleen Keogh, Minh Van Nguyen, Nicole Ronald, Peter McBurney, and Samin Karim. Special thanks go to Golriz Rezaei, Mohsen Jafari Songhori, and Tim Miller, who have co-authored papers with me. I am also grateful to Tim Miller and Shanika Karunasekera for giving me the opportunity to serve as tutor for two interesting subjects.

I would like to acknowledge the generous financial support that I have received from the University of Melbourne (MIRS scholarship) and the Australian government (Endeavour IPRS scholarship). The Melbourne School of Engineering, NICTA, the Complex Open Systems Research Network, and the European Social Simulation Association have supported me generously in my participation of conferences, summer schools, and short courses.

The last few years in Melbourne have been an amazing time that leaves me with a lot of unforgettable memories. This is mainly due to the company and support of my dear friends Agata and Stefan Pohl, Daniel Kless, Fariba Khezri, Klaus Boldt, Mani Abedini, Mohsen Jafari Songhori, Sabrina and Stephan Mehlhase, Silvia Graumann, and Zhiqi Tao. I am also thankful to my previous and current office mates, for shared lunches and laughters.

I would like to express my deepest gratitude and appreciation to my wife Agnieszka, for her enduring love, support, trust, and devotion. It has been an exceptional pleasure and joy to share this Australian adventure with her. I am also indebted to both our families, who have always supported us and contributed to our well-being from afar.

Jens Pfau
Melbourne, Australia, August 2012
# Table of Contents

## 1 Introduction
1.1 The Grounding Model of Cultural Transmission .......................... 3
1.2 Computational Models of Cultural Dynamics .......................... 4
1.3 Research Questions and Contributions .......................... 5
1.4 Thesis Structure .......................... 7

## 2 Culture
2.1 Culture and Cultural Dynamics ........................................... 11
   2.1.1 The Micro-Macro Problem ........................................... 12
   2.1.2 Levels of Analysis in Cultural Dynamics .......................... 13
   2.1.3 Cultural Dynamics as a Complex System .......................... 14
2.2 The Grounding Model of Cultural Transmission .......................... 16
   2.2.1 Common Ground and Grounding Revisited .......................... 19
   2.2.2 Example Dialogues ........................................... 21
   2.2.3 Premises and Implications Revisited ................................ 23
2.3 Conclusions ........................................... 26

## 3 Methodology
3.1 Computational Models of Social Dynamics .......................... 28
   3.1.1 Models, Theories, and Simulations .......................... 28
   3.1.2 The Case for Computational Models of Social Dynamics .......................... 35
   3.1.3 Existing Computational Models of Social Dynamics .......................... 37
3.2 Developing Agent-based Models ........................................... 43
   3.2.1 Purpose Identification ........................................... 44
   3.2.2 Conceptual Modeling and Validation ................................ 46
   3.2.3 Implementation and Verification ................................ 51
   3.2.4 Analysis and Operational Validation ................................ 53
   3.2.5 Summary ........................................... 56
3.3 The Limits and Pitfalls of Computational Modeling .......................... 57
3.4 The Models in this Thesis ........................................... 58
3.5 Conclusions ........................................... 61
4 An Agent-based Model of the Co-evolution of Cultures, Social Network Communities, and Agent Locations 63
4.1 Model ........................................ 66
4.1.1 Interaction .................................... 67
4.1.2 Common Ground and the Evolution of Cultural Space ............ 68
4.1.3 The Evolution of Social and Physical Space ................. 68
4.2 Formal Analysis .................................. 69
4.3 Sensitivity Analysis ................................ 72
4.3.1 The Time-Evolution of Cultures and Networks ............ 73
4.3.2 The Final Configuration of Cultures and Networks .......... 76
4.4 Observations ..................................... 80
4.5 Discussion and Conclusions ................................ 83
4.5.1 Summary of Results ............................ 83
4.5.2 The Effect of Globalization on Cultural Diversity .......... 84
4.5.3 Future Work .................................... 85

5 An Agent-based Model of Stereotype Communication 87
5.1 Experiments by Lyons and Kashima .............................. 90
5.2 Connectionist Models of Stereotype Communication .......... 92
5.3 Model ............................................. 95
5.3.1 Overview ....................................... 95
5.3.2 Memory Storage and Reproduction ....................... 96
5.3.3 Network Processing ............................. 97
5.3.4 Communication .................................. 103
5.4 Parameter Estimation .................................. 107
5.4.1 Evolutionary Algorithms .......................... 107
5.4.2 Method ......................................... 112
5.4.3 Results .......................................... 117
5.5 Validation ......................................... 119
5.6 Sensitivity Analysis .................................. 122
5.7 The Influence of Network Configuration on Stereotype Communication .......... 124
5.7.1 Model .......................................... 124
5.7.2 Results .......................................... 125
5.8 Discussion and Conclusions ................................ 128

6 Towards a Formal Model of the Grounding Model of Cultural Transmission 131
6.1 Common Ground .................................... 133
6.1.1 An Analysis of Common Ground .......................... 134
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.6.3 Implicit Transmission of Embodied Culture</td>
<td>221</td>
</tr>
<tr>
<td>7.6.4 Future Work</td>
<td>222</td>
</tr>
<tr>
<td>8 An Implementation of the Joint Action Architecture</td>
<td>223</td>
</tr>
<tr>
<td>8.1 Model</td>
<td>224</td>
</tr>
<tr>
<td>8.1.1 Associative Memory</td>
<td>224</td>
</tr>
<tr>
<td>8.1.2 Symbolic Memory</td>
<td>227</td>
</tr>
<tr>
<td>8.1.3 Interaction between Symbolic and Associative Memory</td>
<td>228</td>
</tr>
<tr>
<td>8.1.4 The Non-intentional Level</td>
<td>230</td>
</tr>
<tr>
<td>8.1.5 The Intentional Level</td>
<td>231</td>
</tr>
<tr>
<td>8.2 Evaluation</td>
<td>232</td>
</tr>
<tr>
<td>8.2.1 Setup</td>
<td>232</td>
</tr>
<tr>
<td>8.2.2 Sample Trial</td>
<td>237</td>
</tr>
<tr>
<td>8.2.3 Parameter Estimation</td>
<td>239</td>
</tr>
<tr>
<td>8.3 Related Work</td>
<td>240</td>
</tr>
<tr>
<td>8.3.1 MOSAIC</td>
<td>241</td>
</tr>
<tr>
<td>8.3.2 SCM</td>
<td>244</td>
</tr>
<tr>
<td>8.3.3 HiTEC</td>
<td>246</td>
</tr>
<tr>
<td>8.3.4 DiPRA</td>
<td>249</td>
</tr>
<tr>
<td>8.3.5 Summary</td>
<td>251</td>
</tr>
<tr>
<td>8.4 Discussion and Conclusions</td>
<td>252</td>
</tr>
<tr>
<td>9 Conclusions</td>
<td>255</td>
</tr>
<tr>
<td>9.1 Contributions</td>
<td>256</td>
</tr>
<tr>
<td>9.2 Limitations</td>
<td>259</td>
</tr>
<tr>
<td>9.3 Future Work</td>
<td>260</td>
</tr>
<tr>
<td>9.4 Concluding Remarks</td>
<td>261</td>
</tr>
<tr>
<td>Bibliography</td>
<td>263</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Coleman’s understanding of micro and macro level ...................... 12
2.2 Macro, micro, and nano level in cultural transmission and dynamics .... 14

3.1 Models, theories, and target systems .................................. 30
3.2 The process of developing agent-based models .......................... 45

4.1 Possible situation in Axelrod’s model of the dissemination of culture .... 65
4.2 Interactions between cultural, social, and physical space ................ 70
4.3 Example of a Markov chain ............................................. 72
4.4 Sample snapshot of a situation in the model ............................. 74
4.5 Average number of communities over time ............................. 75
4.6 Average intra-community diversity over time ........................... 76
4.7 Average inter-community diversity over time ........................... 77
4.8 Average number of communities and cultures at equilibrium .......... 77
4.9 The $q$-dependent phase transition at equilibrium ..................... 79
4.10 Location of the critical value $q_c$ .................................... 79
4.11 Distribution of community and culture sizes at equilibrium .......... 80
4.12 Sample trajectories of inter- and intra-community diversity over time .... 80

5.1 Experimental setup of Lyons and Kashima ............................. 90
5.2 Example of a fully recurrent connectionist network ..................... 93
5.3 Network architecture of the model ..................................... 97
5.4 Unrolling of a recurrent network ..................................... 100
5.5 Local search illustration ............................................. 109
5.6 Evolutionary algorithm process ...................................... 110
5.7 Evolutionary algorithm illustration ................................... 111
5.8 Fitness of parameters during all generations ........................... 117
5.9 Distribution of parameter values ...................................... 118
5.10 Information reproduced at each position in shared and unshared conditions 119
5.11 Overall information reproduced in ignorance and knowledge conditions ... 119
5.12 Information reproduced at each position in shared-ignorance, shared-partial, and shared-knowledge conditions ......................... 121
5.13 Sensitivity analysis .................................................. 123
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.14</td>
<td>Information transmitted by members of two communities over time</td>
<td>126</td>
</tr>
<tr>
<td>6.1</td>
<td>The main components of the grounding model of cultural transmission</td>
<td>132</td>
</tr>
<tr>
<td>6.2</td>
<td>An example of a Common Ground Logic model</td>
<td>149</td>
</tr>
<tr>
<td>6.3</td>
<td>Counter-example for negative introspection</td>
<td>157</td>
</tr>
<tr>
<td>7.1</td>
<td>The individual Simon task</td>
<td>188</td>
</tr>
<tr>
<td>7.2</td>
<td>Interaction between symbolic and associative memory</td>
<td>197</td>
</tr>
<tr>
<td>7.3</td>
<td>Control and information flow at non-intentional and perception-action level</td>
<td>200</td>
</tr>
<tr>
<td>7.4</td>
<td>The functions of the non-intentional level</td>
<td>201</td>
</tr>
<tr>
<td>7.5</td>
<td>Information flow and dependencies of the intentional level</td>
<td>203</td>
</tr>
<tr>
<td>7.6</td>
<td>Elements of symbolic and associative memory and their relationships</td>
<td>209</td>
</tr>
<tr>
<td>7.7</td>
<td>Intention structure in the no-conflict condition</td>
<td>212</td>
</tr>
<tr>
<td>8.1</td>
<td>Elements of symbolic and associative memory and their relationships</td>
<td>235</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Overview of models in this thesis .............................................. 60
5.1 Patterns learnt by agents in priming phase ................................. 114
5.2 Overview of parameters ......................................................... 116
6.1 Summary of the intended interpretations of operators .................. 145
7.1 Setup of the experimental conditions in the individual Simon task .... 189
7.2 Setup of the experimental conditions in the social Simon task .......... 190
7.3 Mapping between representations in symbolic and associative memory ... 208
7.4 Activation of features in individual and social Simon task ............... 216
8.1 Sample parameters ............................................................... 241
## List of Symbols

### Chapter 2

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>A proposition (informal) expressed by a speaker</td>
<td>17</td>
</tr>
</tbody>
</table>

### Chapter 4

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_p$</td>
<td>Physical mobility</td>
<td>69</td>
</tr>
<tr>
<td>$\beta_s$</td>
<td>Social mobility</td>
<td>67</td>
</tr>
<tr>
<td>$\delta(x,y)$</td>
<td>Kronecker’s delta</td>
<td>67</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Change of link weight after interaction</td>
<td>68</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Bias for interaction between agents that do not have any social link</td>
<td>67</td>
</tr>
<tr>
<td>$A$</td>
<td>Population of agents</td>
<td>66</td>
</tr>
<tr>
<td>$C(A)$</td>
<td>Set of communities within the network of agents $A$</td>
<td>73</td>
</tr>
<tr>
<td>$\sigma_{ai}$</td>
<td>The $i$th cultural trait of agent $a$</td>
<td>67</td>
</tr>
<tr>
<td>$a, b, c$</td>
<td>Agents</td>
<td>66</td>
</tr>
<tr>
<td>$c(a,b)$</td>
<td>Cultural similarity between agents $a$ and $b$</td>
<td>67</td>
</tr>
<tr>
<td>$C_i$</td>
<td>A community of agents in the social network</td>
<td>73</td>
</tr>
<tr>
<td>$D(G)$</td>
<td>Cultural diversity of the group of agents $G$</td>
<td>73</td>
</tr>
<tr>
<td>$d_L(a,b)$</td>
<td>Normalized Euclidean distance between agents $a$ and $b$</td>
<td>66</td>
</tr>
<tr>
<td>$F$</td>
<td>Number of cultural features per agent</td>
<td>64</td>
</tr>
<tr>
<td>$i, j$</td>
<td>States in the Markov chain</td>
<td>70</td>
</tr>
<tr>
<td>$I(a,b)$</td>
<td>Probability of interaction between agents $a$ and $b$</td>
<td>67</td>
</tr>
<tr>
<td>$L$</td>
<td>Lattice width</td>
<td>66</td>
</tr>
<tr>
<td>$M(a,t)$</td>
<td>Probability of agent $a$ to migrate to cell $t$</td>
<td>69</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of agents in the population</td>
<td>66</td>
</tr>
<tr>
<td>$P(C_i)$</td>
<td>Cultural prototype of community $C_i$</td>
<td>73</td>
</tr>
<tr>
<td>$q$</td>
<td>Number of cultural traits per cultural feature</td>
<td>64</td>
</tr>
<tr>
<td>$q_c$</td>
<td>Critical value of $q$</td>
<td>78</td>
</tr>
<tr>
<td>$s(a,b)$</td>
<td>Weight of social link between agents $a$ and $b$</td>
<td>67</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>$T$</td>
<td>Transition matrix of the Markov chain</td>
<td>70</td>
</tr>
<tr>
<td>$Z$</td>
<td>Normalization constant</td>
<td>67</td>
</tr>
<tr>
<td>$Z'$</td>
<td>Normalization constant</td>
<td>69</td>
</tr>
<tr>
<td><strong>Chapter 5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{\text{quality}}$</td>
<td>Strength of the maxim of quality</td>
<td>108</td>
</tr>
<tr>
<td>$\beta_{\text{quantity}}$</td>
<td>Strength of the maxim of quantity</td>
<td>105</td>
</tr>
<tr>
<td>$\delta_i(t)$</td>
<td>Error term for unit $i$ at tick $t$</td>
<td>99</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Learning rate</td>
<td>99</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Factor for time-averaging of net input</td>
<td>98</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Number of offsprings created in a generation</td>
<td>110</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>Set of hidden units</td>
<td>97</td>
</tr>
<tr>
<td>$\mathcal{I}$</td>
<td>Set of input-output units</td>
<td>97</td>
</tr>
<tr>
<td>$\mathcal{I}_C$</td>
<td>Set of input-output units representing content</td>
<td>98</td>
</tr>
<tr>
<td>$\mathcal{I}_S$</td>
<td>Set of input-output units representing a subject</td>
<td>98</td>
</tr>
<tr>
<td>$\mathcal{I}_T$</td>
<td>Set of input-output units representing a target</td>
<td>98</td>
</tr>
<tr>
<td>$\mathcal{M}$</td>
<td>Set of models given all possible parameter vectors</td>
<td>109</td>
</tr>
<tr>
<td>$\mathcal{N}$</td>
<td>Set of all units in the connectionist network</td>
<td>97</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Number of parameter vectors in the population</td>
<td>109</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Number of parents for each offspring</td>
<td>110</td>
</tr>
<tr>
<td>$\sigma_g$</td>
<td>Random noise for mutation in generation $g$</td>
<td>116</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>Set of possible parameter vectors</td>
<td>109</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Parameter vector</td>
<td>107</td>
</tr>
<tr>
<td>$a$</td>
<td>Agent</td>
<td>113</td>
</tr>
<tr>
<td>$a_i(t)$</td>
<td>Activation of unit $i$ at tick $t$</td>
<td>98</td>
</tr>
<tr>
<td>bias</td>
<td>Bias unit</td>
<td>97</td>
</tr>
<tr>
<td>$d$</td>
<td>Number of input-output units per pool</td>
<td>115</td>
</tr>
<tr>
<td>$D_\theta$</td>
<td>Data generated by model given parameter vector $\theta$</td>
<td>107</td>
</tr>
<tr>
<td>$D_{\text{obs}}$</td>
<td>Observed empirical data</td>
<td>107</td>
</tr>
<tr>
<td>$F$</td>
<td>Fitness function</td>
<td>109</td>
</tr>
<tr>
<td>$f_i$</td>
<td>Activation function of unit $i$</td>
<td>98</td>
</tr>
<tr>
<td>$g_i(t)$</td>
<td>Target value for unit $i$ at tick $t$</td>
<td>98</td>
</tr>
<tr>
<td>$h$</td>
<td>Number of hidden units</td>
<td>115</td>
</tr>
<tr>
<td>$i, j$</td>
<td>Units in the connectionist network</td>
<td>98</td>
</tr>
<tr>
<td>$k$</td>
<td>Number of times that a pattern is learnt</td>
<td>108</td>
</tr>
<tr>
<td>$l$</td>
<td>Number of ticks per interval</td>
<td>102</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of intervals</td>
<td>102</td>
</tr>
<tr>
<td>$M(\theta)$</td>
<td>Model under parameter $\theta$</td>
<td>109</td>
</tr>
</tbody>
</table>
\( n \) Total number of ticks ................................................. 98
\( o \) Overlap (proportion) between Jamayans and Jay pattern .... 113
\( p \) Extent to which stereotype-consistent information is assumed to be shared .............................................. 120
\( q \) Probability for first-hand experience ................................ 124
\( r \) Probability for within-community information exchange .... 124
\( R^2 \) Correlation coefficient .............................................. 116
\( s \) Probability for between-community information exchange .... 124
\( s_i(t) \) Net input of unit \( i \) at tick \( t \) ................................. 98
\( t \) Time tick ...................................................................... 98
\( T(t) \) Set of input-output units with a target at tick \( t \) ........ 98
\( u \) Number of intervals with inputs applied ......................... 107
\( v \) Number of intervals with targets applied ......................... 107
\( w_{ij} \) Weight of link from unit \( j \) to unit \( i \) .......................... 98
\( X(t) \) Set of input-output units with external input at tick \( t \) .... 98
\( x_i(t) \) External input of unit \( i \) at tick \( t \) ................................. 98

**Chapter 6**

\( \alpha, \beta, \gamma \) (Joint) actions ................................................. 144
\( \alpha \) Type of action \( \alpha \) ..................................................... 165
\( \langle H, h \rangle \) Argument with support \( H \) and conclusion \( h \) .... 168
\( A \) Finite, non-empty set of agents ..................................... 144
\( C_i \) Dialectical commitments of agent \( i \) .............................. 169
\( \mathcal{L}_{CG} \) Common Ground Logic .................................. 144
\( \mathcal{L}_{SP} \) SharedPlan Logic .............................................. 165
\( \phi \) Proposition (informal) ................................................... 135
\( \pi \) Valuation function ....................................................... 146
\( \Sigma \) Set of propositional variables in \( \mathcal{L}_{CG} \) or \( \mathcal{L}_{SP} \) .......... 144
\( \text{Eff}(\alpha) \) Effects of \( \alpha \) ..................................................... 165
\( \text{Prec}(\alpha) \) Preconditions of \( \alpha \) ......................................... 165
\( \varphi, \psi, \rho \) Sentences in \( \mathcal{L}_{CG} \) or \( \mathcal{L}_{SP} \) .................. 144
\( a, b, i, j \) Agents ................................................................. 144
\( A_i^c \varphi \) Agent \( i \) accepts \( \varphi \) in context \( c \) ......................... 144
\( B_i \varphi \) Agent \( i \) believes \( \varphi \) ................................................. 144
\( C \) Finite set of contexts ...................................................... 144
\( c \) Context .......................................................... 144
\( CA_c \varphi \) Collective acceptance of \( \varphi \) among agents involved in context \( c \) .... 144
\( CBA(i, \alpha, R_\alpha) \) Agent \( i \) can bring about action \( \alpha \) with recipe \( R_\alpha \) ...................... 166
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CG_\alpha \varphi)</td>
<td>Salient common ground of (\varphi) among the agents that are involved in joint action (\alpha)</td>
</tr>
<tr>
<td>(Conc(e))</td>
<td>Conclusion of argument (e)</td>
</tr>
<tr>
<td>(DCG_\alpha \varphi)</td>
<td>Distributed common ground of (\varphi) among the agents that are involved in joint action (\alpha)</td>
</tr>
<tr>
<td>(Do_G \alpha)</td>
<td>Group (G) does (\alpha)</td>
</tr>
<tr>
<td>(Do_i \alpha)</td>
<td>Agent (i) does (\alpha)</td>
</tr>
<tr>
<td>(e, f)</td>
<td>Arguments</td>
</tr>
<tr>
<td>(EA_c \varphi)</td>
<td>Every agent involved in context (c) accepts (\varphi) in that context</td>
</tr>
<tr>
<td>(EB_c \varphi)</td>
<td>Every agent involved in context (c) believes (\varphi)</td>
</tr>
<tr>
<td>(G, H, I)</td>
<td>Groups of persons/agents</td>
</tr>
<tr>
<td>(I_i \alpha)</td>
<td>Intention of agent (i) to perform (\alpha)</td>
</tr>
<tr>
<td>(I_i \varphi)</td>
<td>Intention of agent (i) that (\varphi)</td>
</tr>
<tr>
<td>(M_{CG})</td>
<td>(L_{CG}) model</td>
</tr>
<tr>
<td>(MB_c \varphi)</td>
<td>Mutual belief of (\varphi) among agents involved in context (c)</td>
</tr>
<tr>
<td>(MB_G \varphi)</td>
<td>Mutual belief of (\varphi) among agents in group (G)</td>
</tr>
<tr>
<td>(p, q)</td>
<td>Propositional variables</td>
</tr>
<tr>
<td>(P_i \alpha)</td>
<td>Potential intention of agent (i) to perform (\alpha)</td>
</tr>
<tr>
<td>(P_i \varphi)</td>
<td>Potential intention of agent (i) that (\varphi)</td>
</tr>
<tr>
<td>(Q)</td>
<td>Finite set of joint actions</td>
</tr>
<tr>
<td>(R_{A}^{i,c})</td>
<td>Acceptance accessibility relation for agent (i) in context (c)</td>
</tr>
<tr>
<td>(R_{B}^{G})</td>
<td>Relation corresponding to the (G_B)-reachability for group (G)</td>
</tr>
<tr>
<td>(R_B^i)</td>
<td>Belief accessibility relation for agent (i)</td>
</tr>
<tr>
<td>(R_{\alpha})</td>
<td>Recipe to perform action (\alpha)</td>
</tr>
<tr>
<td>(R_{CA}^c)</td>
<td>Relation describing collective acceptance in context (c)</td>
</tr>
<tr>
<td>(SP(P_\alpha, G, \alpha))</td>
<td>SharedPlan with name (P_\alpha) of group (G) to perform action (\alpha)</td>
</tr>
<tr>
<td>(Supp(e))</td>
<td>Support of argument (e)</td>
</tr>
<tr>
<td>(u, v, w)</td>
<td>Possible worlds</td>
</tr>
<tr>
<td>(W)</td>
<td>Non-empty set of possible worlds</td>
</tr>
<tr>
<td>(X(c))</td>
<td>Group of agents involved in context (c)</td>
</tr>
<tr>
<td>(Y(\alpha))</td>
<td>Set of groups that are salient in joint action (\alpha)</td>
</tr>
</tbody>
</table>

**Chapter 7**

\(\alpha\) (Joint) action | 207 |
attr\((x)\) Function returning the constant symbol that represents the attribute for the constant symbols `Color` and `Location` | 207 |
Color Constant symbol representing the color attribute of the stimulus | 207 |
\(Do_{(x,y)}(\alpha)\) Agents \(x\) and \(y\) do action \(\alpha\) | 207 |
<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Goal_{(x,y)}(\varphi)$</td>
<td>Goal of agents $x$ and $y$ that $\varphi$.</td>
</tr>
<tr>
<td>Green</td>
<td>Constant symbol representing the attribute “green”.</td>
</tr>
<tr>
<td>$green$</td>
<td>Feature representing the concept “green”.</td>
</tr>
<tr>
<td>index-finger-left</td>
<td>Feature representing the left index finger.</td>
</tr>
<tr>
<td>index-finger-right</td>
<td>Feature representing the right index finger.</td>
</tr>
<tr>
<td>$Int_{(x,y)}(\alpha)$</td>
<td>Intention of agents $x$ and $y$ to execute $\alpha$.</td>
</tr>
<tr>
<td>Left</td>
<td>Constant symbol representing the attribute “left”.</td>
</tr>
<tr>
<td>left</td>
<td>Feature representing the concept “left”.</td>
</tr>
<tr>
<td>Location</td>
<td>Constant symbol representing the location attribute of the stimulus.</td>
</tr>
<tr>
<td>Me</td>
<td>Constant symbol representing this participant.</td>
</tr>
<tr>
<td>nice($x$)</td>
<td>Predicate symbol representing the concept “nice”.</td>
</tr>
<tr>
<td>push-left</td>
<td>Action to push the left button.</td>
</tr>
<tr>
<td>push-right</td>
<td>Action to push the right button.</td>
</tr>
<tr>
<td>pushed-left</td>
<td>Predicate symbol representing that the left button was pushed.</td>
</tr>
<tr>
<td>pushed-right</td>
<td>Predicate symbol representing that the right button was pushed.</td>
</tr>
<tr>
<td>Red</td>
<td>Constant symbol representing the attribute “red”.</td>
</tr>
<tr>
<td>red</td>
<td>Feature representing the concept “red”.</td>
</tr>
<tr>
<td>Right</td>
<td>Constant symbol representing the attribute “right”.</td>
</tr>
<tr>
<td>right</td>
<td>Feature representing the concept “right”.</td>
</tr>
<tr>
<td>Simon-task</td>
<td>Action to execute the Simon task.</td>
</tr>
<tr>
<td>tactile-left</td>
<td>Feature representing a tactile perception on the left side.</td>
</tr>
<tr>
<td>tactile-right</td>
<td>Feature representing a tactile perception on the right side.</td>
</tr>
<tr>
<td>$x, y$</td>
<td>Arbitrary constant symbols.</td>
</tr>
<tr>
<td>$x_1, x_2, x_3$</td>
<td>Features representing the push-left action.</td>
</tr>
<tr>
<td>$y_1, y_2, y_3$</td>
<td>Features representing the push-left action.</td>
</tr>
<tr>
<td>You</td>
<td>Constant symbol representing the other participant.</td>
</tr>
</tbody>
</table>

**Chapter 8**

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Action.</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>Change threshold.</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Learning rate.</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>Feature activation threshold.</td>
</tr>
<tr>
<td>$\gamma_{\text{execution}}$</td>
<td>Increase of input to features due to execution of a motor command.</td>
</tr>
<tr>
<td>$\gamma_{\text{planning}}$</td>
<td>Increase of input to features due to action planning.</td>
</tr>
<tr>
<td>$\gamma_{\text{representation}}$</td>
<td>Increase of input to features due to attitude representation.</td>
</tr>
<tr>
<td>$\gamma_{\text{stimulus}}$</td>
<td>Increase of input to features due to stimulus perception.</td>
</tr>
<tr>
<td>$\langle\alpha, H\rangle$</td>
<td>Intention to perform action $\alpha$ associated with group of agents $H$.</td>
</tr>
</tbody>
</table>
\((\alpha, H, s)\)  Performance goal with action \(\alpha\) to perform, set of agents \(H\) associated with that goal, and annotation \(s\) whether the goal is pursued ........ 228

\((\varphi, F)\)  Primitive action with effect \(\varphi\) and corresponding features \(F\) ...... 227

\((\varphi, G)\)  Complex action with effect \(\varphi\) and a set of subgoals \(G\) ............ 227

\((\varphi, H, s)\)  Achievement or maintenance goal with goal state \(\varphi\), set of agents \(H\) associated with that goal, and annotation \(s\) whether the goal is pursued actively ........................................ 228

\(\mathcal{A}\)  Set of all actions ............................................................. 228

\(\mathcal{A}^C\)  Set of all complex actions ........................................... 227

\(\mathcal{A}^P\)  Set of all primitive actions ........................................... 227

\(\mathcal{G}\)  Set of all possible goals ................................................ 228

\(\mathcal{G}^A\)  Set of all achievement goals ...................................... 228

\(\mathcal{G}^M\)  Set of all maintenance goals ...................................... 228

\(\mathcal{G}^P\)  Set of all performance goals ...................................... 228

\(\mathcal{H}\)  The set of agents considered to be involved in the joint action by the focal agent .............................................................. 228

\(\mathcal{T}\)  Set of all possible intentions ........................................... 228

\(\mathcal{L}\)  A propositional logic .................................................... 227

\(\mathcal{N}\)  Set of all units in the connectionist network ....................... 224

\(\mathcal{T}\)  Set of threshold units .................................................. 224

\(\mathcal{U}\)  Set of units corresponding to stimulus and action effect features .................. 224

\(\mathcal{V}\)  Set of units corresponding to motor command features .................. 224

\(\Omega\)  Activation threshold of threshold units ................................. 225

\(\Sigma\)  Set of propositional variables in \(\mathcal{L}\) .................................. 227

\(\varphi\)  Sentence in \(\mathcal{L}\) ............................................................. 227

\(\mathcal{A}\)  Set of actions available to the agent .................................. 228

\(a_i(t)\)  Activation of unit \(i\) at tick \(t\) .............................................. 224

\(D\)  Activation decay .............................................................. 226

\(d\)  Feature representing the concept “red” ................................ 235

\(E\)  Excitation level of an input ................................................ 225

\(e\)  Intention .......................................................... 229

\(F\)  Set of features .............................................................. 227

\(f\)  Activation function ........................................................ 224

\(G\)  Set of the current goals of the agent ................................... 228

\(g\)  Goal .......................................................... 229

\textit{Green-Stimulus}  Proposition representing a green stimulus .......... 234

\(H\)  A subset of the agents in \(\mathcal{H}\) ............................................. 228
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$</td>
<td>Set of the current intentions of the agent</td>
</tr>
<tr>
<td>$i, j$</td>
<td>Units in the connectionist network</td>
</tr>
<tr>
<td>$k$</td>
<td>Number of iterations for which feature associations are learnt</td>
</tr>
<tr>
<td>$KB$</td>
<td>Knowledge/Belief base</td>
</tr>
<tr>
<td>$l$</td>
<td>Feature representing the concept “left”</td>
</tr>
<tr>
<td>$Left$-Proposition</td>
<td>Proposition representing a stimulus on the left side</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of additional motor command units</td>
</tr>
<tr>
<td>$n$</td>
<td>Feature representing the concept “green”</td>
</tr>
<tr>
<td>$p, q$</td>
<td>Propositional variables</td>
</tr>
<tr>
<td>$push$-left</td>
<td>Action to push the left button</td>
</tr>
<tr>
<td>$push$-right</td>
<td>Action to push the right button</td>
</tr>
<tr>
<td>$Pushed$-left</td>
<td>Proposition representing that the left button was pushed</td>
</tr>
<tr>
<td>$Pushed$-right</td>
<td>Proposition representing that the right button was pushed</td>
</tr>
<tr>
<td>$q_i$</td>
<td>Threshold units</td>
</tr>
<tr>
<td>$r$</td>
<td>Feature representing the concept “right”</td>
</tr>
<tr>
<td>$Red$-Proposition</td>
<td>Proposition representing a red stimulus</td>
</tr>
<tr>
<td>$Right$-Proposition</td>
<td>Proposition representing a stimulus on the right side</td>
</tr>
<tr>
<td>$S$</td>
<td>Set of sentences in $L$</td>
</tr>
<tr>
<td>$s$</td>
<td>Feature representing the concept of a stimulus</td>
</tr>
<tr>
<td>$s_i(t)$</td>
<td>Net input of unit $i$ at tick $t$</td>
</tr>
<tr>
<td>$SimonTask$</td>
<td>Complex action representing the Simon task</td>
</tr>
<tr>
<td>$t$</td>
<td>Time tick</td>
</tr>
<tr>
<td>$u_1, u_2$</td>
<td>Features representing the perception of pushing the left button</td>
</tr>
<tr>
<td>$v_1, v_2$</td>
<td>Features representing the perception of pushing the right button</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>Weight of link from unit $j$ to unit $i$</td>
</tr>
<tr>
<td>$x_i, y_i, z_i$</td>
<td>Features representing motor commands</td>
</tr>
<tr>
<td>$x_i(t)$</td>
<td>External input of unit $i$ at tick $t$</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This thesis investigates the computational modeling of the transmission of cultural information and of the resulting emergent cultural dynamics. Cultural information can be conceptualized as any information that is socially, as opposed to genetically, transmitted between people (Kashima and Gelfand, 2011). Such information includes amongst others opinions, norms, practices, and stereotypes. The culture of a society is then at least partly determined by the particular distribution of cultural information among its members. In representing the “collective programming of the mind” (Hofstede, 1984), culture forms part of the backdrop against which any human social interaction takes place. Not only does culture prescribe how to behave in particular situations but it also affects how our cognitive system processes, stores, and retrieves information (DiMaggio, 1997).

Likewise, cultural information is frequently materialized in physical form during social interactions and thereby transmitted between individuals (Kashima, 2008). For instance, people create, communicate, and recognize symbols and they memorize and reproduce stories that they hear or read. Therefore, culture is by no means static but undergoes constant revision and modification while cultural information is reproduced, transmitted, and perceived. The view that cultural transmission is critical to the change of culture is shared by a number of prominent thinkers (Kashima, 2008), including Dawkins (1976), Cavalli-Sforza and Feldman (1981), and Boyd and Richerson (1985).

The analysis above suggests that culture is a macro-level property of complex social systems that emerges from the behavior of individuals during micro-level social interactions, driven by their mental representations and processes. We term this process, which underlies the formation, maintenance, and change of culture, cultural dynamics. It is clear that a comprehensive study of cultural dynamics has to consider the bi-directional interaction between both levels, namely the bottom-up effect of cultural transmission on the formation, maintenance, and change of culture as well as the top-down influence of culture on cultural
Chapter 1. Introduction

transmission in social interactions. This suggests a two-component research program: First, a detailed understanding of cultural transmission during social interactions is required, accounting also for the relationship of micro-level cultural transmission with macro-level properties of culture and the role of mental representations and processes in micro-level cultural transmission. Second, the effect of cultural transmission on a larger scale needs to be considered, based on the results of the first component.

The first component of this research program has traditionally been addressed by fields such as psychology and cognitive science. These fields can build on proven research methodologies for the study of human social behavior in temporally and spatially constrained micro-level settings. The second component of the research agenda requires a methodology that can cope with the bi-directional interaction between micro and macro level, and with spatially and temporally extended settings.

We adopt the perspective that computational modeling can contribute to both parts of this agenda. Computational models are based on languages that can express appropriately the mechanisms governing social systems and dynamics (Sawyer, 2004; Thagard, 2008, 2009). These include cognitive mechanisms at the individual level, mechanisms underlying the interactions between individuals, and mechanisms concerning the dynamics of social groups. By virtue of providing precise descriptions, computational models can contribute to the refinement of models and theories that address the first component of the suggested research agenda; and they provide a promising opportunity for considering the bi-directional interaction between micro and macro level required by the second component. This makes computational modeling an intuitively and methodologically suitable tool for the study of social dynamics in general and cultural dynamics in particular. In the context of social dynamics more generally, computational modeling and simulations have gained increasing acceptance as a valuable tool (see for example Gilbert, 2007; Goldstone and Janssen, 2005; Smith and Conrey, 2007).

This thesis investigates how a particular social-psychological model of the micro-level transmission of cultural information—the grounding model of cultural transmission by Kashima et al. (2008)—can be represented in computational models and how this effort contributes to the understanding of this particular model and to cultural dynamics in general. Thereby, this thesis takes up and contributes to the results of the first component of the outlined research agenda in order to contribute to the second component.

In Section 1.1, we provide a brief overview on the grounding model of cultural transmission. In Section 1.2, we comment briefly on existing computational models of cultural dynamics. In Section 1.3, we present the research questions to be addressed by this thesis and finally outline the structure of this thesis in Section 1.4.
1.1 The Grounding Model of Cultural Transmission

The grounding model of cultural transmission by Kashima et al. (2008) emphasizes the role of complex social interactions in the transmission of cultural information. According to this model, the participants of a social interaction are involved in a joint action (or joint activity) and rely on their common ground—the information that they share and also assume to share (Clark, 1996c)—when they coordinate towards the achievement of this activity. Progress towards the joint activity is made not only by fulfilling tasks but also by communicating in order to achieve and maintain a shared understanding of the joint activity and of the world. This communication process adds to the common ground of the participants in the joint action (whom we also call coactors) and is therefore called grounding.

Kashima et al. distinguish between three different “types” of common ground that determine the information shared in an interaction: Context-specific common ground is the shared information relevant to the current joint activity. Personal common ground is the information shared by the coactors due to previous common experiences. Communal common ground is the information that coactors share because of their common community memberships. For example, the name of the Australian prime minister is part of the communal common ground of the group of Australians (with exceptions). The important part is that the communal common ground of a particular community determines a major part of this community’s culture; and communal common ground and therefore culture play a role in social interactions by virtue of being part of the common ground of the coactors. This reflects the top-down influence of culture on social interactions. At the same time, the grounding process in an interaction does not only add to context-specific and personal common ground but also to communal common ground. Therefore grounding plays an important role in the formation, maintenance, and change of culture. This reflects the bottom-up influence of social interactions on culture.

In summary, the grounding model of cultural transmission provides a detailed account of micro-level cultural transmission. The model also postulates ways in which micro-level cultural transmission is driven by mental representations and processes and interacts bi-directionally with macro-level properties of culture. Thus, the grounding model of cultural transmission contributes to the first component of the research program outlined above. Adopting this model to underpin the development of computational models of cultural dynamics therefore creates an opportunity to contribute to the second component of this research program. In addition, computational models can help refine the grounding model of cultural transmission. This amounts to a contribution to the first component of the research program outlined above, namely a contribution to the understanding of cultural transmission in micro-level social interactions.
Chapter 1. Introduction

We provide a more detailed review of the grounding model of cultural transmission in Chapter 2. Following the grounding model of cultural transmission, this thesis adheres to the view that cultural transmission is mainly a consequence of complex social interactions and that the culture of a community is largely described by its communal common ground. In consequence, we consider other perspectives on culture and cultural dynamics as outside of the scope of this thesis. These perspectives include, for example:

- A view that considers the culture of individuals in a group to be described by a point in a multi-dimensional space of factors such as the level of individualism or collectivism, the level of uncertainty avoidance, and power distance (e.g. Hofstede, 1984; Hofstede et al., 2010; Triandis, 1989).

- A view of culture as “a collection of concrete everyday practices that occur in everyday life” that is according to Kashima (2001) advanced, for example, by Cole (1996); Rogoff (1990); Wertsch (1991).

- A view of culture that is manifested in the symbols, heroes, rituals, and values of a group (Hofstede et al., 2010).

- The views on cultural dynamics by Boyd and Richerson (1985); Cavalli-Sforza and Feldman (1981); Dawkins (1976), which conceptualize cultural transmission as a simple process of replication or mutation and not as the product of complex social interactions (see also Section 2.1.2).

1.2 Computational Models of Cultural Dynamics

As indicated by the introduction of the term “culture” above, various types of information can be considered cultural. Accordingly, a variety of computational models of cultural dynamics have been produced in various fields of research. This includes models of the emergence of social norms, often framed in the context of evolutionary game theory (e.g. Axelrod, 1984, 1986; Nowak, 2006; Szabó and Fáth, 2007). In these models, individuals in a population repeatedly play a certain game against each other and change their strategies according to a given rule. The typical question is under which conditions a population evolves to be dominated by a particular strategy. Other models of cultural dynamics are models of the diffusion of opinions (e.g. Hegselmann and Krause, 2005; Sznajd-Weron and Sznajd, 2000), the emergence of language (e.g. Nowak and Krakauer, 1999), or the emergence of culture in general (e.g. Axelrod, 1997b).

A complete review of this literature is beyond the scope of this thesis. However, we can summarize that the commonality of most of these models is that they do not adhere rigorously
to theories that describe in detail the micro-level mechanisms that drive cultural transmission during social interactions. Instead, cultural transmission is typically conceived as a simple contact-process with simple rules governing the transmission of cultural information between interacting individuals. By no means do we claim that these models do not have any value. However, the focus of this thesis lies on drawing on the grounding model of cultural transmission in the construction of computational models of cultural dynamics; and the grounding model of cultural transmission does provide detailed statements on the mechanisms of cultural transmission at the micro level.

Exceptions that build more rigorously on detailed accounts of cultural transmission do exist. For example, there is a long tradition in social psychology of creating computational models of stereotyping processes, i.e. those processes that are involved in the cognitive storage, retrieval, and application of stereotypes. In recent years, this research agenda has been augmented by models that also represent the communication of stereotypes and stereotype-relevant information (Van Overwalle and Heylighen, 2006; Van Rooy, 2009).

There is also an increasing interest in considering the factor of culture more generally in models of human behavior, for example in the case of training simulations for the military (Solomon et al., 2008; Taylor et al., 2007), the simulation of trade negotiations (Hofstede et al., 2011), or the influence of culture on practical reasoning and on behavior more generally (Dignum et al., 2008). While these models recognize the influence of culture on social interactions, they do not aim at representing the transmission of cultural information and the emergence of culture. In that, they lack support for the bottom-up direction from micro to macro and are not strictly models of cultural dynamics.

1.3 Research Questions and Contributions

This thesis aims specifically at building on the grounding model of cultural transmission and its postulates in the development of computational models of cultural dynamics. The following research questions are addressed:

**RQ1** How can the grounding model of cultural transmission be translated into computational models of cultural transmission and dynamics?

**RQ2** How can these computational models contribute to the refinement of the grounding model of cultural transmission and to the understanding of cultural dynamics?

With regard to **RQ1**, this thesis claims that the grounding model of cultural transmission implies three essential building blocks for computational models of cultural transmission: a model of the joint activities people are engaged in, a model of their common ground, and a model of their task-incidental grounding dialogue and its interrelation with their
common ground. We demonstrate how representations of joint activity, grounding, and common ground can be represented by computational models of cultural dynamics. This demonstration relies on a series of computational models as examples. However, our contribution is not just illustrative. The models presented in this thesis can serve as input to other computational models of cultural dynamics. As we shall see, these models have been developed with different purposes in mind. As a corollary to answering RQ1, we will discuss, when possible, the inherent complexity of these models, their computational complexity, their empirical adequacy, and the faithfulness with which they adhere to the grounding model of cultural transmission. All of these factors influence the possible benefits of these models.

With regard to RQ2, this thesis holds that computational modeling can contribute to the advancement of the grounding model of cultural transmission and to the understanding of cultural dynamics in two ways:

1. Formal modeling of verbal-conceptual models almost always highlights ambiguities or gaps in the models and theories relied on. Resolving these issues in cooperation with domain experts does not only benefit the actual modeling effort but also the refinement of these other models and theories. This in itself is clearly a positive result. When appropriate, we describe novel questions that the presented models pose about the grounding model of cultural transmission. This is our first contribution towards answering RQ2.

2. Computational models and formal models more generally allow for the rigorous study of the implications of the mechanisms they represent. Computational models, in particular, lend themselves to simulations that substitute or complement—with limitations—empirical experiments, which might be too costly or time-consuming, ethically questionable, or otherwise unfeasible. The grounding model of cultural transmission is defined at the level of joint activities. The obvious possibility for simulations within the context of this thesis is therefore the exploration of how cultural transmission in joint activities according to this model affects cultural dynamics. Two of the models we present have been developed with the intention to answer questions of this kind. Other types of formal models can be studied analytically, e.g. formal logics models can be analyzed in terms of their deductive consequences. We also present models that belong to this second category. Providing results about the implications of the grounding model of cultural transmission is the second contribution we make towards addressing RQ2.
1.4 Thesis Structure

In brief, Chapters 2 and 3 discuss the background material on culture, cultural dynamics, and computational modeling. Chapters 4 to 8 present a series of computational models of cultural dynamics. Chapter 9 concludes this thesis. We provide a more detailed positioning of Chapters 4 to 8 in Section 3.4 after the necessary background material has been introduced. For now, the following provides a less detailed overview of this thesis:

**Chapter 2** discusses the issues surrounding the emergence of culture as a macro-level property from micro-level social interactions. That chapter also expands on the idea that a two-component research program is necessary to study the bi-directional interaction between culture at the macro level and cultural transmission during micro-level social interactions that are driven by the participants’ mental representations and processes. In addition, the grounding model of cultural transmission is described in more detail. Chapter 2 therefore provides the background information on culture, cultural transmission, and cultural dynamics.

**Chapter 3** elaborates on the role of models in science and on the benefit of computational models in the study of cultural dynamics. The discussion in that chapter takes up the argument that computational models are the appropriate language for representing and reasoning about social dynamics in general and cultural dynamics in particular. Agent-based modeling—a particular kind of computational modeling—is advanced as the most appropriate framework for the study of cultural dynamics. Agent-based modeling allows the direct and intuitive representation of actors and their interactions within their wider social contexts. However, we also emphasize the benefit of cognitive models in the development of agent-based models and in the study of cultural transmission more generally. Our perspective on the process of developing agent-based models is described, and limits and pitfalls of creating and using such models are outlined. This more detailed discussion of the subject matter provides the technical background for the remainder of the thesis and feeds into a more precise positioning of this thesis than possible at this stage.

**Chapter 4** introduces an agent-based model of the co-evolution of cultural, social, and physical space or respectively cultures, social networks, and geographical locations. The phenomena represented by this model include (a) the effect of cultural similarities and differences as materialized during cultural transmission on the quality of social links and on the migration behavior of individuals, and (b) the influence of social links and physical distance on the likelihood of interaction and hence cultural transmission between individuals. We explore the dynamics of this model across varying parameters
Chapter 1. Introduction

and thereby contribute to the study of cultural dynamics. This model offers an idealistic representation of the grounding model of cultural transmission insofar that only a simple model of common ground and grounding is assumed. Therefore, this model cannot be considered to contribute to the refinement of the grounding model of cultural transmission. However, the grounding model is used to represent the effect of cultural transmission on social and physical space, which is a core assumption of the model introduced in that chapter.

Chapter 5 introduces an agent-based model of the communication of stereotype-relevant information as a particular type of cultural information. The model is largely based on an empirical study by Lyons and Kashima (2003) on the transmission of stereotype-relevant information through chains of coactors similar to the Chinese whispers game. Lyons and Kashima found that the extent to which stereotype-consistent or stereotype-inconsistent information is transmitted through such a chain depends on the actual and perceived common ground of that stereotype.

The agents in our model are able to acquire stereotype-relevant information first- and second-hand, while taking into consideration the common ground they share with their communication partners. The agent model accounts for the integrated storage and reproduction of information about groups and individual group members, which is crucial in stereotyping processes. The rigorous representation of these aspects distinguishes our model from previous work. We thereby contribute in particular to the development of models that allow investigating the diffusion of stereotype-relevant information depending on social network structure. Our model offers the possibility to account not only for network composition (i.e. the existence of social links) but also network configuration (e.g. the common ground associated with the individuals adjacent to a particular link, their group membership, and the group membership of the third party that the communication is about).

The model is calibrated to one data set obtained by Lyons and Kashima and this estimated model is validated using another data set. We then explore the dynamics of the model with a larger population of communicating agents, investigating the role of the underlying social network in the diffusion of stereotype-relevant information. The model contributes to a refinement of the grounding model of cultural transmission because it provides evidence that the mechanisms postulated by that model can indeed give rise to the patterns observed by Lyons and Kashima.

Chapter 6 provides a detailed semi-formal description of the grounding model of cultural transmission. The core of that chapter is a formal description of common ground and its role in social interactions in terms of a modal logic. We study the properties of this
1.4. Thesis Structure

logic. In addition, this chapter presents a semi-formal description of how our model of common ground would interact with formal models of joint activities and of the grounding process. This chapter contributes to a refinement of the grounding model of cultural transmission by elaborating on the epistemic status and the properties of common ground and by providing a more precise account of the interaction between common ground, grounding, and joint activities. This model lends itself to informing the development of other computational models of cultural transmission and dynamics.

Chapter 7 presents a detailed semi-formal architecture of joint action based on the observation that an individual’s behavior during joint actions is determined by interacting higher- and lower-level mental processes. Throughout that chapter, a particular experimental task provides the backdrop against which the architecture is developed. Chapter 7 contributes to the refinement of the grounding model of cultural transmission insofar that it expands on the notion of joint action, which forms a central component of the grounding model of cultural transmission. The chapter contributes to the study of cultural dynamics by discussing the consequences of this model on the implicit transmission of embodied culture and on the symbolic transmission of culture, which is addressed by the grounding model of cultural transmission.

Chapter 8 describes an implementation of the architecture presented in Chapter 7. The implementation is evaluated with respect to the mentioned experimental task. By showing that the implementation is able to reproduce the results of that experimental task qualitatively, that chapter underlines the value of the proposed architecture.

Chapter 9 closes this thesis with an outline of contributions and directions for future work.

In summary, this thesis makes the following concrete contributions:

- An agent-based model of the co-evolution of cultures, social networks, and geographical locations and an analysis of this model through computational simulations.

- An agent-based model of the communication of stereotype-relevant information, its calibration and validation against empirical data, and, using computational simulations, an analysis of the model’s dynamics given the communication of stereotype-relevant information in social network structures.

- A logic-based semi-formal description of the grounding model of cultural transmission, and a modal logic formalization and analysis of common ground.

- A semi-formal architecture of the interaction between higher-and lower-level coordination mechanisms in joint action, an implementation of this architecture, and a demonstration of its viability through the reproduction of empirical data.
Chapter 2

Culture

This chapter elaborates on the concept of culture and on the concept of cultural dynamics as the process that yields the formation, maintenance, and change of culture. In particular, we describe culture as a macro-level property of a complex social system and cultural dynamics as the bi-directional interaction between culture and micro-level social interactions. We revisit the argument that the study of cultural dynamics requires a two-component research program: an analysis of micro-level cultural transmission and an analysis of how this transmission affects culture at the macro level. We also discuss in more detail the grounding model of cultural transmission, which provides a description of the mechanisms that govern micro-level cultural transmission. This chapter introduces the background material about culture, cultural transmission, and cultural dynamics that is necessary for this thesis.

Section 2.1 elaborates on our understanding of culture and cultural dynamics. Section 2.2 discusses the grounding model of cultural transmission and Section 2.3 concludes the chapter.

2.1 Culture and Cultural Dynamics

In the previous chapter, we have argued that the study of cultural dynamics requires the simultaneous consideration of micro and macro level. In this section, we expand on this argument. We briefly discuss how the interaction between micro and macro level has been accounted for in social scientific research. Then, we discuss how the different levels of analysis need to be considered in the study of cultural dynamics. Finally, we relate cultural dynamics to a complex social system.
2.1.1 The Micro-Macro Problem

The existence of two different levels of analysis in social science was already evident in the work of the founders of modern sociology, e.g. Durkheim (1982) and Weber (1978). The realm of institutions, norms, and structures has been termed the “macro level”. The realm of actors, individual actions, desires, beliefs, etc. has been termed the “micro level” (Salgado and Gilbert, 2008). A central debate in sociology is which level needs to be addressed to generate knowledge about social systems. Methodological individualism argues in favor of analyzing how social phenomena arise from individual action (Weber, 1978). Methodological collectivism postulates that macro-level properties have causal powers on individual action and therefore need to be the focus of the analysis (Durkheim, 1982).

These two perspectives, however, fall short of considering the micro-macro problem as one that spans both levels (Goldspink and Kay, 2009). The micro-macro problem is “the relationship between the actions of individuals and resulting social structures and the reciprocal constraint those structures place on individual agency” (Goldspink and Kay, 2009). This bi-directional interaction between micro and macro level was prominently described by Coleman (1990). Coleman suggests that macro properties produce conditions on the micro level, which motivate or constrain individual behavior (downward causation). The entanglement of the behavior of multiple individuals, in turn, has a causal effect on the macro level (upward causation). This idea is illustrated by Coleman’s “boat” (Coleman, 1990, p. 8), as shown in Figure 2.1.

There is still no consensus as to what is the appropriate perspective on the micro-macro problem (Goldspink and Kay, 2009) and whether the notion of levels is helpful at all or whether there are more levels to be considered (Troitzsch, 2009b). Therefore, we do not adhere to any particular account and rely on the idea of different levels of analysis mainly to
guide our discussion and structure this thesis. More elaborate discussions of this topic are provided by Goldspink and Kay (2009); Salgado and Gilbert (2008); Sawyer (2009).

2.1.2 Levels of Analysis in Cultural Dynamics

Traditionally, culture has been studied from two different perspectives: as a relatively stable system of meaning shared by a group of people, and as processes of meaning making that people engage in (Kashima, 2000). The former macro-level perspective underlies most cross-cultural comparative research and provides an analysis in terms of concepts such as cultural groups. The understanding of the culture of a society in relation to the macro-level distribution of information among its members, which we adopted in Chapter 1, falls into this category. The latter micro-level perspective facilitates the understanding of the situated expression and acquisition of cultural information in social interactions.

However, ultimately it is micro-level interactions between individuals that give rise to the macro-level distribution of cultural information and cultural groups and institutions. Likewise, the macro-level distribution of information affects how, when, and whether this information is communicated at all in a given situation. This idea was already reflected by Coleman’s “boat” as described in the previous subsection (Figure 2.1) and is represented again in Figure 2.2. Studying the bi-directional interaction between micro and macro level is necessary for understanding cultural dynamics, namely, the formation, maintenance, and transformation of culture over time. For example, stereotypes as an instance of culturally transmitted information are frequently invoked during social interactions. They enable sense-making of other individuals’ behavior and leverage the communication about others. At the same time, the use of stereotypical information during social interactions affords the cultural transmission of that stereotype and its maintenance (see Chapter 5).

A prominent contemporary meta-theoretical approach that attempts to examine cultural dynamics is neo-diffusionism (Kashima, 2008; e.g. Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981; Dawkins, 1976), the view that takes culture as a collection of non-genetic information prevalent in a group of people, which is socially transmitted from one generation to the next. In this view, the critical issue is the process of social transmission of information between individuals. Frequently transmitted information at the micro level becomes shared within a group, and is therefore likely to become part of its culture at the macro level. While these approaches have yielded valuable insights into cultural dynamics, most of them tend to simplify the process of cultural transmission as primarily one of replication or mutation.

However, a closer look at the actual process of cultural transmission reveals that it is mainly a consequence of joint activities in complex social interactions. This is the main assumption of the grounding model of cultural transmission by Kashima et al. (2008), which
Chapter 2. Culture

2.1.3 Cultural Dynamics as a Complex System

A closer inspection of the interactions between macro, micro, and nano level shows that we are in fact dealing with a complex system (Manson, 2001). The key property of a complex system is that its dynamics arise from relationships and interactions between its constituents. We find this property at the macro and micro level of cultural dynamics: The dynamics of the macro level arise from the micro-level interactions between individuals, and the micro-level
behavior of individuals arises from nano-level processes operating on mental representations. Four other properties often attributed to complex systems (Manson, 2001) can easily be identified in social systems in general and cultural dynamics in particular: non-linearity, feedback, emergence, and self-organization.

**Non-linearity** Mental processes operating on mental representations at the nano-level and social interactions relating individuals at the micro level obviously follow highly non-linear dynamics.

**Feedback** Figure 2.2 indicates the possibility for feedback between the different levels but certainly we can find other feedback effects within the same level. For example, there is plenty of opportunity for feedback effects to occur in repeated interactions between the same individuals.

**Emergence** Macro-level and micro-level properties emerge from micro-level and nano-level processes respectively. That is, these properties cannot be described by properties at the lower level nor can they be analyzed by measurements that apply to the lower level (Gilbert, 2007). Even with complete knowledge about lower-level properties it is impossible or at least practically impossible to predict the outcomes at the higher level (Goldspink and Kay, 2009; Sawyer, 2009). An important observation is that humans are capable of recognizing emergent properties, communicating about them, and considering them in their reasoning processes. We can identify a norm as a norm and a tradition as a tradition, both possibly having different implications on our attitude towards this structure. This ability causes the emergence in a social system to be of a different order and quality than the emergence in a physical system, and it contributes to the system’s complexity (Castelfranchi, 1998; Goldspink and Kay, 2009). Note that the epistemological and ontological status of emergence is by no means clear (Salgado and Gilbert, 2008).

**Self-organization** Self-organization is a result of the circular dependencies between the different levels (Goldspink and Kay, 2009; Salgado and Gilbert, 2008). Higher-level properties constrain and enable lower-level properties but at the same time emerge from these lower-level properties. For example, we will see in Chapter 5 how the perception that a particular stereotype is shared increases the exchange of stereotype-consistent information, which, in turn, further contributes to that perception and hence the maintenance of that stereotype.

Clearly, this analysis implies that the study of cultural dynamics in particular and social systems in general is an inherently intricate one. As outlined in the introduction, we believe
that a two-component research program is necessary to advance our understanding of cultural
dynamics. First, micro-level cultural transmission and obviously its relation with macro-level
properties and nano-level mental processes needs to be understood. Second, the effect of
cultural transmission on larger scales needs to be studied. As we shall see in the next section,
the grounding model of cultural transmission contributes to the first part of this agenda. The
distinct feature of that model is that it does comment on the relation between micro-level
cultural transmission and nano and macro level.

2.2 The Grounding Model of Cultural Transmission

The previous section has put forward an image of culture as a macro-level property of
complex social systems that emerges from the interactions between individuals at the micro
level, driven by their nano-level mental processes. Kashima et al. (2008) claim that the
micro-level social interactions that host cultural transmission are typically everyday joint
activities. Their grounding model of cultural transmission is based on the assumption that
cultural transmission depends on (a) the purpose and context of the joint activity that coactors
are involved in, (b) the information that they share and assume to share, and (c) the process
by which information is exchanged in order to carry out the joint activity. We describe the
grounding model of cultural transmission in the following. Later in this section, we expand
on the discussion of the model’s main components, provide some examples that illustrate
the ideas behind this model, and finally identify the main premises of the model and the
implications that can be derived from those premises.

For the development of computational models of cultural dynamics, the grounding model
of cultural transmission provides a detailed description of the mechanism by which micro-
level interactions drive cultural transmission. The grounding model of cultural transmission
also comments on the relation between micro level on the one hand and macro and nano level
on the other hand. In this section, we describe the concepts and processes of the grounding
model of cultural transmission that are relevant to this thesis.

From the perspective of the grounding model of cultural transmission, a joint activity
can be as simple as having small-talk with a co-worker but it can also consist of multiple,
hierarchically organized sub-activities. Two or more actors are involved in a joint activity,
contributing by their individual actions which are regulated by their intentions. However,
these intentions obviously need to be coordinated. Philosophers have theorized about joint
activities in terms of joint intentions or we-intentions, for example the intention that we write
a paper together, distributing writing work but aligning content, style and language, and
proofreading each other’s contributions. Some argue that joint intentions can be reduced
to individual intentions and mutual beliefs (Bratman, 1992) while others disagree (Gilbert,
2.2. The Grounding Model of Cultural Transmission

1992; Searle, 1990; Tuomela, 2006). We do not endorse one view or the other, but note that participants need to properly intend to perform their parts of the joint activity in coordination with their partners.

In order to coordinate, participants need to communicate and align their beliefs about the information that is relevant to the successful execution of the activity. Building on Clark’s theory of grounding in language use (1996c), Kashima et al. apply the term “grounding” to describe this alignment process. Grounding thus is a subordinate process to the participants’ joint activity. We sometimes use the verb “to ground” to denote the process of adding information to common ground.

Clark postulates that during communication interlocutors rely heavily on their common ground. He (1996c, p. 93) defines common ground of two coactors as “the sum of their mutual, common, or joint knowledge, beliefs, and suppositions”. When interlocutors begin to engage in their joint activity, they start with a certain initial common ground, which is due to their previous shared experience (personal common ground) or their common group membership (communal common ground). Subsequent grounding during the interaction adds new information to their common ground. Grounding consists of at least two different phases: (1) the presentation of some information, say, a proposition $\phi$ by a speaker, and (2) the acceptance by the listener, which signals that the speaker’s intent with regard to the presentation of $\phi$ has been understood.

The acceptance by the listener can be a simple nod or any other non-verbal confirmation. However, often the listener’s acceptance is in fact itself the presentation of a proposition that is relevant to the speaker’s presentation. This proposition then needs to be accepted by the first speaker. Thus, in principle, the presentation-acceptance pair can continue indefinitely; however, when the interlocutors regard a certain proposition as mutually understood to the extent sufficient for the current purpose (as defined by the joint activity), they stop the presentation-acceptance exchange, and treat it as common ground. As part of this exchange, both interlocutors can request clarifications from each other when they cannot understand their partner sufficiently for the current purpose. The proposition eventually accepted by all interlocutors is added to their common ground. Thus, common ground is constructed collaboratively, and the proposition eventually added to common ground is not necessarily the proposition $\phi$ that the speaker intended to communicate originally.

Clark (1996c) developed his model of grounding in order to explain language use at the utterance level. That is, his theory is about how interlocutors establish a sufficient basis that they have understood an utterance, but not about whether they mutually agree on the communicated content or not. It is the latter version of the grounding process that is in our opinion most important for manipulating common ground. We will return soon to a
discussion of grounding as a process that establishes mutual agreement or acceptance of communicated information.

Kashima et al. call the common ground created during a particular joint activity context-specific common ground. Context-specific common ground is indexed by the time and location of the activity as well as the identities of the participants. However, context-specific common ground can be generalized temporally, spatially, or socially: Interlocutors usually assume that grounded information constitutes context-specific common ground for the next interaction, and should be mutually accessible again if the interaction continues at another location or time, thus becoming part of personal common ground. This links context-specific to personal common ground.

Likewise, interlocutors can infer that their context-specific common ground is actually shared by a wider community, i.e. that this information is part of that community’s communal common ground. This process of generalization links context-specific to communal common ground. In that, we use the term “communal common ground” not only to denote the information that is shared by the participants of an interaction because of their common ground memberships, but also to denote the information that is shared within a wider community. Throughout this thesis we ensure that the intended meaning of this term is clear from the context.

Kashima et al. assume that not only target information is grounded during an interaction, but also presuppositional and relational information. Target information is information that is explicitly grounded; presuppositional information is information presupposed by the target information; and relational information concerns the social relationship between the interlocutors or with other individuals or groups implied by the target information. This information is individually inferred to be part of common ground. Hence, interlocutors might come to different views of what their common ground is and we need to distinguish between actual common ground and perceived common ground. In fact, according to Kashima et al. in common ground is only that information which is in actual and perceived common ground. We will return in Chapter 6 to a more comprehensive analysis of the difference between actual and perceived common ground.

From the perspective of the grounding model of cultural transmission the culture of a social group is largely determined by the communal common ground of that group. That is, a large part of culture is described by the information actually shared and assumed to be shared within that group. Therefore, grounding during joint activities plays a crucial role in establishing and maintaining the culture of social groups. Likewise, communal common ground or in fact culture plays a role in interactions as the information that is assumed to be shared by virtue of common group membership.
2.2. The Grounding Model of Cultural Transmission

We emphasize here that we consider primarily unstructured groups as the holders of communal common ground. Unstructured groups are, for example, the participants of a casual interaction, the group of all medical doctors, all Texans, all stamp collectors, or all inhabitants of Melbourne. In the case of unstructured groups, communal common ground emerges without being determined by a ruling minority or formal group decision-making.

Kashima et al. assume that although the joint activity largely dictates which information needs to be communicated, and therefore determines epistemic goals, the joint activity typically implies certain relational goals as well, that is, goals of regulating social relationships among the interlocutors. Epistemic goals are managed by generic strategies such as Grice’s (1975) communication maxims or Sperber and Wilson’s (1995) principle of relevance. Relational goals can be managed, for example, by Levinson’s (1983) politeness rules. However, these goals might be incompatible at times, thus posing a dilemma: Sharing information that is accurate (epistemic goal) but inconsistent with common ground might require more effort during the grounding process. This might have an adverse effect on the relationship between the interlocutors (Clark and Kashima, 2007). Modifying such information so that it is more consistent with common ground might lead to a smoother grounding process and might be socially-connective (satisfying a relational goal), but may also amount to the dropping of some relevant information. Interlocutors need to manage what they communicate in order to achieve these possibly competing goals. In general, a larger common ground typically implies that grounding new information is easier (Wu and Keysar, 2007).

2.2.1 Common Ground and Grounding Revisited

To further elaborate on the characteristics of the grounding model of cultural transmission, we expand on the notions of common ground and grounding. We deem this critical to a full understanding of the model. In fact, we will elaborate even more on this analysis in Chapter 6, when we work towards a comprehensive formal account of the grounding model of cultural transmission.

Common Ground Common ground is the information that interlocutors presuppose or take for granted in a conversation. Stalnaker (2002) traces the notion of common ground back to Paul Grice’s William James lectures in 1967. Various epistemic states such as common/shared/mutual belief or knowledge have been used to describe common ground, which has contributed to a confusion around the term (Lee, 2001). As mentioned before, Clark defines the interlocutors’ common ground as “the sum of their mutual, common, or joint knowledge, beliefs, and suppositions” (Clark, 1996c, p. 93). Lee, however, emphasizes that each of these epistemic states has different implications for the interlocutors’ cognition and calls for a crisper definition. He rules out mutual belief or knowledge as psychologically
implausible because of the infinite regression these expressions imply (see the more detailed discussion in Chapter 6).

Others argue that belief or knowledge are inappropriate descriptions for common ground altogether, calling on the observation that what interlocutors presuppose is not necessarily what they mutually believe. Most prominently, Stalnaker (2002) postulates that common ground is a state of collective acceptance, in particular that a proposition is common ground of a group if everyone in that group accepts this proposition and that acceptance is mutually believed. Note that acceptance here denotes an epistemic state that is conceptually different from the acceptance of a presentation considered by Clark. Interlocutors can accept a proposition for a particular reason, e.g. because it is adequate to advance their joint activity, without them believing in the truth of this proposition. If the actual truth value of the proposition is irrelevant for the progression of the joint activity, its quiet acceptance can still facilitate interaction. According to Stalnaker, “the simplest reason to treat a proposition as true is that one believes that it is true” (Stalnaker, 2002, p. 716). Others emphasize the independence of belief and acceptance (Hakli, 2006). Some argue that acceptance can be context- and goal-dependent and can be adopted voluntarily (Hakli, 2006). The context-dependence of acceptance is obviously related to the different notions of common ground: Context-specific common ground has a validity limited to the current context or activity. Personal and communal common ground are restricted to the contexts of particular groups.

We subscribe to the view of common ground as collective acceptance insofar that believing and presupposing are distinct epistemic states and that common ground can be goal- and context-dependent. In particular, we follow Stalnaker’s conceptualization of common ground (see also Section 6.1). Furthermore, we consider personal and communal common ground as the information shared and perceived to be shared beyond a particular interaction.

**Grounding** According to Clark (1996c), coordination during communication happens at four distinct levels, from the execution of a behavior that carries a signal to establishing the signal’s meaning and uptake of its perlocutionary act. Nevertheless, Clark is mainly concerned with the processes by which interlocutors achieve the mutual belief that they have the same understanding of a signal and its meaning. Achieving mutual understanding in this sense is a prerequisite to, but fundamentally different from, the achievement of acceptance—agreeing on going with a proposition for the purpose of the joint activity. It is at this level that the joint activity is advanced and common ground is modified more permanently.

Achieving acceptance is a higher level process that we believe is best described as the negotiation of the truth value of exchanged propositions. We base this claim on the observation that unless an utterance is explicitly disagreed or questioned, any proposition implied by this utterance is normally encoded as true by interlocutors (Gilbert et al., 1990,
2.2. The Grounding Model of Cultural Transmission

This result indicates the role of negotiation in determining what becomes common ground. In the following, we will refer by grounding to the process of achieving the mutual acceptance of a proposition, not mutual understanding in Clark’s sense.

An important aspect of the grounding model of cultural transmission is the assumption that interlocutors do not only ground the content of their communication—target information—but also presuppositional information (Kashima et al., 2008). Similar rules likely apply to the grounding of target and presuppositional information. For example, an interlocutor might not only challenge the truth of target information but also of the presuppositional information she infers from the target information. However, this inference is private to the interlocutors and not established as common ground explicitly. Thereby, interlocutors might come to diverging views about their common ground.

2.2.2 Example Dialogues

We illustrate the previous discussion with three sample dialogues and describe how they can be explained by the grounding model of cultural transmission. The background is that Alice, an employee of the city’s football club, was made aware that Gary, one of the club’s players, was caught drunk-driving the night before after he drank a bottle of wine by himself. Even though all dialogues are based on this same prior event, their contexts (common ground, epistemic and relational goals) differ, thus leading to vastly different outcomes.

The day after the incident, Alice and her work colleague Maria work on the problem of managing the reputation of some players including Gary’s. Their overarching joint goal is to work out how to improve Gary’s reputation. Gary’s identity is in their common ground but the incident from last night is not. Thus, the information that Gary was caught drunk-driving is novel and highly relevant to the joint goal and therefore contributed by Alice according to her epistemic goals. Relational goals—the improvement of Alice’s and Maria’s relationship—play only a minor role, if at all.

(1) Alice: Unfortunately Gary got caught drunk-driving just yesterday.
(2) Maria: What happened? I didn’t hear about that.
(3) Alice: He emptied a bottle after he heard of his grandma’s death but then decided to visit his grandfather.
(4) Maria: I didn’t know he was that close to them.
(5) Alice: His parents were out of town quite often because of their jobs and his grandparents looked after him then.
Chapter 2. Culture

(6) Maria: So he actually is a decent fellow?
(7) Alice: Yes it seems.
(8) Maria: Oh. This incident really is bad luck for him.

In (1) and (3), Alice contributes just as much information as she deems necessary for the successful completion of her joint activity with Maria. For example, she uses the phrase “a bottle” instead of “a bottle of wine”. She prefers suppressing the information that Gary drank wine, assuming that football players are generally seen as beer and not wine drinkers (part of communal common ground). The information that Gary drank wine does not appear relevant to the task to work out how to improve Gary’s reputation. Moreover, this information is incompatible with the common stereotype that football players do not drink wine. Hence, the communication of this information might require additional discussion that would impede the flow of the conversation. Maria cannot make sense of Alice’s story repeatedly and asks for clarifications: First, she indicates that she is not aware of the incident (2), then she indicates that she was not aware of Gary’s relationship with his grandparents (4). Provided further information, Maria summarizes her judgment of Gary being a decent person (6). Amongst others, Alice and Maria ground information about the incident, about Gary’s relationship with his grandparents, and about him being a good but unlucky person in this case.

Later Alice communicates with her husband Bob about her day. The joint goal is to have a casual conversation. However, relational goals are not important because their relationship is already strong. Alice assumes that Gary’s identity is part of their common ground as well as the stereotype that a majority of football players does not drink wine. Therefore, according to epistemic goals, the information that Gary got drunk would be redundant with their stereotypes and hence irrelevant. However, the information that he drank wine is relevant because it is novel to Bob.

(1) Alice: Did you know that Gary is a wine drinker?
(2) Bob: Who is Gary?
(3) Alice: One of our players, the one that we met the other day at Jimmy’s. You remember?
(4) Bob: Mh, no. But... a football player drinking wine?
(5) Alice: Some of them seem to like wine.
(6) Bob: Mh.

Alice makes the wrong assumption that Gary’s identity is part of her common ground with Bob. Only after this conversation, Gary’s identity is part of Alice’s and Bob’s common ground. They also ground that there are football players who do like wine. In effect, they establish that there are exceptions to the stereotype of football players drinking only beer.
Another day Alice has a conversation with her casual acquaintance Stacy. Again, the joint goal is to have a casual conversation. Relational goals are strong because the women are only casual acquaintances. Likewise, their personal common ground is small. Transmitting information that is assumed to be consistent with stereotypes in communal common ground contributes to relational goals. Therefore, Alice tells Stacy that Gary was caught drunk-driving without mentioning the more novel information that Gary actually drank wine, which would not contribute to their epistemic goals any further.

(1) Alice: Gary ..., a player from our club, got caught by the police the other day.
(2) Stacy: What happened?
(3) Alice: The usual story. He got drunk, drove his Porsche at 150, and abused the police when he got caught. He ended up in jail for the night.
(4) Stacy: These football players are all the same.

This conversation grounds some information about the incident and the common attitude towards football players as careless and irresponsible people. In fact, Alice makes use of the phrase "the usual story" to indicate that this stereotype of football players is actually shared on a communal level, i.e. that it is part of communal common ground. Stacy confirms this view in (4).

2.2.3 Premises and Implications Revisited

In the following, we revisit the main premises of the grounding model of cultural transmission and their implications. In the remainder of this thesis, we refer back to these points repeatedly when we identify which of them are accounted for in our models. These are the main premises:

(PREM1) Coactors have some information in common/share this information—their actual common ground.

(PREM2) Coactors make assumptions about which information is part of their common ground—the perceived common ground. Coactors are not aware of their actual common ground and therefore it is perceived common ground that they consider to be common ground.

(PREM3) Common ground consists of communal common ground, which is shared by virtue of common group memberships, personal common ground, which is due to previous common experiences, and context-specific common ground—the information that is shared only within the current context.
Chapter 2. Culture

(PREM4) The culture of a community is largely described by the information which is actually shared and perceived to be shared by the members of the community.

(PREM5) During a joint activity, coactors align their understanding so that they can carry out their activity successfully. This highly interactive alignment process adds to their common ground and is called “grounding”.

(PREM6) Grounding requires effort, which generally decreases with increasing common ground and with increasing compatibility of the exchanged information with existing common ground.

(PREM7) The less effort the grounding process requires from coactors, the less likely grounding is to fail and the more socially-connective and the less socially-disruptive is the interaction.

(PREM8) A joint activity implies epistemic and relational goals. Epistemic goals require the grounding of some information, relational goals require the formation or maintenance of the social relationship of the coactors. Because of the role that grounding plays in the formation and maintenance of social relationships (PREM7), epistemic and relational goals can be incompatible.

(PREM9) Coactors can generalize their personal common ground to the wider community, thereby contributing to the change of the actual and perceived communal common ground of that community.

From these premises follow, more or less directly, a few main implications:

(IMPL1) Cultural transmission does not only depend on what is actually shared, but also on what is perceived to be shared. This follows directly from PREM1 and PREM2. While the distribution of information within a community naturally determines what can be communicated, cultural transmission depends also on what is perceived to be shared. For example in her conversation with Stacy above, Alice makes her communication dependent on the public image of football players that she assumes to be shared.

(IMPL2) Whether some information is exchanged between individuals depends on its compatibility with the information that these individuals share because of their common personal experience and their group memberships. There are two points here: First, the compatibility of transmitted information with common ground and the size of existing common ground determine the effort required for grounding that piece of information (PREM6), which interacts with epistemic and relational goals in determining whether this
2.2. *The Grounding Model of Cultural Transmission*

Information is transmitted (*PREM7, PREM8*). Information that is more compatible with the information assumed to be shared might be more likely to be transmitted because it is easier to exchange and to add to common ground. Second, common ground consists of personal common ground due to common personal experiences and communal common ground due to common group memberships (*PREM3*). Hence, while cultural transmission often occurs during dyadic or small group interactions, it is rarely isolated from the wider context of cultural groups.

(IMPL3) **Cultural transmission at the micro level contributes to the change or maintenance of culture at the macro level.** Cultural transmission occurs during the so-called grounding process in joint activities (*PREM5*). A feature of this grounding process is that it can change actual and perceived communal common ground (*PREM9*), which is a large part of the culture of a community (*PREM4*).

(IMPL4) **Cultural transmission at the micro level can have an effect on the structure of social networks at the macro level.** According to (*PREM7*), cultural transmission can affect the quality of a social relationship. Conceiving a social relationship to be part of a larger social network, cultural transmission therefore has an effect on the social network that this relationship is part of.

(IMPL5) **The culture of communities at the macro level affects cultural transmission at the micro level.** The existing common ground of coactors determines the information that they share or assume to share (*PREM1, PREM2*). Part of common ground is due to communal common ground (*PREM3*), which is the culture of a community (*PREM4*). Through (*PREM6, PREM7*), existing common ground interacts with relational goals and through (*PREM8*) with epistemic goals in determining which information is transmitted during grounding. Therefore, the culture or communal common ground affects cultural transmission.

(IMPL6) **The composition and configuration of social networks at the macro level determine cultural transmission at the micro level.** Obviously, cultural transmission occurs along links in social networks. However, the quality and purpose of these links is of equal importance as their composition. It is the social relationship of coactors that determines their relational goals in joint activities (*PREM8*) and their existing common ground due to previous experiences or common group memberships (*PREM3*). Close social contacts are likely to exchange information with each other that is deemed inconsistent with common ground because they have a rich history of interactions and anything else would be identified
as redundant communication. Remote acquaintances, on the other hand, are likely to be more conservative in what they communicate to improve or not to threaten their social relationship. Therefore, communication along these links will seek to be in line with perceived communal common ground because this is the information that is assumed to be shared by default due to group membership.

A social relationship also determines the joint activities carried out in the context of this relationship, which entail particular epistemic goals (PREM8). For example, interactions within the context of professional relationships would have strong epistemic goals and the success of the joint activity would be weighed higher than the formation or maintenance of social relationships. Ultimately, this means that coactors would more likely accept the risk of transmitting information that is novel and incompatible with their common ground.

### 2.3 Conclusions

This chapter has described culture as a macro-level property of a complex social system and cultural dynamics as the interaction between culture, micro-level cultural transmission, and nano-level mental processes. We have stated the requirement of a two-component research program for the study of cultural dynamics. First, micro-level cultural transmission and its relation with macro-level properties and nano-level processes needs to be understood. Second, based on this understanding, the role of cultural transmission in the change and maintenance of culture at the macro level on a larger scale is to be studied. We have discussed the grounding model of cultural transmission as a promising attempt at targeting the first part of this agenda. The grounding model of cultural transmission builds on Clark’s model of grounding in language use to describe the transmission of cultural information as an implication of joint activity. The grounding process, which amounts to cultural transmission, depends on the context of the interaction and the existing common ground. A more comprehensive treatment of the grounding model of cultural transmission is provided by Kashima et al. (2008).
Chapter 3

Methodology

The previous chapter has described culture as a macro-level property of complex social systems and cultural dynamics as the bi-directional interaction between culture and micro-level cultural transmission driven by nano-level cognitive mechanisms. We have discussed the grounding model of cultural transmission as a detailed description of the micro-level mechanisms that govern cultural transmission. The purpose of this chapter is to describe how this thesis addresses the research questions posed in Chapter 1, namely how the grounding model of cultural transmission can be translated into computational representations and how this effort can contribute to the refinement of that model and to the wider understanding of cultural dynamics. To do this, we also discuss in detail the role of computational modeling in scientific research and the process of translating the grounding model of cultural transmission into computational representations.

We suggest in particular that the development of less detailed computational models of the grounding model of cultural transmission benefits from the availability of more detailed models. The idea is essentially that more detailed models serve the dialogue between modelers and domain experts because they allow for a richer representation and because they sharpen the stakeholders’ understanding of the domain. Thereby it becomes clearer which additional assumptions each model introduces and how it differs from the grounding model of cultural transmission. Hence we encourage a modeling process that includes the development of more detailed auxiliary models. With decreasing level of detail, a model’s contribution to the refinement of the grounding model of cultural transmission decreases. Likewise, however, its (computational) complexity decreases, which facilitates simulations and hence contributes to the exploration of the operational consequences of the grounding model of cultural transmission in the context of cultural dynamics. We support our suggestion in two ways: We argue for our point later in this chapter and we show in Chapter 8 how the development of the model presented there is supported by the model introduced in Chapter 7.
So far a vague idea of what a computational model or a computational simulation is has sufficed to follow our discussion. However, for what follows, this is an essential concept and we cannot defer elaborating on it. Therefore, we continue this chapter in Section 3.1.3 with a discussion of computational models of social dynamics, considering that cultural dynamics fall in this category of phenomena. We identify one type of model, so called agent-based models, as particularly appealing for the study of social phenomena. In Section 3.2, we discuss the processes and issues involved in creating agent-based models of social dynamics and advance our own perspective. In doing so, we indicate how agent-based modeling can benefit from other computational approaches to modeling social dynamics. We follow with a brief discussion of the problems of computational modeling in Section 3.3. Based on these discussions we justify subsequently in Section 3.4 how the next chapters contribute to answering our research questions. We conclude the chapter with a brief summary in Section 3.5.

3.1 Computational Models of Social Dynamics

This section elaborates on the role of computational models in the study of human social behavior and dynamics in general and cultural dynamics in particular. We discuss approaches that differ in their purpose and in the scientific discipline from which they have emerged. Throughout our discussion, we converge to agent-based models as an appropriate tool for the study of cultural dynamics. Before we delve into these details, however, we take a step back to discuss the nature and purpose of models in general and in particular their interrelation with theories, other models, and the systems they represent. Section 3.1.1 provides this background discussion. Section 3.1.2 argues the case for computational models for the analysis of social dynamics in general and cultural dynamics in particular. Section 3.1.3 describes and compares the most prominent computational approaches for studying social dynamics and behavior with an emphasis on their suitability for the study of cultural dynamics.

3.1.1 Models, Theories, and Simulations

Models abound in scientific practice although their semantic, ontological, and epistemological status is still debated in philosophy (Frigg and Hartmann, 2009). An in-depth engagement in this debate would be beyond the scope of this thesis. Therefore, we provide but a shallow treatment of this topic, which is however sufficient for our purpose and at the same time not incompatible with current mainstream philosophy of science. In the following, we
3.1. Computational Models of Social Dynamics

refer to Figure 3.1. In science, models can fulfill at least two representational functions:¹ They can represent a theory or they can represent a part of the world (Frigg and Hartmann, 2009). We call the first class theoretical models and the second one representational models. Under the umbrella of representational models we also consider those models that represent other models. This can be the case, for example, when a formal model is derived from a verbal-conceptual model.²

Theoretical Models

Different relations between theories and theoretical models have been proposed. According to the so called syntactic view (Frigg and Hartmann, 2009, sometimes also received view, Contessa, 2011), a theory is a set of sentences (axioms) in a formal language. An interpretation of this theory is a model of this theory if all axioms of that theory are true in that interpretation. For example, Newton’s laws of motion and his law of universal gravitation form a theory. One can construct a model of planetary movement in our solar system in which this theory is true. Or one could construct a model of the movement of an object falling from the leaning tower of Pisa in which Newton’s laws are true. In that, a theoretical model is a concrete instantiation of that theory. According to the semantic view, theories are not collections of sentences but families of models, which are commonly considered to be set-theoretic structures (Frigg, 2006). Hence from this perspective the models of a theory are the theory. An alternative suggestion is that models are autonomous mediators between theory and data (Cartwright, 1999; Morgan and Morrison, 1999). This view, which can be called the models-as-mediators view, allows for the consideration of models that are not rigorously deducible from the theories they represent. The independence of models from theory enables them to fulfill functions they could otherwise not, i.e. complementing theories that are incompletely specified, substituting theories when these are too complex to study or unavailable, or serving as preliminary theories (Frigg and Hartmann, 2009). These are, in fact, common uses of models in scientific practice.

Because of the assumed independence between theories and models, the models-as-mediators perspective allows for models as representations of theories that are descriptions of mechanisms (instead of sets of axioms or families of models). A mechanism is often conceptualized as a system of parts that interact according to particular processes, activities, or operations to bring about a particular behavior (e.g. Bechtel and Abrahamsen, 2005; Hedström and Ylikoski, 2010; Machamer et al., 2000).

¹This thesis is primarily concerned with the representational functions of models but not with other functions, listed for example by McBurney (2012).
²While this case is not explicitly considered in the literature cited in the following, we feel that it is an important one.
Figure 3.1: The role of a model in the context of a theory, other models, and a target system.

The crucial point for this thesis is that theories relevant at all levels of modeling social dynamics can be understood as descriptions of mechanisms (Elsenbroich, 2012; Sawyer, 2004; Smith and Conrey, 2007; Thagard, 2008, 2009). Moreover, the components of mechanisms obviously relate to the components of complex systems as discussed in Section 2.1, namely constituents and their interactions. This relationship was already discussed by Glennan (1996). Mechanistic theories offer a kind of causal explanation, instead of mere propositions about the covariations between initial conditions and outcomes (Elsenbroich, 2012; Mayntz, 2004). However, a mechanistic explanation is not necessarily a full causal explanation (Elsenbroich, 2012). A comprehensive overview of this topic was recently provided by Hedström and Ylikoski (2010).

While Thagard (2009) admits that the mechanistic view of theories is disputed in the philosophy of science, we find this perspective helpful because parts and processes of mechanisms do not only relate to complex systems but also to the primitives of computation: data structures and operations on these data structures. When a model is couched in terms of computational entities, non-technical stakeholders who endorse the mechanistic perspective of theories can more easily understand the mapping between theory and model. In fact, the mechanistic perspective of theories is the one implicitly taken by many researchers in psychology, neuroscience, and biology (Thagard, 2009). Explaining phenomena mechanistically is also compatible with methodological individualism, which postulates that macro-level properties are to be explained from the interactions of individuals and their intentional states (Weber, 1978). Likewise, it is possible to reconcile mechanistic explanation with the notion of emergence (Sawyer, 2004).
3.1. Computational Models of Social Dynamics

Representational Models

A representational model is an object that represents a part of the world (the target system) in a simplified manner. The hydraulic model of economic activity is a representational model because a system following the laws of hydraulics is used to represent an economic system. A model can simplify the target in different ways, for example by representing the target on a smaller scale or by using idealized representations of its features (Gilbert, 2008).

We are interested in representational models of cultural dynamics as representations of a part of the world. Yet it is important to note that theoretical models often serve as the building blocks of representational models (Contessa, 2011). Hence the distinction between theoretical and representational models is often blurred, which explains the frequent entanglement of the terms “theory” and “model”. Two subclasses of representational models can further be distinguished, depending on the nature of the target: models of phenomena and models of data (Frigg and Hartmann, 2009). Phenomena that could be represented by models are, for example, language production in the human mind or disease spreading within a society. Models of data are those obtained by statistical methods, e.g. linear regression. Clearly, when modeling cultural dynamics, we deal with a phenomenon. However, we shall also encounter models of data in this thesis. Given raw data, one typically employs modeling to extract general trends and to compare them to the predictions of theory (Frigg and Hartmann, 2009). As we shall see, it is also possible to conceive a model of phenomena from the perspective of models of data when the model is to be calibrated to existing data.

Ontology and Epistemology of Models

Models can be described verbally or formally, sometimes with the help of visual descriptions.3 Most social scientific models are verbal-conceptual ones, i.e. they are described in informal language. Mathematical and computational models are subclasses of formal models. Mathematical models are described in terms of variables and equations, computational models in terms of data structures and operations. This point immediately highlights a distinction between formal and verbal-conceptual models: Formal models are necessarily more precise. But precision in the case of a complex model might require cumbersome tinkering with details. Verbal-conceptual models allow us to describe complex issues with less detail by relying on the reader’s understanding and intuition to link all aspects together implicitly.

Having described what models are, we need to discuss what their epistemic benefit is. In doing so, we refer again to Figure 3.1. The lifecycle of a model consists of its construction, use, and interpretation with regard to the target system and/or underlying theories and models.

---

3 Other types of models such as physical ones are not relevant to this thesis.
Chapter 3. Methodology

(Frigg and Hartmann, 2009). These steps correspond to the *denotation, demonstration, and interpretation* of models as considered by Hughes (1997).

The construction of a model requires the researcher to fit different pieces together, i.e. theories and/or other models and/or information about the phenomena and data describing a target system. In that way, a conceptual model is made up of components that we call *assumptions*. Specifying the model’s assumptions is a creative process that also relies on the input of the modeler. As Grüne-Yanoff and Weirich (2010) have it, modeling requires “ingenuity, sense of purpose, technical ability, and luck”. Therefore, models are typically not constructed only from theory and data, which highlights their role as mediators; and the very process of constructing a model affords an opportunity for learning about the model (Morgan, 1999). In fact, there is an argument that computational models of human cognition are theories in their own right because they extend existing theories in the way described here (Simon, 2002; Sun, 2009b). We feel more comfortable with maintaining the distinction between theories and models because both serve different purposes. Therefore, we do not agree that such a model is a theory in its own right but we accept that the model implies a corresponding theory of which this model is but one possible model.

The use of a model depends on its type and purpose. A verbal-conceptual model can be used, for example, by thought experiments. Mathematical models can be used by deriving results or solving their equations, either analytically or with numerical means. Computational models are primarily employed for simulation to study their dynamics, a point we will return to later. Essentially, the model is used to derive its operational consequences or implications. The model serves as a surrogate that is studied instead of a theory or a target system (Swoyer, 1991). Obviously, the use of the model offers another opportunity for learning about the model itself (Morgan, 1999).

Finally, insights gained from analyzing the model need to be interpreted with respect to the underlying theories or models or the target system. Also this step depends on the type of model and the purpose of the study. No generally accepted methodology exists (Frigg and Hartmann, 2009). With regard to a target system, a model’s purpose is often prediction, whereby different types of prediction are possible (Troitzsch, 2009a). Another frequent purpose is explanation. We emphasize that while prediction is often taken to be the main benefit of models, it is not the only one. In particular, this thesis mostly focuses on uses of models other than prediction. Again following Thagard (2009), we consider explanation to be a description of how the operation of a mechanism gives rise to a certain phenomenon. With regard to an underlying theory or model, the construction of a model often highlights gaps, conflicts, and inconsistencies, especially if the theory or model is described informally and does not provide all the necessary details. Models can also serve as test-beds to explore the consequences of theories or other models if real experiments are infeasible. Models of
3.1. Computational Models of Social Dynamics

data provide rigorous descriptions of data obtained from the target system, which can be used to guide the development of theories. Hence, there are multiple ways in which models can contribute to the revision and construction of theories and other models (Sun, 2009b). As pointed out in Section 3, we rely on this with our contribution to RQ2—how computational models of cultural dynamics can contribute to the grounding model of cultural transmission.

Many more reasons for modeling exist (Epstein, 2008; McBurney, 2012). For example, models can guide data collection in the way theories do, or they can facilitate communication about phenomena in the target system because they provide explicit statements about the behavior of the target. Often, the use of a model does not even make explicit reference to theories, other models, or a target system (McBurney, 2012). As indicated previously, however, we focus here on the representational functions of models.

Models and Targets

So far we have not discussed the relationship that holds between model and target. Philosophers ask two questions about this relationship: What are the properties of the relationship between model and target so that one can say that the model represents the target; and what makes this relationship successful (Knuuttila, 2011)?

According to Knuuttila (2011), there are two classes of answers to these questions. The first class is the one held primarily by proponents of the semantic view of models. This perspective typically holds that a model represents a target if model and target produce mathematical structures that are related to each other through some kind of morphism, e.g., an isomorphism. If such a morphism exists, the representation is also considered successful. There are various problems with this idea, in particular with regard to the role of models in current scientific practice. For example, there is no room for models that cannot be related to their target system through a morphism. However, constructing models that are not perfectly accurate is common practice.

The second class of answers makes a representational relationship dependent on the user of a model intending to use this model for a representational purpose. A model represents a target system if at least the model’s user intends and denotes it to. Hence, a model relates to a target not through structural similarity but through a third entity (the user) in this relationship. However, this obviously raises the question of how we can learn anything about the target system from the model. This is sought to be remedied by giving more weight to the second question, which is concerned with the success of a representational relationship. We elaborate on this in the following.

It is safe to claim that the success of a representation depends on the purpose of the study. A success criterion could be that the representation be faithful to a certain degree. Faithfulness could be defined in terms of the strength of the morphism that holds between the
mathematical structures produced by the model and the target (Contessa, 2011). Obviously, this success criterion is borrowed from the first class of answers to the two questions above (Knuuttila, 2011). Contessa offers a map of London’s subway network as an example: Every circle would correspond to a subway station and lines between each circle to an existing connection between the respective stations. For an up-to-date map, one could construct an isomorphism between the map and the actual subway system. An out-dated map might miss new stations or connections and therefore the morphism that one can construct is weaker than an isomorphism.

However, faithfulness is something difficult to establish in practice, considering the potential complexity of a target system. Moreover, only under certain conditions one would want to construct a faithful and hence complex model. Typically, simpler models are preferred because their dynamics are more easily understood, following the KISS (“Keep it simple, stupid!”) principle and Occam’s razor.4 Alternatively, a model could be deemed successful if it exhibits reliability, empirical adequacy, explanatory power, etc. (Knuuttila, 2011). As we shall see later, practitioners have developed methods to determine success based on these kind of aspects. Later in this chapter we will discuss validation as a tool for establishing that a model is a successful representation of the target in terms of empirical adequacy.

Computational Models

We mentioned before that simulation is the primary use of computational models. Let us expand on this point. For now, we assume that a computational model accepts an input (independent variables) including parameters and initial conditions and provides an output (dependent variables) based on the representation of a mechanism (Sawyer, 2004; Thagard, 2008). That is, a computational model or alternatively simulation model imitates a mechanism and therefore has a temporal dimension. As mentioned before, parts and processes as the constituents of a mechanism are naturally mapped onto data structures and operations, which are the primitives of computational models. A computational model is implemented by a computer program. We call the execution of the program a simulation or sometimes a simulation run. The purpose of simulation is to study the relationship between inputs and outputs of the program and thus the model, possibly in both directions. This activity yields insights into the operational consequences of the model.

Clearly, our definition of simulation is a simplified one. For example, we do not account for simulations that are not implemented computationally or simulations of differential equations, which is a frequent instrumental application of computational simulations but not a case of representing a mechanism directly. However, this definition is sufficient for our

4But see our later discussion about this point in the context of agent-based models.
3.1. Computational Models of Social Dynamics

purpose. The reader might refer to Grüne-Yanoff and Weirich (2010) for a comprehensive
discussion on various perspectives on the term “simulation”.

In our view, simulations are just a particular use of a particular type of model. In a
similar way to other models, simulation models are a tool for learning about theories, other
models, and target systems. In a limited way, one can employ simulations as substitutes for
experiments, which can be helpful when actual experiments are impossible, e.g. because
of scale or because of temporal or financial limitations. However, the ontological and
epistemological differences between experiments and simulations are still debated (Grüne-
Yanoff and Weirich, 2010). It is clear that the knowledge that can be gained from experiments
with real individuals and objects is different from the knowledge that can be gained from
simulating a model of these entities. Despite this philosophical debate, the practical value of
computational simulations in science is generally agreed on. For example, simulations are
used for proof, prediction, explanation, discovery, or exploration.

In summary, models are vehicles for learning about target systems, theories, and other
models. Simulations are essentially the execution of computational models, which represent
mechanisms. The important messages to take from this section are:

- We are concerned with representational models that might be informed by theories,
  other models, and data.
- Models can be used in various ways to learn about theories, other models, and targets.
- The usefulness of a model is determined by the purpose of the study.
- Simulations are yet another use of a particular kind of model to study theories, other
  models, and targets.

Much more could be said about models, theories, and simulations. We shall elaborate on our
discussion when describing computational models of social dynamics.

3.1.2 The Case for Computational Models of Social Dynamics

Now that we have established the nature and purpose of computational models and simu-
lations, we argue the case for their use in the study of social dynamics. We first contrast
computational models with other common types of models and then compare their epistemic
benefit with the one of other common methods for the analysis of social systems, which
obviously constitute the target systems in the study of social dynamics.

Human social dynamics are studied in fields as diverse as psychology, sociology, eco-
nomics, anthropology, biology, cognitive science, or philosophy. Depending on the field,
verbal-conceptual or formal models as representations of the theories and systems under
study prevail. Verbal-conceptual models are expressive in the sense that they are flexible with regard to the specification of the model but they are also vague and imprecise, and therefore typically leave a lot of interpretation to the reader. Formal models are usually mathematical ones. In contrast to verbal-conceptual models, mathematical models are precise but less expressive. Mathematical models usually provide descriptions of the relationship between different variables in terms of equations. Therefore their use with regard to explaining which mechanism gave rise to a particular observed behavior is often limited.

Computational models retain the benefits of verbal as well as mathematical models because computational languages are as precise as and typically more expressive than mathematical ones. Strictly speaking mathematical models can be as expressive as computational ones but they are hardly ever used as such. The reason is that complex mathematical models quickly become intractable or difficult to manipulate, especially when nonlinear, qualitative, and conditional effects are considered. Computational models, in contrast, do not suffer from this shortcoming. Therefore, computational models lend themselves to the representation of complex theories and targets such as social systems. Moreover, at all levels of the nano-to-macro axis theories are in fact couched in terms of mechanisms, which hence map onto computational models (Sawyer, 2004; Thagard, 2008, 2009). For example, psychological theory stipulates mental representations and processes operating on these representations. Similarly, mechanisms at the macro level consist of individual actors and the processes such as influence or communication that act between them. Further advantages of computational models over verbal-conceptual ones are that they require assumptions to be laid out explicitly and completely and that they offer the possibility of exploring the consequences of these assumptions exhaustively by simulation (Gilbert, 2008).

As argued in the previous chapter, the study of social dynamics in general and cultural dynamics in particular is an exceptionally intricate one. One of the main reasons for this is that a complete account of cultural dynamics appears to require the simultaneous inspection at all levels of the nano-to-macro axis. The primary means for the study of social dynamics are experiments. As an experiment we understand a situation in which a number of participants engage in a task under controlled conditions while data about their behavior and performance is being collected in order to test a particular hypothesis. Experiments do allow us to obtain observations from social systems directly, also taking into consideration the inherent dynamic nature of the objects under study. Therefore, experiments enjoy a unique epistemological status. However, the scope of experiments along the nano-to-macro axis is severely restricted by temporal, financial, ethical, and technical constraints. For example, experiments with more than a few dozen participants become typically infeasible. To some extent, these limitations are mitigated by other methods. Surveys, for instance, can be conducted to inspect social systems at a larger scale but surveys can only provide snapshots and thereby
3.1. Computational Models of Social Dynamics

normally lose temporal information. Surveys also do usually not probe for information about
the interactions between individuals. As discussed in the previous subsection, simulations
can substitute or at least complement experiments to some extent.

There are at least two benefits to be gained from employing computational models:
First, computational simulations allow us to explicitly represent the mechanisms assumed
to yield particular observations. Experiments, in contrast, make only implicit reference to
mechanisms via the hypotheses they are testing. An experiment delivers data but does not
directly point to the mechanism that produced this data. Second, one can easily conceive
computational models that represent a larger scope on the nano-to-macro axis. Exactly
this kind of model enables a study of cultural dynamics that considers all levels of the
nano-to-macro axis simultaneously.

In conclusion, computational modeling is a promising tool that can complement other
methods in the study of social dynamics. Computational models are the most appropriate
tool to represent the mechanisms underlying social dynamics and they offer the unique
opportunity to bridge all levels of the nano-to-macro axis.

3.1.3 Existing Computational Models of Social Dynamics

Obviously, the argument we made in the previous subsection is not a new one. In fact,
computational modeling of social dynamics has been a busy research program throughout
all the disciplines mentioned previously for at least two decades. The first computational
models of social dynamics were already developed in the 1960’s and 70’s (see references in
Cioffi-Revilla, 2010a), long before the recent surge in computing power enabled large-scale
implementations. The term “computational social science” has been advanced as the name
of this emerging field, which also includes other computational approaches than simulation
for the study of social dynamics (e.g. Cioffi-Revilla, 2010a; Gilbert, 2007; Macal and North,
2010). In the following, we identify the most important approaches and their relationship
with the disciplines mentioned before. We describe each approach in a bit more detail and
discuss how it can contribute to bridging the different levels of the nano-to-macro axis in the
study of social dynamics.

System Dynamics Models

Together with agent-based models, which we discuss soon, system dynamics models are
arguably the most prevalent types of computational models of social dynamics that account
for a large number of actors (Cioffi-Revilla, 2010a). System dynamics models were first
introduced by Forrester (1961) and have since been used abundantly for the study of complex
systems, including social systems (Garson, 2009). The elements of system dynamics models
are state variables (stocks) and the influence on the change of one variable by another (flows). Other elements are feedbacks and time delays. Stocks could, for example, denote the numbers of employed and unemployed people in a society and the flows between them the rates at which unemployed people become employed and vice versa. System dynamics models obviously do implement mechanisms but they fall short of taking into account more than the macro level. To this thesis, therefore, system dynamics models are not of relevance.

**Computational Cognitive Models**

Computational cognitive models are an important tool in cognitive science (Sun, 2008; Thagard, 2009). Cognitive science is a highly interdisciplinary field that builds on psychology, neuroscience, anthropology, artificial intelligence, and philosophy and is concerned with the analysis of the mechanisms that drive cognitive processing (Thagard, 2008). In accordance with our previous discussion, computational modeling in cognitive science rests on the correspondence between the constituents of mechanisms and computational models. The main working hypothesis in cognitive science is that cognitive processing is analogue to computational operations on mental representations (Thagard, 2011).

Various types of mechanisms have been proposed as the representations of cognitive processing. Among the most prevalent ones are *logic-based processing*, *rule-based processing*, and *connectionist processing* (Thagard, 2011).

Logics-based approaches assume that inferences people make can be represented by inductive and deductive procedures on sentences in formal logics (Bringsjord, 2008; Thagard, 2011). Hence, this approach relies on a *symbolic* representation of knowledge. While it is intriguing to assume that people’s reasoning follows rules of logical inference, there are some problems with this approach (McClelland, 2009): First, it is difficult to conceive how cognitive processing beyond symbolic reasoning, e.g. perception, can be explained in terms of logical inference only. Second, even in strict logical inference tasks, humans do not necessarily perform according to theoretical predictions.

Another symbolic approach is rule-based processing, which relies on the encoding of declarative knowledge by propositions and on the representation of procedural knowledge by IF-THEN or so called *production rules* (Bringsjord, 2008). Usually a difference is made between long-term memory that includes all declarative knowledge and working memory that contains only that declarative knowledge which is active at the moment. Processing consists of matching the conditions of rules against working memory content and then executing activated rules. Activated rules can perform changes on the memory structures or lead to actions in the environment of the system. In particular, the implication of an activated rule can change the content of working memory and thereby trigger the execution of further rules. In this basic form, rule-based processing corresponds to inference in simple logical
systems. However, the accessibility of production rules is typically conditioned on recency of use, expected utility, or other factors assumed to affect human cognition (McClelland, 2009). Hence, these models are usually more closely aligned with various functions of human cognition.

Connectionist approaches postulate that cognitive processing is best described by the spreading of activation between neuron-like processing units and the storage of information in the weight of the connections between them (Rumelhart et al., 1986). In contrast to the two previous modes of processing, connectionist processing is often said to happen on a subsymbolic level. This type of model is conducive to the representation of automatic processes in which high-dimensional input is reconciled with prior knowledge encoded in the network of processing units (McClelland, 2009). While connectionism borrows from neuroscience, claiming any direct correspondence between connectionist models and the actual neural structure in the human brain is far-fetched (McClelland, 2009). The drawback of connectionist models is that their processing is often difficult to understand. Given the low representational level, it is also difficult to see how mechanisms of higher cognition could be represented in connectionist models.

For some, the ultimate goal of computational cognitive modeling is a unified computational framework of human cognition that enables systems to exhibit intelligent behavior comparable to the one of humans. Such attempts are called cognitive architectures. Prominent examples of cognitive architectures are ACT-R (Anderson, 1993), Soar (Newell, 1994), and CLARION (Sun, 2002). The former two rely primarily on rule-based processing, the latter on a combination of rule-based and connectionist processing. Cognitive architectures provide a framework of mechanisms of human cognition that are assumed to be invariant across time, task, and individuals (Langley et al., 2009; McClelland, 2009; Sun, 2008). In that, cognitive architectures embody general theoretical assumptions about human cognition. These architectures can then be instantiated with domain-dependent content. Aspects common to most of these architectures are short-term and long-term memory as well as processes that operate on the mental representations in these memory structures (Langley et al., 2009). Common capabilities include decision-making, learning, perception, problem solving, and planning (Langley et al., 2009). Cognitive architectures have a successful history of alignment with empirical data (Sun, 2009b).

One of the criticisms of cognitive science is that it fails to consider the individual mind to be situated and embodied in a physical and social context (Thagard, 2011). Indeed, computational cognitive models are typically not embedded in any environment or enabled to interact with other systems. Thagard (2008) argues that this challenge does not invalidate cognitive science as a research program but rather calls for an extension of current practices. If one accepts, as we do here, that mechanisms are the means for describing social dynamics
Chapter 3. Methodology

at all levels of the nano-to-macro scale, then cognitive science is an appropriate research program for the study of social dynamics in general and cultural dynamics in particular. However, the extension of cognitive science research beyond the individual mind has not been advanced until recently (Goldstone and Janssen, 2005; Sun, 2009a). An obvious challenge that such an effort is going to face is the complexity of computational cognitive models, which renders large-scale simulations difficult. Regardless of this, cognitive science has developed promising tools for describing human cognition at the individual level. Therefore, cognitive models can provide valuable insights into the role of individual cognition in cultural dynamics or in particular on their role in the transmission of cultural information during interactions. Cognitive models allow us to take seriously the intra-personal processes that govern joint activities.

Agent-based Models

As much as cognitive modeling is concerned with studying how mental representations and processes give rise to the behavior of individuals, agent-based modeling is concerned with studying how macro-level phenomena emerge from the micro-level interactions of individuals. To this end, agent-based models explicitly represent individuals and their interactions, hence providing a representation of the mechanisms that underlie social dynamics at a higher level than the one considered by cognitive modeling. Agent-based modeling emerged originally from a number of fields, including distributed artificial intelligence and multi-agent systems. Agent-based modeling has now found wide acceptance in various fields such as ecology (Grimm and Railsback, 2005), psychology (Smith and Conrey, 2007), sociology (Macy and Willer, 2002), and economics (Tesfatsion, 2006). Alternative terms for agent-based models are “multi-agent-based simulations” or also “multi-agent social simulations” or “agent-based social simulations” when the focus lies on the modeling of social systems. In ecology, agent-based models are also called “individual-based models” (Grimm and Railsback, 2005).

The multi-agent systems community defines an agent as a software system that is autonomous, reactive and proactive in pursuit of its goals, and able to interact with other agents and its environment (Wooldridge and Jennings, 1995). An agent-based model consists of a number of agents that represent actors in the target system such as individuals or organizations. These agents “inhabit” a virtual space that represents the environment of the actors in the target system. Agents might interact both with their environment or other agents. Hence, the agents’ behavior can be coupled directly through interactions or indirectly through changes they inflict on the environment.

An important potential ability of these agents is that they might be able to learn, that is, they could be capable of accumulating knowledge about their environment and adapting their behavior (Gilbert, 2008; Macy and Willer, 2002). This implies the availability of some
We agree with Troitzsch (2009b) that in general cultural transmission between individuals requires them to store models of their environment (e.g. perceived common ground) in memory. However, we emphasize that not all types of cultural transmission require explicit models of the environment (see Section 7.6.3 for an example) and that agent-based models not representing this capability can nonetheless contribute to the study of cultural dynamics (see Chapter 4).

Agents normally have limited information and a bounded rationality, which contrasts with the typical assumption of economic models that agents have infinite computational power to optimize their behavior (Smith and Conrey, 2007). Another important characteristic of agent-based models is that agents are typically heterogeneous, e.g. with respect to their attributes, behavior, information available to them, or their location in the environment. Examples of agent-based models include a model of racial segregation in US cities (Schelling, 1971), and models of cultural diffusion (Axelrod, 1997b) and farming practices (Kaufmann et al., 2009).

Usually, a distinction is made between agent-based modeling, cellular automata, and microsimulation models. We agree with Williamson (2007) that the differences between these types of models have become blurred and therefore we will address them all as agent-based models. This is not to deny microsimulation and cellular automata their status but to avoid unnecessary confusion about names.

We accept that agent-based modeling provides the appropriate framework for this thesis because we are interested in the emergence of culture as a macro-level phenomenon from the micro-level interactions of individuals. Sawyer (2009) argues that agent-based models are the appropriate tool for studying the interaction between micro and macro level of social dynamics. Agent-based models explicitly represent interactions between individuals and allow us to study how these interactions give rise to macro-level phenomena. In fact, while most agent-based models do not represent any macro-level properties explicitly and therefore implicitly adopt the position of methodological individualism (Epstein, 1999; Sawyer, 2004), explicit representations of macro-level properties are not ruled out categorically (see for example Dignum et al., 2008). Others also come to the conclusion that agent-based models are the appropriate tool for studying social dynamics (e.g. Gilbert, 2008; Smith and Conrey, 2007; Squazzoni, 2008; Troitzsch, 2009b).

Agents and their interactions, which are the components of agent-based models, map promisingly onto the constituents of the mechanisms we are interested in: actors and their interactions. The view that agent-based modeling fits well with the mechanism account of explanation is also advanced by others (e.g. Elsenbroich, 2012; Hedström and Ylikoski, 2010). In fact, there is an argument similar to the theory-as-mechanism perspective first mentioned in Section 3.1.1 that agent-based models provide candidate explanations of macro-
Chapter 3. Methodology

level phenomena by generating these phenomena from the interactions of agents (Epstein, 1999). It is not entirely clear though what kind of explanations agent-based models and simulations of social systems can provide. In this case, the underlying laws that govern the system’s dynamics are not known or not well confirmed (as opposed to a crash simulation, for example). Grüne-Yanoff (2009) argues that agent-based models do not offer full explanatory power because they can never be fully validated. He suggests instead that these models can provide potential explanations. Elsenbroich (2012) holds against this argument that the problem is not at all unique to agent-based modeling of social systems but characteristic of any social scientific research.

Agent-based models allow for the integration of lower-level mechanisms, i.e. agents can in fact implement more complex cognitive processing as considered in cognitive modeling. Enriching agents’ cognitive abilities and therefore aligning them closer with human cognitive abilities has been identified as promising (Goldstone and Janssen, 2005). This would afford the opportunity for covering a larger scope on the nano-to-macro axis. Consequently, agent-based models are the appropriate tool for studying complex social systems in general and cultural dynamics in particular.

The toolkit of agent-based modeling is impressively large in the sense that modelers draw on a wide variety of possible representations for different model components. For example, some models adopt a simple lattice as the representation of the physical environment assuming that each agent occupies one tile of the lattice. Others rely on real geo-spatial information to create an environment that resembles more closely a part of our world. Some models assume that interactions between agents are guided by an underlying social network, others assume that interactions depend on spatial proximity with regard to the environment. Sometimes the environment is an entirely social one, for example representing an organization. The behavior of agents can be based on simple rules or more complex agent models operating on mental attitudes such as beliefs, desires, intentions, and goals. Sometimes such modeling decisions imply lesser or greater computational complexity, or lesser or greater faithfulness with respect to the target system.

A benefit of this diverse toolkit is that agent-based models lend themselves to the representation of various intra- and inter-personal processes (Smith and Conrey, 2007). A disadvantage of the diversity is that most agent-based models are custom-made and cannot be compared to any other agents-based models directly. Most models are in fact “one-off” models (Carley, 2009), especially those that represent particular situations (but see Scerri et al., 2010, for an effort to develop a generic framework of agent-based models with reusable components). Note that domain dependence is in stark contrast to the main property of cognitive architectures, namely independence from particular domains. Because of the diversity of models, we do not believe that a categorization of agent-based models along
3.2 Developing Agent-based Models

particular dimensions is meaningful or in any way helpful to the discussion here. However, we provide references to such efforts because they provide a colorful illustration of the diversity we indicate here (e.g. Garson, 2009; Gilbert, 2007).

We conclude this subsection by emphasizing once again the key characteristics of agent-based models: Agents are autonomous actors and typically the population of agents is heterogeneous. The rationality of agents is bounded. Agents interact with each other and with a virtual environment.

Summary

Various computational approaches are deployed to study social dynamics. We have identified agent-based models as the most appropriate framework for this thesis. Agent-based models enable the explicit representation of individuals and their interactions with each other and with their environment. Agent-based models allow us to study how the mechanism formed by these micro-level components leads to the emergence of cultural dynamics and properties at the macro level. In the rest of this chapter and in the rest of this thesis we shall therefore primarily refer to agent-based models for studying cultural dynamics.

However, we have also noted the benefits of cognitive modeling. On the one hand, there is a clear overlap between agent-based and cognitive modeling. Cognitive modeling lends itself to a more precise representation of the nano-level cognitive mechanisms that influence cultural transmission in joint activities. This encourages the integration of cognitive models into agent-based models. On the other hand, cognitive models deserve their own place in this thesis as models that by virtue of their detail contribute to the refinement of the grounding model of cultural transmission.

3.2 Developing Agent-based Models

In this section, we provide our take on the process of developing agent-based models. While we specifically address agent-based modeling in this section, most of the presented material is also applicable to cognitive modeling. Thereby we lay the foundations for Section 3.4 in which we explain how we intend to answer our research questions by constructing computational models of cultural dynamics based on the grounding model of cultural transmission. Let us emphasize upfront that due to the recency of agent-based modeling there is not yet a commonly accepted method suitable for all agent-based modeling projects (Hales, 2010).

Recall the perspective of a model as an intermediary between theory, other models, and target system presented in Section 3.1.1. We now expand on the lifecycle of models—construction, use, and interpretation with regard to theory and/or other models and/or a target system—in the case of agent-based models. In particular, we describe in more detail
how an agent-based model is informed by theory, data, and other models, and how the analysis of an agent-based model yields insights into theory, models, and target system. Following others before us (e.g. Gilbert, 2007), we adopt a process-view on the development of agent-based models. We illustrate our understanding in Figure 3.2 and describe each of the steps involved in more detail below. Briefly, a study involving agent-based modeling requires the identification of a purpose or research question, the development of a conceptual model, an implementation, and finally the analysis or use of the model. As indicated in the Figure, it is conceptual modeling and analysis that relates the model to theory, other models, and the target system.

We consider each of the steps in this process to have a goal, which is achieved by following associated modeling principles and techniques. It is important to remark that this modeling process is iterative in the sense that a return to earlier stages will often be necessary. Agent-based modeling projects are typically carried out by teams with stakeholders coming from different disciplines. In some cases, there might be only two groups of stakeholders, e.g. domain experts and computer scientists. In other cases, input from different fields is required so that domain experts might come from different disciplines as well. We will point out a few principles and techniques to facilitate interdisciplinary agent-based modeling projects.

In parallel to the construction and use of an agent-based model, it is essential to establish that this model is valid for the purpose of the study. In the language of our previous discussion this means that we need to establish whether the model is a successful representation, because it is sufficiently faithful to the target system and to underlying theories and other models, or because it exhibits sufficient empirical adequacy, or because it fulfills other criteria of success (Knuuttila, 2011). In scientific research, faithfulness (also accuracy or fidelity) and empirical adequacy are the most prevalent aspects according to which the validity of a model is judged. Accordingly, each step in the process outlined above is matched with a subprocess responsible for establishing the validity of its result. In particular, these steps are the conceptual validation of the conceptual model, the verification of the implementation, and the operational validation of the outputs produced by this model (Sargent, 2008). We discuss these validation processes and their goals and techniques together with their paired constructive processes. We focus here on those techniques that are most prevalent and most relevant to this thesis. Other validation techniques are described by Sargent (2008).

Each of the next subsections is concerned with one of the steps illustrated in Figure 3.2.

3.2.1 Purpose Identification

The goal of the first step in the development process is to identify the purpose that the study serves, which in scientific research is normally the research question to answer. In general, there are many different purposes a model can serve (Carley, 2009; Epstein, 2008),
3.2. Developing Agent-based Models

Purpose
Identification
Conceptual
Modeling
Implementation
Analysis
Conceptual
Validation
Verification
Operational
Validation

Figure 3.2: The process of developing agent-based models. Rectangles with rounded corners correspond to subprocesses and normal rectangles to entities that this process depends on. Solid lines denote the order in which steps can be taken, dashed lines to the relations between subprocesses and dependent entities.

e.g. theory development or hypothesis generation. In the normal course of computational modeling in scientific research, one typically aims at answering a question that pertains to one of three types (Burton, 2003): What is (positive)? What might be (plausible)? What should be (normative)?

Let us exemplify the types of questions belonging to each of these three categories and how they influence the later course of the development process. This discussion is based on Burton (2003). If we ask a what is question, we are interested in how a target system really works. We might want to explain how a certain mechanism yields a particular outcome, or we might want to predict a previously unobserved outcome produced by a mechanism. In both cases, the mechanism represented by our agent-based model is initially nothing more than a hypothesis. It is up to us to demonstrate that the model is valid, e.g. that it is a sufficiently faithful representation of the target system, other models, or theories, or that it exhibits sufficient empirical adequacy. Sufficiency obviously depends on the question we are asking. The model only needs to be as faithful and empirically adequate as is required by the research question.

If we ask a what might be question, we are interested in how a target system might behave if a particular mechanism was at work. This question is the domain of abstract models that allow us to investigate the consequences of very general mechanisms which are difficult to relate to a target system closely. These models usually do not require high faithfulness or
empirical adequacy. However, one can imagine applying what might be questions to models with high faithfulness and empirical adequacy, too. For example, one could be interested in the outcome if a mechanism represented by an agent-based model was changed in one particular aspect only. This allows us to investigate how certain policy changes might affect the mechanism.

If we ask a what should be question, we build on what might be and seek to identify a mechanism that produces a favorable outcome with respect to a particular quality measure. This is frequently the issue in policy modeling. It is obvious that if one wants to understand the effect of particular policies in a target system, the model should exhibit sufficient faithfulness and empirical adequacy.

The key principle for this step is to keep track of the purpose of the study carefully during the development process because it influences all other steps of the process and because the purpose can change during the development. It is helpful to spell out the purpose explicitly to ensure that all stakeholders agree on what they are doing. In case the description of the purpose is a more complex one, agreements as to the meaning of individual terms should be established explicitly and early. To use the language of the last chapter, this establishes the initial common ground of stakeholders and reduces the gap between actual and perceived common ground, which serves the prevention of misunderstandings. When results are reported, the purpose of the model needs to be stated clearly to avoid overestimating the power of the model (Helbing and Balietti, 2011).

3.2.2 Conceptual Modeling and Validation

The goal of the second step in modeling is a conceptual model, or actually the model if one considers the implementation of this model not to be more than a computer program. The conceptual model provides the information that is necessary for the implementation and it is the artifact that is usually reported when the results of the study are presented. Developing the conceptual model is the process that we referred to as the construction of the model in Section 3.1.1. We described this process as a creative endeavor that draws on theory, data, other models, and the modeler’s input. We discussed that the construction of the model, in turn, enables learning about the underlying theories and models and about the target system.

Complexity vs. Simplicity

From a high-level perspective, the construction of the conceptual model is mainly determined by trading off simplicity against complexity, guided by the purpose of the model (Carley, 2009; Goldstone and Janssen, 2005). At one end of the scale, models with simple environments, agents, and interactions afford more easily interpretable outcomes and explanations,
3.2. Developing Agent-based Models

usually a better generalizability of results, and simpler implementations that allow computationally less expensive simulations and therefore more agents. These models would typically cover a smaller scope on the nano-to-macro axis, usually by representing agent behavior by simple rules. At the other end of the scale, models with detailed environments, agents, and interactions afford the grounding of the model in empirical data, which serves validation. However, complex models also consume more processing time per agent interaction, which prevents simulations with a large number of agents, and they would normally require more effort during implementation. The interpretation of results delivered by complex models is more complicated because of the larger number of parameters and their non-linear interactions. In terms of the learning that the construction process affords, more complex models are likely to evoke more questions about underlying theories and models. Therefore, complex models are particularly helpful in reflecting on theories and other models used in their construction.

This trade-off basically yields a scale from models that represent abstract processes, to empirically-grounded models, to history-friendly models, and finally to one-to-one simulation models (Gilbert and Ahrweiler, 2009). While some fiercely argue for models that are grounded in empirical data (e.g. Carley, 2009; Edmonds and Moss, 2005; Edmonds, 2010), others uphold the value of simple, possibly abstract models (e.g. Axelrod, 1997a; Macy and Willer, 2002; Smith and Conrey, 2007). Yet these two sides of the discussion just seem to reflect two different purposes of agent-based modeling: On the one side are those that see agent-based models primarily as representations of target systems (which is particularly true for policy modeling, for example). On the other side are those that see agent-based models as test beds for theories or hypotheses. In this latter case, insights are often easier gained from simpler models. We agree with Gilbert and Ahrweiler (2009) that each side in this discussion is but one point on a continuum between abstract models and one-to-one simulation models. The particular point one adopts on this scale depends on the purpose of the study at hand.

A few principles can be derived easily from this discussion. For example, if the results of the study are to be generalizable, one should typically aim for a simple model. Other principles follow directly from the discussion above. The conventional technique for finding an appropriate spot on the complexity-simplicity scale is to start with a simple model and to add more detail when required (Gilbert, 2007; Macy and Willer, 2002). This method is justified by the KISS principle. The opposite approach is taken if the KIDS (“Keep it descriptive, stupid!”) principle is followed (Edmonds and Moss, 2005). Edmonds and Moss claim that because of the complexity of social dynamics, often it is more reasonable to start off with more complex models and simplify only if warranted by model and evidence. Another alternative is pattern-oriented modeling, which was developed by agent-based modeling researchers in ecology (Grimm et al., 2005). In pattern-oriented modeling, the
model is constructed such that it can accommodate structurally for multiple patterns in the target system (in fact the model is also validated against these patterns). The idea is that thereby the model becomes more similar to the real mechanism at work. This improves the testable predictions the model makes and therefore its validation.

The Environment, Agents, Objects, and Interactions

From a low-level perspective, the construction of the conceptual model is concerned with the specification of the environment, objects, agents, and their interactions (Gilbert, 2007). These entities can typically be deduced from the purpose of the model, the underlying theories and models, as well as empirical evidence. An approach to inform modeling in particular by input from stakeholders (e.g. domain experts) is termed *companion modeling* (Barreteau et al., 2003; Bousquet et al., 1999). It is the role of conceptual validation to determine whether a model’s assumptions appropriately represent the theories, models, and available evidence for the purpose of the study. In that, conceptual validation is a tool for establishing the faithfulness of the model to theories and other models. A common technique for conducting conceptual validation is *face validation*, which relies on inspections of the model by domain experts. Companion modeling ensures that the model is conceptually validated with respect to stakeholders’ views (Moss, 2008).

Some seem to postulate that agents and their attributes and behavior should be specified before their interactions (e.g. Gilbert, 2007; Maxwell and Carley, 2009). We find that this is only warranted when the information utilized in creating the model (e.g. theories, models, data) is centered on the attributes and behavior of individuals. Instead, if this information pertains mainly to the interactions between actors, we claim that it is more helpful to specify interactions first. The types of interactions that are required presuppose what type of social behavior agents can and should exhibit and therefore which cognitive capabilities are required from them.

A crucial part of conceptual modeling is the identification of independent variables (parameters and initial conditions) and dependent variables. It is important to identify how both relate to the target system and the purpose—research question—of the study. That is, already during conceptual modeling some thinking should be spent on the analysis that needs to be performed to achieve the purpose. Building on existing theories, models, and empirical evidence obviously increases the confidence one can have in the faithfulness of the model. So does the grounding of variables in the target system.

Those that assume causal powers on macro-level properties might also represent these properties explicitly (Sawyer, 2004). Even if one adopts the position of methodological individualism, representing macro-level properties explicitly can facilitate the implementation of the interactions between agents and macro-level properties.
Auxiliary Models

We discuss in the following how other models might be employed to inform model construction. In particular, we suggest the construction of two types of models in support of the modeling effort: computational-conceptual models and formal logics models. Our suggestion is that both types of models are particularly suited for guiding the interdisciplinary modeling dialogue and for informing the construction of agents and their interactions in agent-based models of social dynamics. Before we make this argument, let us first describe what we mean by these two types of models.

By a computational-conceptual model we mean a semi-formal description not unlike a verbal-conceptual model that, however, relies particularly on computational concepts. In contrast to computational models, a computational-conceptual model is still not a precise description but its language is closer to the one of computational models than is the language of verbal-conceptual models. Yet at the same time a computational-conceptual model retains the benefit of verbal-conceptual models because it is more expressive than a computational model and because it is more accessible to non-technical stakeholders. By a formal logics model we mean a precise and complete description of an issue in terms of formal logics. Formal-logics models allow an analysis by logical inference but because their complexity can be quite substantial, we do not assume that they are directly implementable.

How can these two types of models inform the development of agent-based models? In the case of agent-based models of social dynamics, modeling can rely on a rich portfolio of research concerned with the mental attitudes that drive human behavior and the interactions between actors. Among these mental attitudes are beliefs, desires, goals, intentions, and commitments. This research has its roots in philosophy, for example based on a folk psychology of human reasoning (the BDI or Belief-Desire-Intention model, Bratman, 1987). The multi-agent systems community has provided computational and formal logics models based on computational representations of these ideas (refer to Wooldridge, 2009, for a comprehensive overview). We suggest to harness this research for the construction of computational-conceptual and formal logics models that inform the construction of agents and their interactions in agent-based models of social dynamics.

Employing insights from multi-agent systems research and philosophy for the computational modeling of human cognition and social dynamics is not novel (see for example Castelfranchi, 2006; Dignum et al., 2008). In fact, there is a recent argument for aligning formal philosophical research closer with empirical research, and in particular cognitive science (van Benthem, 2008). Philosophers can contribute to formal logics of higher-order reasoning about mental states of others such as their beliefs and intentions (Verbrugge, 2009), which is obviously an essential aspect of joint activities.
Chapter 3. Methodology

Note that computational-conceptual models are underspecified. However, they can still serve as frameworks that can be evaluated rigorously through concrete models that instantiate these frameworks.\(^5\)

From our perspective, computational-conceptual and formal logics models sit in between verbal-conceptual and computational models. They allow for expressiveness, which is important for describing the complexity of social dynamics, and they also rely on computational concepts, which map so neatly onto mechanisms at all levels of social dynamics. However, these types of models do not go all the way down to the simplifications necessary for actual implementations. In that, they allow for intermediary steps in the modeling process. Employing these auxiliary models has at least the following benefits:

- The gap between verbal-conceptual models, which typically describe social behavior and dynamics, and computational models is bridged by computational-conceptual and formal logics models, which establish computational concepts as a kind of lingua franca of the modeling process.

- Discussions about the model can be conducted in terms of computational concepts very close to the understanding that non-technical stakeholders have of the modeling problem. This helps establishing a common ground based on computational concepts as the appropriate language for these issues.

- Computational-conceptual and formal logics models can serve as frameworks for different concrete implementations. This helps to understand which parts of a theory or model the final computational model represents and which parts it does not represent.

- Formal-logics models enable the analytic exploration of the consequences of the mechanisms that they represent.

- Formal logics models are well suited to specify the representations and processes of cognitive models that rely on logic-based processing.

Obviously, these benefits weigh more in the case of complex modeling tasks when an understanding of the problem at multiple levels of detail might be helpful.

In effect, computational-conceptual and formal logics models undergo a similar modeling process as the one described here, except that they are not implemented. As indicated before, formal logics models afford a more rigorous analysis than computational-conceptual models do. We expand later on this discussion. One can also consider multiple of these models being

\(^5\)This observation is due to Hommel et al. (2001) in their description of the theory of event coding, which in our terminology is a computational-conceptual model. The theory of event coding plays a role in Chapters 7 and 8.
3.2. Developing Agent-based Models

arranged in a hierarchy with the more faithful ones at the top and those closer to the final computational model at the bottom. This would represent the iterative specification of the model. In general, however, we do not require that all these models cover the same scope as the final model. For example, one could employ one auxiliary model to represent one part of the final model and another auxiliary model for another.

Constructing models at varying levels of detail to facilitate the modeling process is not new in the context of agent-based models. At least three advances similar to our suggestion have been made (Cioffi-Revilla, 2010b; Edmonds and Moss, 2005; Hales, 2010). Cioffi-Revilla proposes that in order to mitigate the complexity of agent-based modeling tasks, researchers should create a sequence of increasingly complex models, with the final model being the one to be investigated. In contrast to that, our suggestion is that in fact more detailed models can support the construction of less detailed ones. This is similar to the KIDS approach mentioned before (Edmonds and Moss, 2005). Yet it appears that the relationship between models that we postulate is weaker than the one postulated by Edmonds and Moss. In our case, auxiliary models are not necessarily supposed to represent the target system at the same scope as the final model does. Instead, auxiliary models can provide insights into parts of the modeling task. Hales (2010) discusses chains of different models in which two models are linked if they share a common property. This perspective could be considered to subsume the two other approaches mentioned before and in fact ours as well. However, we consider how particularly computational-conceptual and formal logics models can be used to inform further modeling. At the very least, we claim that the construction of computational-conceptual and formal logics models contributes to the general understanding of a theory, model, or target system. This serves a scientific or policy modeling process that extends beyond the particular model (Moss, 2008).

3.2.3 Implementation and Verification

The goal of the implementation is to produce a computer program that represents the conceptual model. A computer program is expressed in terms of source code in a chosen programming language. In effect, this is a software engineering problem and much has been said about how to properly engineer software. However, we would like to state a few principles that are particularly relevant to agent-based modeling.

Almost always someone is at sometime going to come back to the source code produced during an agent-based modeling project, be it to extend the code or to gain a better understanding of its mechanisms or to reproduce its result. This someone can of course be the person that created the original version but it does not have to. Either way, it is essential to keep this observation in mind during the implementation. It is easier to work with the source code later if a few principles are adhered to during its initial development:
Chapter 3. Methodology

**Simplicity** Simple source code refrains from making excessive use of features particular to the programming language used but not shared by other common languages. Simple source code expresses complex calculations in multiple lines of code and does not summarize them in obscure one-line statements. Simple source code relies on the smallest necessary set of custom data structures (e.g. classes).

**Understandability** Understandable source code is well documented, including the relation between variables in the code and those used when reporting the results of the model. Understandable source code uses meaningful names for variables, data structures, and functions. Understandability is also improved if a common programming language is used instead of an exotic one.

**Modularity** Modular source code represents unrelated concerns in separate modules and ensures low coupling between different modules but high cohesion of the functionality within a module. Modularity can be facilitated by abstraction, which allows functionality shared between different modules to be extracted.

Obviously, these principles can be at odds sometimes. It is up to the developer to make informed decisions by weighing the consequences of choices against each other, also taking into account the purpose of the study at hand. Together, these principles increase the chances that someone can follow what happens when the program is executed and that someone can extend the original code, for example by exchanging a module by a different implementation.

An essential requirement of the implementation is that it is correct with respect to the conceptual model. The process of establishing that this state holds is called *verification* or *internal validation*. *Testing* and *replication* are two essential techniques of verification.

Testing is the activity of triggering failures of the program by specifying inputs and determining whether outputs are as expected according to the conceptual model. If a failure is observed, *debugging* is performed to determine the fault(s) in the source code that caused the failure. Testing is facilitated by modular source code because individual modules can be tested independently of others.

Replication is the process of developing two or more independent computer programs for the conceptual model by different research groups (Edmonds and Hales, 2003). If all replicates produce the same output or statistically indistinguishable outputs\(^6\), the confidence in the internal validity of the models is improved. If replication fails, at least one of the implementations is faulty or the assumptions on which the replicates are based have not been laid out accurately. From this follows that in order to allow other researchers to replicate an

---

\(^6\)Care is to be taken with regard to differences between the outcomes of replicates that are due to constraints of the programming languages used, e.g. floating point number precision or random number generation (Hales, 2010).
3.2. Developing Agent-based Models

agent-based model, the presentation of the conceptual model should lay out all assumptions correctly and unambiguously. Sometimes this is not possible within the limited space of a publication. In that case, full details of the conceptual model should be provided as supplementary material.

Care is to be taken when the conceptual model is translated into data structures and algorithms in order to avoid creating faults; and programming languages and toolkits have to be chosen wisely. Higher-level languages shield the unexperienced programmer from dirty mistakes but they also come with a performance trade-off. Lower-level languages often lead to more efficient machine code but they also complicate the implementation. A good way to go seems to implement the model in a higher-level language first in order to gain experience with the computational problems involved, and then to translate this code into a lower-level language. As a byproduct of following this principle one obtains a weak form of replication (the models were not developed by independent research groups). Of course, simplicity, understandability, and modularity also reduce the chance of introducing new faults when working with the source code.

Because the source code is likely to be modified frequently during its lifetime, the code should be entrusted to a version control system, and results should be clearly attributed to the respective version of the code with which they were generated.

We encourage publishing source code alongside the description of the conceptual model. This gives others the chance to reproduce results and might lead to interesting extensions of the original model. It also serves the credibility of reported results because the method with which they were generated is fully accessible. Furthermore, replications are facilitated (even though this might lead to replicates that are not independent).

### 3.2.4 Analysis and Operational Validation

As indicated in Figure 3.2, analysis serves two goals: Setting up the initial conditions and parameters of the model, and learning about underlying theories, models, or a target system, which amounts to providing an answer to the research question of the study. Associated with the analysis is the operational validation of the model. Operational validity is sometimes called *external validity*. In the typical course of agent-based modeling, operational validation is about establishing empirical adequacy, i.e. whether the model reproduces evidence from the target system with sufficient accuracy. Both analysis and operational validation are obviously determined by the purpose of the study (Carley, 2009). Because there is not a single clear process to follow, we comment on various issues and techniques independently and outline in the individual cases how they might relate to other points in this section. We do not elaborate on the point of analysis because it is too dependent on the purpose of the study.
However, analysis typically amounts to hypothesis testing on data obtained by simulating the model.

**Stochasticity**

An important point to consider during analysis and validation is that agent-based models typically contain stochastic elements, sometimes also called *random noise*. Stochastic elements stand for those parts of the represented mechanism that are not explicitly modeled, e.g. because there is no sufficient information available about these parts (Edmonds, 2012). The implicit assumption is that in the average the introduced noise does not affect the outcome of simulating the model. Because of these stochastic elements, agent-based models can usually not be studied as deterministic systems. Any numerical study of the model’s dependent variables needs to be conducted with statistical tools on the outputs of multiple executions of the simulations. When claims are made about numerical differences between the dependent variables under varying parameters, for example, the statistical significance of the differences between the sets of simulation runs needs to be established. At least, the location and variance (for example means and standard variance) of the data needs to be reported (Helbing and Bani, 2011).

**Empirical Evidence**

Both for analysis and validation, empirical evidence is often crucial. Therefore, we consider briefly what type of empirical evidence might be available. On one hand, one can distinguish qualitative and quantitative data: Qualitative data is usually obtained from role-play with stakeholders, focus group discussions, interviews or by broadly generalizing quantitative observations (*stylized facts*, Kaldor, 1961). Qualitative data cannot be measured on a numerical scale. Quantitative data can. On the other hand, one can categorize empirical evidence along two other dimensions: whether the evidence is generalized or context-specific, and whether the evidence relates to few or many individuals (Janssen and Ostrom, 2006). Stylized facts are generalized information about many individuals. Lab experiments yield generalized evidence about few individuals. Role-play and companion modeling yield context-specific evidence about few individuals. Case studies yield context-specific evidence about many individuals.
3.2. Developing Agent-based Models

Parameter estimation

Parameter estimation or equivalently\(^7\) calibration employs empirical evidence to restrict the range of initial conditions and parameters that needs to be considered during analysis. This can happen in two ways: First, some model parameters are closely linked to known parameters in the target system and therefore they can be set directly with available data. The same applies to some initial conditions. Second, some parameters and initial conditions can be estimated indirectly. Indirect estimation proceeds by simulating the model under varying parameters and initial conditions and retaining only those instances for which the output of the model corresponds to some available empirical evidence. Sometimes model output is compared only qualitatively to the target system, for example with stylized facts. If empirical evidence can be compared quantitatively to the model output, a more rigorous parameter estimation is possible. Techniques enabling a rigorous estimation are based on statistics, for example approximate Bayesian computation (Beaumont, 2010). Instead of conducting parameter estimation with a statistical perspective, one can also conceive this problem to be one of parameter optimization, i.e. finding that setting of parameters and initial conditions that yields model outputs that align most closely with the empirical evidence. We apply this technique in Chapter 5. This perspective makes tools for parameter optimization, e.g. evolutionary algorithms, available for parameter estimation. As we indicated in Section 3.1.1, parameter estimation puts the model in the role of a model of data.

Sensitivity Analysis

Sensitivity analysis is a host of techniques that determine how changes in model assumptions and inputs (parameters and initial conditions) affect model outputs. Possible changes in model assumptions and inputs include stochastic elements (as discussed above), variations of the random distributions from which these stochastic elements are drawn, parameter variation, temporal variation of model elements, variation in the level of data aggregation, variation in the mechanisms implemented by the model, and variations of sample size (Richiardi et al., 2006). By far the most common variations used in sensitivity analysis are stochastic and parameter variations.

The extent to which sensitivity analysis is conducted depends on the purpose of the study (Grüne-Yanoff and Weirich, 2010). In the simplest case, sensitivity analysis amounts to an exhaustive enumeration of parameters and initial conditions to study a sufficiently large proportion of the space of independent variables. Alternatively, sensitivity analysis can be applied locally around the estimated parameters (Richiardi et al., 2006). Hence, sensitivity\(^7\) We side with Richiardi et al. (2006) that a distinction between parameter estimation and calibration is not helpful or at best artificial.
analysis can happen before or after parameter estimation. Sensitivity analysis can be part of the actual analysis of the model, contributing to an understanding of the behavior of the model. But sensitivity analysis also contributes to operational validation because it can indicate how robust a particular result is with respect to changes in model assumptions and inputs. In effect, sensitivity analysis determines the range of parameters in which the model behaves as expected.

Sensitivity analysis is helpful in detecting whether assumptions in the model that are not deemed crucial have any significant effect on the model’s output. Such artifacts can occur because usually when specifying the conceptual model or the implementation, additional assumptions have to be made to obtain a fully specified model or an executable implementation respectively (Galán et al., 2009).

**Operational Validation**

Operational validation determines empirical adequacy, i.e. whether the model’s output is sufficiently accurate with respect to data generated from the mechanism it represents. The data used for determining operational validity can obviously not be the same as the one used for (indirect) calibration. For if a model is sufficiently complex, it can be calibrated to pretty much any data set. In that, testing for operational validity determines whether the model is appropriate beyond the particular case to which it was calibrated. Therefore, validation is an essential step to ensure that the model is not over-fitted. Operational validity can be established, for example, by face validation or by statistical comparison between model outputs and empirical data.

**Auxiliary Models**

We mentioned before that we would return to the analysis of computational-conceptual and formal logics models. Computational-conceptual models can be analyzed by thought experiments. Similarly they can be validated by demonstrating how they could qualitatively reproduce some empirical evidence. Formal logics models can be analyzed by formal proof procedures and it can be demonstrated analytically that they are capable of reproducing empirical evidence.

**3.2.5 Summary**

This section has elaborated on the process of developing agent-based models. We refer back to the discussion here in Section 3.4, when we outline how this thesis answers the research questions presented in Section 3. For more in-depth discussions of the issues surrounding
3.3 The Limits and Pitfalls of Computational Modeling

Before discussing how we intend to answer our research questions, let us reflect on the limits of computational modeling of social dynamics.

The key “limitation” of any modeling effort in general is that no model is ever going to be true or correct in an absolute sense. The purpose of modeling is to create a manageable representation of a theory, model, or target that is sufficiently valid for a given purpose until proven otherwise. Failing to understand this leads to the fallacy of creating ever more complex models that are falsely assumed to be more true or correct than simpler ones. In the hypothetical extreme case, a model becomes a one-to-one representation of a target system. But this model obviously is not any more manageable than the target system itself and therefore does not have any value as a model. It is crucial to find the trade-off between simplicity and complexity that is most appropriate for the purpose at hand.

Another general problem of modeling are those assumptions a modeler includes in the model without awareness because they are part of the implicit view that the modeler has adopted about the problem. Similarly, there might be assumptions that the modeler is aware of but either their number prohibits to include all of them in the model or there is not sufficient evidence as to how they should be included (Edmonds, 2010).

Computational modeling affords a few problems not found in other types of modeling. One of the most serious problems is probably faults in the source code. Implementations of complex models require extensive testing and even then a fault-free program is not guaranteed. The problem is likely to be more severe when the implementation is produced by an inexperienced programmer. In the opposite case when the implementation is created by a programmer and/or the conceptual model by a computer scientist, the modeling process runs the risk of being dominated by technical considerations (Padgett, 2000). This creates a situation in which peculiarities of computational tools become illegitimate constraints to the modeling process. This is in particular the case when computational efficiency is an issue.

In the case of computational models of social dynamics, we face yet further problems. Among the most important problems are that there is no single unifying theory of human behavior that could be leveraged and that there is usually no single data set with sufficient detail to calibrate and validate a model (Carley, 2009). These restrictions have a few consequences (Louie and Carley, 2008): The first restriction requires that multiple theories be used in the construction of a model but that might cause problems if one theory invalidates the assumptions of another. However, computational modeling makes it possible to study
the interaction of different theories at all. The second restriction impedes the grounding of the model in the target system. As we discussed before, however, the purpose of a model does not necessarily require a perfect grounding and even without, it is possible to explore the parameter space to learn about the implications of the model. Another crucial problem is that the intelligence of humans is extremely complex and that humans have the unique ability to develop models of the world on their own or possibly even to inspect the model that the modeler created (McBurney, 2012).

The complexity of typical computational models of social dynamics renders their analysis difficult. A large parameter space usually prevents an exhaustive coverage of the space of independent variables. Therefore, it is difficult to identify the part of the parameter space in which the model operates as expected (Louie and Carley, 2008). It also implies that the analysis of the data obtained from the model is difficult.

Ultimately, computation as we know it today is just another tool for modeling and it might turn out that there is a more appropriate one for the representation of social dynamics. Above all, social dynamics is an extremely complex target system and it is still debated whether any formal modeling can ever be successful. Yet considering the topic of this thesis, we obviously believe in the value of computational modeling of social dynamics.

A more complete discussion of the pitfalls of agent-based modeling is provided by Barth et al. (2012) and a more complete discussion of problems with modeling in general by McBurney (2012).

3.4 The Models in this Thesis

In this section, we relate the content of the following chapters to the discussion of this chapter and in particular outline how the following chapters address our research questions. Each of the following chapters introduces a model of cultural transmission or dynamics based on the grounding model of cultural transmission. In each of those chapters, we identify the premises and implications from Section 2.2.3 that the respective model accounts for and indicate how the model advances our understanding of cultural dynamics and of the grounding model of cultural transmission. The first part contributes to answering RQ1, the second part contributes to answering RQ2. The following paragraphs provide an overview of those chapters.

Chapter 4 addresses the question of how the co-evolution of cultures, social networks, and geographical locations can be simulated computationally. To this end, we propose an agent-based model that represents the actual common ground between agents, the grounding process happening during interactions, and the effect of this process on
3.4. Towards Computational Representations of the Grounding Model

the migration behavior of agents and the change of their social contacts. Thereby, we contribute to answering RQ1. We study the model in detail, in particular considering a varying mobility of agents and different population sizes. This analysis contributes to answering RQ2 by shedding light on cultural dynamics.

Chapter 5 addresses the question of what is a suitable model for studying the communication of stereotype-relevant information in the context of the grounding model of cultural transmission. As an answer to this question, we propose an agent-based model in which agents adjust their communication about stereotype-relevant information based on a representation of the extent to which the stereotype is common ground with their interaction partners. We calibrate this model to empirical data and then validate it on another, previously unknown data set. Thereby, we contribute to answering RQ1. Providing confirmation that the mechanisms postulated by the grounding model of cultural transmission can give rise to the empirical data contributes to answering RQ2. In a second part of that chapter, we explore how the rate with which agents are subjected to first-hand or second-hand information affects the communication of stereotype-relevant information in a network of these agents. This is another contribution towards answering RQ2.

Chapter 6 addresses the question of what is an appropriate more formal description of the basic grounding model of cultural transmission. The proposed answer to this question is a computational-conceptual model of the interplay between the joint activity, the grounding process, and common ground. At the core of that chapter is a formal logics model of common ground. The chapter contributes to RQ1 by providing a more comprehensive model of the grounding model of cultural transmission. We analyze the properties of the common ground model formally and we demonstrate how the overall model can represent the interplay between the joint activity, the grounding process, and common ground. This analysis as well as the precision contributed to the grounding model of cultural transmission is that chapter’s contribution to answering RQ2.

Chapter 7 addresses the question of what are the essential ingredients for a computational architecture of joint activities that considers both coordination due to higher-level cognitive processes and coordination due to lower-level processes based on direct links between perception and action. We propose an answer to this question in terms of a computational-conceptual model. Thereby that chapter contributes to answering RQ1 because one of the essential ingredients of the grounding model of cultural transmission are joint activities. We show that this model can incorporate the model of grounding presented in chapter 6. We also show that the model takes into account the relationship between the different linguistic and non-linguistic levels of alignment.
### Chapter 3. Methodology

Table 3.1: This table illustrates the relation between the models presented in the following chapters and their type and coverage of the nano-to-macro axis.

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Model</th>
<th>Type</th>
<th>Nano-to-macro axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Co-evolution</td>
<td>Agent-based</td>
<td>Micro/Macro</td>
</tr>
<tr>
<td>5</td>
<td>Stereotypes</td>
<td>Agent-based / Connectionist</td>
<td>Nano/Micro/Macro</td>
</tr>
<tr>
<td>6</td>
<td>Joint Activity, Common Ground, &amp; Grounding</td>
<td>Computational-conceptual / Formal logics</td>
<td>Nano/Micro</td>
</tr>
<tr>
<td>7</td>
<td>Joint Action Architecture</td>
<td>Computational-conceptual</td>
<td>Nano/Micro</td>
</tr>
<tr>
<td>8</td>
<td>Joint Action Implementation</td>
<td>Logics-based / Connectionist</td>
<td>Nano/Micro</td>
</tr>
</tbody>
</table>

in joint activities, which is relevant to a more holistic perspective on grounding and common ground. By discussing the implications of this model for the grounding model of cultural transmission, that chapter also contributes to answering RQ2.

**Chapter 8** provides a simplified version of the model from Chapter 7. This simplified version is implemented to represent an experimental task that has been found to evoke an interaction between high-level cognition and low-level perception-action links. This chapter mainly strengthens the argument in favor of the architecture proposed in Chapter 7. Also, that chapter demonstrates that a computational-conceptual model is a helpful intermediary between a verbal-conceptual model and a concrete simulation model. By providing an implementation of the joint action architecture, that chapter also contributes to RQ1.

Table 3.1 summarizes the role of each of the following chapters in the context of this thesis. In essence, we start off by investigating how the grounding model of cultural transmission can be employed to inform the development of larger scale agent-based models of cultural dynamics, which consider a large scope on the nano-to-macro axis including the nano level. Then we turn to the development of computational-conceptual and cognitive models that are more faithful to the grounding model of cultural transmission than the previous models and restricted to the nano and micro level. In the last part we present cognitive models that in fact even surpass the current theorizing around the grounding model of cultural transmission by drilling into the details of joint action.
3.5 Conclusions

This chapter has elaborated on the role of models in scientific discovery with a focus on computational models. We have adopted an understanding of models as intermediaries between theory and data. We have emphasized the relationship between computational primitives and the constituents of mechanisms, highlighting that mechanisms are suitable descriptions of social dynamics at all levels of the nano-to-macro axis. This contributes to the value of computational models in the study of social dynamics. Because agent-based models in particular afford representations along the whole spectrum of the nano-to-macro axis, we have identified them as the appropriate type of computational models for studying cultural dynamics. However, we have also highlighted the value of cognitive models. Cognitive models can be used to add detail to agent-based models and they can provide more detailed formal descriptions.

We have described our understanding of the process of developing agent-based models and suggested utilizing two types of additional models in their construction: Computational-conceptual models are verbal descriptions that make particular use of computational concepts, thus enabling expressive descriptions of phenomena without the restricting requirement of being implementable. Formal-logics models are expressive and can at the same time be studied rigorously, yet typically without being implementable. Both computational-conceptual and formal logics models can bridge the gap between verbal-conceptual models, which are the prevalent type of descriptions for social dynamics, and computational models. They establish computation as the dominating language in the modeling process.
Chapter 4

An Agent-based Model of the Co-evolution of Cultures, Social Network Communities, and Agent Locations

As discussed in Section 3.4, this chapter is concerned with the question of how the co-evolution of cultures, social networks, and geographical locations can be simulated computationally. The main premise here is that cultures co-evolve with the geographical locations of people and with their social contacts. For example, the transmission of cultural information is often directed by physical proximity and social relationships (Latané, 1996). Cultural differences or similarities, on the other hand, have often caused people to change their physical location or social environment (Castles, 2002; Centola et al., 2007). Hence, an understanding of cultural dynamics requires an understanding of the interaction between cultural, social, and physical space. This observation is particularly relevant to the effect of globalization on the evolution of culture, considering that globalization has opened a whole new world of opportunities for migration and communication. By providing an agent-based model, this chapter contributes to this research agenda.

Given the abstract nature of the question, we build on an abstract agent-based model of cultural dynamics that was originally proposed by Axelrod (1997b). We modify this model by making use of the premises and implications of the grounding model of cultural transmission in order to incorporate the co-evolution of cultural, social, and physical space.

Axelrod’s model is essentially a cellular automaton. The elements of this automaton, which we call agents, are arranged on a lattice. Every agent is represented by a vector of
Chapter 4. A Model of the Co-evolution of Cultures, Networks, and Geographical Locations

$F$ cultural features, each of which holds one of $q$ nominal cultural traits at a time. As an example, consider a feature vector with $F = 3$ features, standing for religion, language, and meat preference respectively.\(^1\) Assume $q = 3$ cultural traits per feature. The possible traits for the religion feature could be Islamic, Jewish, and Christian. The traits for language could be Arabic, Hebrew, and English. The traits for meat preference could be pork, beef, and chicken. Note that $q$ determines the possible cultural diversity in the population. The more traits per feature there are, the more different cultures can be constructed.

Through a number of iterations, agents interact with their direct neighbors according to two simple rules: (1) Agents are more likely to interact the more similar their feature vectors are; (2) an interaction leads to one agent copying a feature of the other agent. Hence these rules represent two well established empirical regularities: Individuals prefer to interact with similar others (homophily, Lazarsfeld and Merton, 1954; McPherson et al., 2001) and interactions typically cause individuals to become more similar (social influence, Mason et al., 2007).

Figure 4.1 depicts a possible situation in this model. In this part of the lattice, three cultures coexist. While the culture represented by the green square shares some features with both other cultures and is therefore culturally compatible with them, the ones represented by the blue diamond and the red circle do not share any features. Therefore, these latter two cultures are culturally incompatible—members of these two cultures cannot interact. Note that in this sense, Muslims and Jews as described in the example above would be culturally compatible while Jews and Christians would not be. Of course, this is a hypothetical case and a simplified description. Obviously, there are other potential interpretations of cultural features, similarity and compatibility. However, we stick here to Axelrod’s interpretation.

The dynamics of the model are influenced by the initial cultural diversity in the population, determined by $q$, and the ordering forces of homophily and social influence. Eventually, the process reaches a final state or set of states, which is characterized by the configuration of cultural regions on the lattice. Agents within the same region share the same culture but agents on the borders between regions do not share any cultural features and therefore cannot interact anymore. It turns out that with increasing $q$ the final configuration of cultures transitions from a mono-cultural to a multi-cultural phase. This finding is interesting from a statistical physics perspective as this transition corresponds to a non-equilibrium order-disorder phase transition (Castellano et al., 2000). In fact, sociophysics, which employs statistical physics tools to conceptualize social systems in terms of interacting particles, has made wide-ranging contributions to the study of social phenomena such as opinion formation and crowd behavior (Castellano et al., 2009). Axelrod’s model has been studied from this perspective as well.

\(^1\)We owe this example to an anonymous reviewer.
Figure 4.1: A possible situation in Axelrod’s model: Agents with different cultures occupy the cells of a lattice (in here only a part of the lattice is shown). The culture represented by the green square shares some features with both other cultures. Members of this culture can interact with members of both other cultures. The other two cultures do not share any features. Therefore, an interaction between members of these two cultures is impossible.

These studies have mainly investigated the order-disorder phase transition on various extensions of Axelrod’s original model (Castellano et al., 2009). For example, the influence of random perturbations of cultural features (Klemm et al., 2003a, 2005), interaction thresholds (De Sanctis and Galla, 2009; Parravano et al., 2007), or external fields as representations of mass media (González-Avella et al., 2005, 2007; Mazzitello et al., 2007; Rodríguez and Moreno, 2010) have been investigated. Others have studied the model’s dynamics on different static network structures (Guerra et al., 2010; Klemm et al., 2003b), with different models of social influence (Flache and Macy, 2008; Kuperman, 2006; Rodríguez and Moreno, 2010), or with larger neighborhood sizes (Greig, 2002).

In Axelrod’s model, the interaction structure or social network between the agents is fixed, as are their locations on the lattice. Some attempts have been made to investigate the co-evolution of culture and social networks or physical locations (Centola et al., 2007; Gracia-Lázaro et al., 2009; Vazquez et al., 2007). However, the simultaneous co-evolution of cultural, social, and physical space has not been considered yet.

This chapter introduces an extension of Axelrod’s model to investigate the co-evolution of cultural, social, and physical space by accounting for the pair-wise interactions between these three spaces. The model is defined in accordance with some of the premises and implications of the grounding model of cultural transmission identified in Section 2.2.3. Thereby this chapter contributes to **RQ1**. The model presented in this chapter does not assume any detailed representation of behavior, so that it can be located around the micro-macro area of the nano-to-macro axis. We analyze the model by simulation and focus on the evolving configuration of cultures and networks under varying conditions: alternative lattice sizes, population densities, initial diversities, and social mobilities. Social mobility is a proxy for the probability of interaction depending on distance. In effect, we conduct
a global sensitivity analysis. The detailed analysis of this model and the discussion of our results is this chapter’s contribution towards RQ2.

We observe the emergence of one or more social network communities each of which corresponds to a particular culture. All members of the same community share the same culture while members of different communities are culturally incompatible, in that they do not share any cultural features. Before communities become culturally homogeneous, however, populations transition through an initial period in which social links are formed and the communities in the network become increasingly culturally heterogeneous. We find that a larger social mobility (higher probability of interaction over longer distances) accelerates the creation of social links and therefore supports the initial peak of diversity within communities. With a large social mobility, the number of communities and therefore cultures in equilibrium increases with increasing initial diversity, but decreases with increasing lattice size and population density. With increasing initial diversity, the dynamics transition from a phase in which a single network community and culture dominates the population to a phase with multiple culturally homogeneous communities. The critical value of initial diversity at which this transition occurs increases with increasing lattice size and population density, and in fact generally with absolute population size (which is determined by lattice size and population density). Hence, larger initial diversities promote heterogenization while larger lattice sizes, population densities, and in consequence absolute population sizes promote homogenization.

We describe our model in Section 4.1. In Section 4.2, we provide a formal analysis of this model. In Section 4.3, we present our analysis of the model by simulation. We synthesize and discuss our results in Section 4.4, considering also the connection with related studies. We conclude the chapter with a discussion in Section 4.5.

**4.1 Model**

In this section, we describe our model as an extension to Axelrod’s original work, accounting for the co-evolution of cultural, social, and physical space.

We retain Axelrod’s feature vector representation for agents and their arrangement on a non-toroidal square lattice of width $L$ that represents physical space. Let $\mathcal{A}$ be the set of all agents on the lattice with $|\mathcal{A}| = N$. We allow $N \leq L^2$ such that not necessarily all $L^2$ cells of the lattice are occupied by agents. By $N/L^2$ we denote the population density on the lattice. Agents are able to migrate to empty cells and to create social links with each other as we shall explain. We denote by $d_L(a, b) \in [0, 1]$ the Euclidean distance between two agents $a$ and $b$ on the lattice normalized by the longest distance on the lattice—the length of the diagonal determined by $L\sqrt{2}$. The social network between agents consists of undirected, weighted links and evolves over time as we will explain further on. The higher
the link weight $s(a,b) \in [0, 1]$, the stronger is the social connection between agents $a$ and $b$. We denote by $\sigma_{ai}$ the entry at index $i$ in the feature vector of agent $a$ and by $c(a,b)$ the cultural similarity between agents $a$ and $b$:

$$c(a,b) = \frac{1}{F} \sum_{i=1}^{F} \delta(\sigma_{ai}, \sigma_{bi})$$

where $\delta(x,y)$ is Kronecker’s delta with:

$$\delta(x,y) = \begin{cases} 
1 & \text{if } x = y \\
0 & \text{otherwise.}
\end{cases}$$

Agents $a$ and $b$ are culturally compatible if $c(a,b) > 0$. In effect, the feature overlap between two agents represents their common ground (PREM1, see Section 2.2.3) and the features shared in a group of agents is that group’s actual communal common ground, i.e. a superset of their culture (PREM4).

All feature vectors and agent locations are initialized randomly from a uniform distribution. The social network does not contain any links initially. The model is run for a number of iterations, each of which consists of the following phases.

### 4.1.1 Interaction

In every iteration, a focal agent $a$ is selected randomly from the population. In Axelrod’s model, one of the agents located in $a$’s von Neumann neighborhood is chosen to interact with $a$. In our case, on the contrary, the probability $I(a,b)$ of agent $b$ to be selected for interaction with a focal agent $a$ depends on their physical and social proximity. We assume that $I(a,b)$ declines with distance following a function of the form $y = d_L(a,b)^{-\beta}$ (Morrill and Pitts, 1967) and increases with the strength of their social tie:

$$I(a,b) = \begin{cases} 
\frac{1}{Z} \left[ d_L(a,b)^{-\beta} \max(s(a,b), \gamma) \right] & \text{if } a \neq b \\
0 & \text{otherwise.}
\end{cases}$$

That is, the composition of the social network at the macro level affects which agents engage in micro-level interactions and possibly cultural transmission (IMPL6).

Here, $Z$ is a normalization factor and $\gamma > 0$ ensures that an interaction is possible between agents that do not have any social link. The parameter $\beta$ controls the social mobility of agents. Larger values correspond to a smaller social mobility. Note that depending on these parameters, any two agents on the lattice could interact, which contrasts with Axelrod’s model in which interactions are restricted to direct neighbors on the lattice.
4.1.2 Common Ground and the Evolution of Cultural Space

In Axelrod’s model, two agents $a$ and $b$ interact with the probability of their cultural similarity $c(a,b)$ and in that case they always exchange some information. We slightly re-interpret Axelrod’s meaning of interaction: We consider every selection of two agents as an interaction. A successful interaction is one in which they do exchange information, an unsuccessful interaction is one in which they do not exchange any information. It remains to be defined when agents exchange information in an interaction: As noted before, a larger common ground increases the probability that information is exchanged successfully (PREM6, PREM7, IMPL2). Therefore, we assume that information exchange between two interacting agents occurs with the probability of their cultural similarity $c(a,b)$. In that case and if $c(a,b) < 1$, one of the agents’ features $i$ with $\sigma_{ai} \neq \sigma_{bi}$ is chosen randomly and $\sigma_{ai}$ is set to $\sigma_{bi}$ such that the two feature vectors become more similar. This represents a successful grounding process, during which cultural information is transmitted (PREM5) and communal common ground, and hence culture, at the macro level is modified (IMPL3). If no information exchange occurs, the grounding process is considered unsuccessful.

Overall, the mechanism that determines whether cultural exchange between two agents takes place is the same as in Axelrod’s model: It depends on their cultural similarity $c(a,b)$. However, agents in our model can interact even if they do not share any cultural features. In the example introduced above, Jews and Christians could interact but they would not exchange any cultural information. Of course, we would expect Jews and Christians to be able to exchange information during interaction. However, here this example represents the hypothetical extreme case of zero cultural overlap. It appears to be reasonable that in such a case, there cannot be any information exchange because the interaction partners do not have any starting point whatsoever for grasping the meaning of each other’s actions. Nevertheless, they might interact (in a primitive and potentially useless way).

4.1.3 The Evolution of Social and Physical Space

If the information exchange was successful, the social relationship between both individuals is likely to improve (Clark and Kashima, 2007), which is represented by an increase of their link weight by $\eta$, up to a maximum of 1. If the agents were unsuccessful in exchanging information, their link weight is decreased by $\eta$, down to a minimum of 0.\(^2\) The strength of the link update is arbitrary but discrete steps ensure that there is a limited number of possible network configurations, which serves the analysis of the model (see Section 4.2). This mechanism reflects the effect of successful and unsuccessful grounding on social relationships (PREM7, IMPL4).

\(^2\)We assume that 1 is a multiple of $\eta$. 
If the information exchange is unsuccessful, the focal agent \( a \) decides to migrate, which represents another consequence of an unsuccessful, socially-disruptive grounding process (PREM7). The agent migrates towards one of its social contacts, reflecting that migration tends to be directed towards social contacts (Castles, 2002). Migration is initiated by the agent selecting one of its social neighbors \( c \) randomly, with the probability of the weight of their relationship. Agent \( a \) then considers all empty cells as possible migration destinations that are closer to the selected social neighbor than the current position of \( a \). We assume that the probability \( M(a,t) \) of agent \( a \) to migrate to cell \( t \) is approximated well by an exponential function of the form \( y = e^{-\beta pd_L(a,t)} \) (Morrill and Pitts, 1967):

\[
M(a,t) = \begin{cases} 
\frac{1}{Z'}e^{-\beta pd_L(a,t)} & \text{if } d_L(t,c) < d_L(a,c) \text{ and } t \text{ is unoccupied}, \\
0 & \text{otherwise.}
\end{cases}
\]

Here, \( Z' \) is a normalization factor and the parameter \( \beta_p \) controls the agents’ physical mobility.

Migration is the part of the model that dominates the computational complexity of agent interactions. The computational complexity of migration is \( O(N+L^2) \) because there is an initial effort of \( O(N) \) to determine the neighbor \( c \) in whose direction to migrate and a subsequent effort of \( O(L^2) \) to determine the migration probabilities \( M(a,t) \) for each of the tiles \( t \) of the lattice. The complexity of all remaining operations is dominated by \( O(N+L^2) \).

This model actually does not carry over the direct representation of homophily from Axelrod’s work. In his model, larger cultural similarity directly leads to a larger probability of interaction. In contrast, we assume that the probability of interaction is primarily determined by geographical and social proximity. However, the model does include a kind of “network homophily” (McPherson et al., 2001) in that successful cultural transmission leads to an improved social relationship, which, in turn, increases the probability of interaction.

Note how cultural, social, and physical space are linked (Figure 4.2): Cultural similarities determine the choice of social contacts and cultural differences cause migration. Social contacts and physical proximity determine the probability of interaction and hence the probability of cultural transmission between agents. Social contacts guide the direction of migration.

### 4.2 Formal Analysis

In this section, we provide a tentative formal analysis of the model described in the previous section. We follow Izquierdo et al. (2009) and conceptualize the process represented by the model as a Markov chain. While this analysis cannot be exhaustive due to reasons
Chapter 4. A Model of the Co-evolution of Cultures, Networks, and Geographical Locations

Figure 4.2: The pair-wise interactions between cultural, social, and physical space in our model. The parts in solid blue lines are present in Axelrod’s original model, the parts in dashed red lines represent the extensions introduced here.

presented later in this section, the information provided here is useful and partly necessary for understanding the remainder of this chapter.

The state space of the Markov chain representing the process introduced above can be described by the allocation of cultures to agents, their physical locations, and their social links. There are $q^F$ possible different cultures, $\binom{L^2}{N}$ different allocations of $N$ agents to $L^2$ lattice cells, and $1 + \frac{1}{\eta}$ possible weights $(0.0, \eta, \ldots, 1.0 - \eta, 1.0)$ for all $\frac{N(N-1)}{2}$ inter-agent links. Thus the state space is discrete with size

$$q^{FN} \times \binom{L^2}{N} \times \left(1 + \frac{1}{\eta}\right)^{\frac{N(N-1)}{2}}$$

This Markov chain is time-homogeneous because the probability for a certain state to occur in the next time step is fully determined by the current state and independent of elapsed time. Note that the transition matrix $T$ is large but sparse. Though tedious, each transition probability $T(i, j)$ for the transition from state $i$ to state $j$ could be derived analytically. $T(i, j) = 0$ if the difference between states $i$ and $j$ requires more than a single step or when this change is not allowed by the rules of the model. The probability of every other transition is a function of the current cultural, social, and physical space. The following absorbing states and classes of states exist:

**Monocultural absorbing states** In these final configurations, we have $c(a, b) = 1$ and $s(a, b) = 1$ for all $a, b \in A$ and all information exchanges are successful. Hence, no change in cultural, physical, or social space occurs anymore and in each such
4.2. Formal Analysis

absorbing state $i$ we have $T(i,i) = 1$ and $T(i,j) = 0$ for $j \neq i$. The network consists of a single connected component. There are $q^F \times \binom{L^2_N}{2}$ of these absorbing states.

**Multicultural absorbing classes** In these final configurations, more than one culture survives. For two agents $a, b \in A$ from the same culture, we have $c(a,b) = 1$ and $s(a,b) = 1$. For two agents $a, b \in A$ from different cultures, we have $c(a,b) = 0$ and $s(a,b) = 0$. Hence, all agents with the same culture are members of the same connected component in the network. Agents within each of the components are clustered physically in such a way that migration is minimized. While network structure and cultural configuration are frozen, there can still be migration caused by unsuccessful interactions. Hence each such absorbing class consists of $\binom{L^2_N}{2}$ states that differ only by the physical locations of the agents. There are $\sum_{k=2}^{\min(N,q)} \binom{N}{k}$ of these absorbing classes, in which the network consists of 2 or up to $\min(N,q)$ connected components.

The Markov chain is reducible: The state space can be partitioned into $q^F \times \binom{L^2_N}{2}$ absorbing states, the $\sum_{k=2}^{\min(N,q)} \binom{N}{k}$ absorbing classes, and the class of all other states. Figure 4.3 provides an example of such a Markov chain. Eventually, the Markov chain reaches one of the absorbing states or absorbing classes, yet which one depends on the initial conditions and on the transition matrix. Initial cultures are drawn from a uniform distribution and the process does not depend on the exact values of cultural features. Hence, the probability for all cultures that are able to emerge from the original cultural distribution to be among the surviving cultures is approximately equal. However, the probabilities for observing particular geographical constellations differ as the process of homogenization affects migration. Furthermore, we can imagine to construct an initial condition manually that favors a certain culture due to geographical positions.

All states that do not belong to the set of absorbing states nor to any of the absorbing classes are just transient, that is, they have a non-zero probability that they will not be visited again. This is not only due to the fact that eventually the process gets stuck in an absorbing state or absorbing class but also due to the fact that for every such state $i$ there is a non-zero probability that later a state $j$ is visited whose population misses a cultural trait that was present in the population of $i$. Due to the rules of the model, a cultural trait that was once lost from the population, cannot be reinstated. In fact, depending on the initial cultural distribution, some states can never be visited at all.

The model dynamics always reach a monocultural absorbing state or multicultural absorbing class. However, it is difficult to obtain an analytical solution for deriving the

---

3 A connected component is a set of nodes in a network such that all pairs of nodes in this set are connected by a path through the network and such that no node in this set is connected to any node outside of this set.

4 $\binom{N}{k}$ denotes the Stirling number of the second kind—the number of ways a set of $N$ elements can be partitioned into $k$ subsets.
parameters under which either of the final configurations is obtained because the transition matrix is large and difficult to compute, i.e. computationally intractable. The reason are the size of the state space as well as non-linearities in the system’s behavior and feedback loops introduced by the co-evolution of the three spaces. Therefore, we opt for a numerical analysis, in line with previous approaches to the study of Axelrod’s model (González-Avella et al., 2006; Gracia-Lázaro et al., 2009; Guerra et al., 2010; Klemm et al., 2003b; Rodríguez and Moreno, 2010).

### 4.3 Sensitivity Analysis

In this section, we analyze the model with simulations and effectively conduct a global sensitivity analysis. The results reported here are discussed in the next section. For all simulations, we set $F = 5$ since studies of Axelrod’s original model indicate that results are qualitatively equivalent for all $F$ with $F > 2$ (Castellano et al., 2000). We fix $\beta_p = 10$, $\gamma = 0.01$, and $\eta = 0.1$ here to facilitate the analysis of the other parameters. We run simulations for varying levels of $L$, $N/L^2$, $q$, and $\beta_s$. Note that populations of size $N = L^2$ fill the lattice and therefore prevent migration. Large values of $\beta_s$ favor local interactions (small social mobility). Small values enable more frequent interactions over longer distances (large social mobility).

We discuss both the model’s behavior over time as well as its state in or near the equilibrium. Our focus lies on describing the emergent network structure and the configuration of cultural diversity within the network. To analyze the social network, we report on the number of communities, calculated by an algorithm that searches greedily for an allocation of network nodes to communities such that the modularity score of the network is maximized.

![Figure 4.3: A Markov chain with states (shapes) and transitions between these states (arrows). This Markov chain can be partitioned into a set of transient states (red circles), two absorbing classes (blue diamonds), and two absorbing states (green squares).](image-url)
4.3. Sensitivity Analysis

(Clauset et al., 2004). In terms of cultural diversity, we measure the overall diversity within the population and also the average diversity both within and between different communities.

The cultural diversity \( D(G) \) of any subgroup \( G \subseteq A \) is defined here by the average pairwise cultural dissimilarity between all members of that subgroup:

\[
D(G) = \frac{2}{|G| \times (|G| - 1)} \sum_{a,b \in G, a \neq b} 1 - c(a, b). \tag{4.1}
\]

Let \( C(A) \) be the set of communities within the network of agents \( A \). For every such community \( C_i \in C(A) \), we determine the cultural prototype \( P(C_i) \) of that group of agents, which is the feature vector that holds those cultural traits that are in the majority within this community. Now we can both calculate the intra-community diversity \( D(C_i) \) of a community \( C_i \) as well as the inter-community diversity \( D(\{P(C_i) | C_i \in C(A)\}) \) among the prototypes of all communities. We assume that \( D(\{P(C_i) | C_i \in C(A)\}) = 0 \) if \( |C(A)| = 1 \).

Figure 4.4 shows a sample snapshot of an agent population with the location of an agent denoted by a node on the lattice, its culture by the node’s color, and social links by the size of edges between nodes. Network communities are marked by light-blue rectangles that cover those nodes that represent the agents that are the members of these communities. In this example, some of the communities exhibit intra-community diversity and there is also inter-community diversity as communities host different cultures.

Note that we analyze community-structure while our discussion on final configurations in the previous section is framed in terms of the configuration of connected components. We believe that the configuration of communities is a better indicator of cultural and social segregation than is the configuration of components. Take Figure 4.4 as an example. All agents are part of the same connected component, which obviously is very diverse in culture. However, the indicated community structure provides a much more accurate picture. This illustration describes the population as one that is socially and culturally segregated, likely to fall into a number of components corresponding to the current communities. With regard to final configurations, the focus on community structure is a limitation only at first glance: In the absorbing states and classes, communities necessarily correspond to components and hence their analysis is interchangeable.

4.3.1 The Time-Evolution of Cultures and Networks

We first describe the evolution of cultures and networks over time. Simulations are run for \( 10^8 \) iterations. In this subsection, we focus on a lattice of width \( L = 25 \) and a few

\(^{5}\)We are aware of the limitations of this approach (Brandes et al., 2008). However, we believe that the problem is less severe when the same algorithm is used to compare the community structure of different networks.
representative levels of population densities $N/L^2$, social mobility $\beta_s$, and initial diversity $q$ but extrapolate our discussion to further levels. In fact, we have conducted a series of numerical simulations using a $50 \times 50$ lattice with the same relative population densities and obtained qualitatively equivalent results. Therefore, we consider the lattice size studied here as representative. Later when we discuss the final configurations at equilibrium, we present results for varying $L$.

Figure 4.5 displays the average number of communities at different iterations for varying levels of $N/L^2$, $\beta_s$, and $q$. For all plots in this chapter, time is presented on a log scale, averages are calculated from 30 runs, and error bars have the length of one standard deviation in each direction. Note that the number of communities is normalized by population size $N$.

We see immediately that for all parameters, there is a decrease in the number of communities over time. While the number of communities does settle to a stable level for $\beta_s = 1$, it is still decreasing for $\beta_s = 10$ and $q \geq 10$ at the end of the simulations. In general, the decrease in the number of communities is faster for smaller values of $q$; and the initial decrease is generally faster for smaller values of $N/L^2$. With regard to small levels of $\beta_s$ we can say that smaller values of $N/L^2$ and larger values of $q$ promote the survival of a substantial number of communities.
4.3. Sensitivity Analysis

Figure 4.5: The average number of communities normalized by $N$ at different iterations for $L = 25$ and selected values of $\beta_s$, $N/L^2$, and $q$. As in all following graphs, averages are calculated from 30 runs and error bars correspond to one standard deviation in both directions.

Figure 4.6 shows the average intra-community diversity in the population. Recall that in all absorbing states and classes intra-community diversity is non-existent: All agents within the same community have the same culture. However, prior to the predicted decrease of intra-community diversity we observe an initial rise and peak in diversity. This peak is less pronounced for larger values of $\beta_s$ and smaller values of $N/L^2$. In line with our previous discussion, it is only in the case of $\beta_s = 1$ that the population’s trajectory approaches the absorbing states and classes in which intra-community diversity vanishes. For some populations with $\beta_s = 10$, on the other hand, intra-community diversity is yet to disappear despite a large number of simulated iterations ($10^8$ iterations).

Figure 4.7 shows the inter-community diversity in the population. Recall that in each monocultural absorbing state, there is only a single community and hence inter-community diversity is non-existent. In each multicultural absorbing class, there are multiple communities with incompatible cultures and therefore inter-community diversity is at its maximum. Hence, we expect inter-community diversity to settle to a minimum or maximum level. And indeed, for $\beta_s = 1$ inter-community diversity generally vanishes with sufficiently small values of $q$ and reaches its maximum with sufficiently large values of $q$. We discuss the exception of intermediary values of $q$ later, which is also going to explain why some of these trajectories show an initial decrease before the final increase. Let us turn to the case of $\beta_s = 10$. With decreasing $N/L^2$ and increasing $q$, inter-community diversity is more prone to maintaining a stable level throughout the time of these simulations. Only for sufficiently large values
of $N/L^2$ and sufficiently small values of $q$ there is a visible decrease in inter-community diversity over time.

### 4.3.2 The Final Configuration of Cultures and Networks

Having described the evolution of cultures and networks over time, we turn now to their configuration at equilibrium. We are primarily interested in cultural and social space here. Therefore, we assume that a population is in equilibrium once intra-community diversity has vanished and inter-community diversity is at its minimum or maximum. Hence, we ignore any ongoing change in the physical locations of agents. In these states, network and cultures have reached a configuration corresponding to a monocultural absorbing state or a multicultural absorbing class. In the former case, no change of physical locations occurs, but in the latter case, agents can still change their locations. However, the reachable states differ only by the physical locations of agents and in all these states, communities correspond to components and cultures.

We found parameter settings for populations with small social mobilities ($\beta_s \geq 2$) for which populations had not converged even after $10^9$ iterations. Therefore, we analyze in detail populations with $\beta_s = 1$ as representatives for large social mobilities, which allow populations to converge within a reasonable number of iterations (in the order of $10^8$ to $10^9$). However, we do expect a qualitatively equivalent behavior at equilibrium independently of the value of $\beta_s$.

Figure 4.8 provides a quick overview of the results by showing the number of communities (and therefore cultures and components) at equilibrium depending on $L$, $N/L^2$, and $q$.

---

**Figure 4.6:** The average intra-community diversity at different iterations for $L = 25$ and selected values of $\beta_s$, $N/L^2$, and $q$. 

---

<table>
<thead>
<tr>
<th>$N/L^2$</th>
<th>0.04</th>
<th>0.2</th>
<th>0.8</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_s$</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time in iterations</th>
<th>10^6</th>
<th>10^7</th>
<th>10^8</th>
<th>10^9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-community diversity</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

---

76
4.3. Sensitivity Analysis

Figure 4.7: The average inter-community diversity at different iterations for $L = 25$ and selected values of $\beta_s$, $N/L^2$, and $q$.

Figure 4.8: The average number of communities (and cultures and components) at equilibrium for $\beta_s = 1$ depending on $L$, $N/L^2$, and $q$.

As observed in the previous section, the number of communities increases with $q$. However, the number of communities decreases with increasing $N/L^2$. While not directly visible from these figures, the number of communities decreases also with increasing $L$ for fixed $N/L^2$ and $q$ (established with independent two-sample t-tests based on a p-value threshold of 0.05). Thus, with a larger absolute population size $N$, which is determined by $L$ and $N/L^2$, there are in general less cultures and communities in relation to population size.

Figure 4.9 shows the number of communities, the size of the largest community, and inter-community diversity at equilibrium depending on $q$ for two particular configurations of $\beta_s$, $L$, and $N/L^2$. These graphs are representative of varying levels of $L$ and $N/L^2$. We do not present the average intra-community diversity here since it is at its minimum level for all parameter settings as predicted by our discussion about the absorbing states and classes.

77
Chapter 4. A Model of the Co-evolution of Cultures, Networks, and Geographical Locations

\[ \beta_s = 1 \quad \beta_s = 2 \]

Figure 4.9: The average number of communities, the average size of the largest community, and average inter-community diversity for independent runs at equilibrium depending on \( q \) for \( L = 25, N/L^2 = 0.4 \), and \( \beta_s \in \{1, 2\} \). The critical value \( q_c \) marks the location at which the variance in the size of the largest community is maximum. The gray area marks the range of \( q \)-values in which some simulation runs settle into a monocultural absorbing state with a single community, while other runs settle into a multicultural absorbing class with multiple communities.

In line with previous observations, the number of communities increases with increasing \( q \). Accordingly, the size of the largest community decreases with increasing \( q \). Furthermore, there is a transition from a minimum level of inter-community diversity at small values of \( q \) to a maximum level at large values of \( q \). This transition is associated with a continuous phase transition in the number of communities and in the size of the largest community. The transition occurs around a critical value \( q_c \) at which the variance in the size of the largest community is at its maximum. For \( q < q_c \), the dynamics settle generally into one of the monocultural absorbing states with a single community or one of the multicultural absorbing classes in which one culture and community dominates the population. For \( q > q_c \), the dynamics settle generally into a multicultural absorbing class with multiple communities. The number of these communities grows monotonically with increasing \( q \). Likewise, the size of the largest community decreases with increasing \( q \). Associated with the phase transition is an area at which both monocultural absorbing states and multicultural absorbing classes are reached (highlighted gray in the figure). It is this range of \( q \)-values that produces a large variance in the average inter-community diversity observed in Figure 4.7. Both populations with an inter-community diversity of 0 and populations with an inter-community diversity of 1 are observed in this range.

A comparison between the case of \( \beta_s = 1 \) and the case of \( \beta_s = 2 \) in Figure 4.9 confirms the previously mentioned expectation that the behavior of the system in equilibrium does not change qualitatively with varying \( \beta_s \). In the following, we proceed with focusing on the case of \( \beta_s = 1 \) as we suggested in the beginning of this section. We do remark, however, that the case of a larger \( \beta_s \) exhibits a lower value of \( q_c \). This is in line with the previous observation.
4.3. Sensitivity Analysis

Figure 4.10: The location of the critical value $q_c$ for $\beta_s = 1$ depending on $L$ and (a) $N/L^2$ or (b) $N$ respectively. Vertical lines mark those populations for which $N = L^2$.

that populations with larger values of $\beta_s$ are generally more resistant against the merging of communities and cultures.

Figure 4.10a shows that the critical value $q_c$ increases both with increasing $L$ and increasing $N/L^2$. In fact, Figure 4.10b shows that $q_c$ increases monotonically with increasing absolute population size $N$ if high population densities ($N/L^2 \gtrsim 0.97$) are ignored. The value of $q_c$ for high-density populations is larger than suggested by the general trend described above. We will return to this paradox later.

While not shown here, we also observe that the gradient in the size of the largest community around the phase transition is steeper for larger $N/L^2$. With larger $N/L^2$, the dominance of the largest community is maintained over a larger range of $q$-values but the subsequent decrease in the size of the largest community with increasing $q$ is more rapid.

Figure 4.11 shows the distribution of community sizes in the area of $q$-values for which both monocultural absorbing states and multicultural absorbing classes are observed for $L = 25$ and different levels of $N/L^2$. Obviously, community sizes are distributed bi-modally with the bulk located at the two ends of the scale. The bimodal character of the distribution is better developed for larger values of $N/L^2$. This is in line with the previous observation that the largest community is more dominant around this area for larger $N/L^2$ and hence does not leave room for other communities of significant size. Figure 4.11 is representative for all levels of $L$ that we studied.

We can now explain the negative peak in inter-community diversity early on in the simulations shown in Figure 4.7 for $\beta_s = 1$ and $q$-values around the phase transition. For each of these $q$-values, there are populations that continue to merge into a single community. But there are also populations that settle into multiple communities. The former populations contribute to a constant decrease in inter-community diversity. The latter populations
Chapter 4. A Model of the Co-evolution of Cultures, Networks, and Geographical Locations

Figure 4.11: The distribution of community and culture sizes for $\beta_s = 1, L = 25$, varying levels of $N/L^2$, and those values of $q$ at which both monocultural absorbing states and multicultural absorbing classes are observed.

Figure 4.12: Inter- and intra-community diversity over time for $\beta_s = 1, L = 25, N/L^2 = 0.8, q = 30$ and 5 sample runs.

Contribute initially to a decrease as well. However, ultimately the emerging communities diverge in their cultures such that inter-community diversity reaches its maximum. The negative peak of inter-community diversity occurs before these populations emerge into the final configuration of culturally incompatible communities with maximum inter-community diversity. Figure 4.12 illustrates this process. The figure shows 5 sample runs, 2 of which converge to an inter-community diversity of 0 and 3 of which converge to an inter-community diversity of 1. The negative peak in the case of the latter runs can clearly be seen.

4.4 Observations

We now discuss key conclusions drawn from the results presented in the previous section. Our key conclusions concern (A) the initial peak in intra-community diversity and the eventual emergence of culturally homogeneous communities; (B) the promotion of heterogenization by larger initial diversities and the promotion of homogenization by larger lattice sizes and population densities; and (C) the role of migration in the promotion of heterogenization.

(A) The agents’ interactions lead to the creation of social links and the formation of network communities. These communities merge with each other until either a single
community or a number of communities evolve. The emerging social connectivity within communities promotes cultural homogenization such that ultimately the process reaches one of the monocultural absorbing states with a single culture and community or one of the multicultural absorbing classes with multiple culturally homogeneous communities. Initially, however, the creation of social links enables agents with different cultures to join the same community, which entails an initial peak in intra-community diversity. It is not until the network settles into its final configuration that cultural homogenization leads to the disappearance of any diversity within communities. If multiple communities survive, their cultures are incompatible such that no further merging is possible.

Large social mobilities allow for long range interactions and hence promote the creation of social links and the formation of communities more so than small social mobilities. Therefore, the initial peak in intra-community diversity is more pronounced in the case of large social mobilities. In this case, the creation of social links clearly outpaces cultural homogenization. For smaller social mobilities, the creation of social links and the merging of communities is much slower such that cultural homogenization is able to confine the initial rise of diversity in the communities.

An analysis of the equilibrium states for a large social mobility has revealed that the number of communities, which correspond to the surviving cultures, increases with increasing initial diversity. In contrast, the number of communities decreases with increasing lattice size and population density, and therefore decreases in general with increasing absolute population size. As in Axelrod’s original model, we find a phase transition depending on initial diversity. Below a critical value, populations reach a monocultural absorbing state or one of the multicultural absorbing classes in which there is one community that dominates the population. Above the critical value, populations reach a multicultural absorbing class with the number of surviving communities growing with increasing initial diversity.

Our observations contrast with findings for a model of the co-evolution of cultures and networks also based on Axelrod’s work (Centola et al., 2007; Vazquez et al., 2007). These studies also note a critical value \( q_c \) at the transition between a phase with a single dominating culturally homogeneous component and a phase with multiple culturally homogeneous components. They observe that the critical value \( q_c \) increases with increasing population size, which parallels our finding that the critical value increases with lattice size and population density. However, they find another transition at \( q^* > q_c \) to a third phase in which a single culturally-heterogeneous network component emerges. In their model, the total number of social links is kept constant because two interacting agents rewire their link randomly after an unsuccessful interaction. For large initial diversities, the majority of interactions are unsuccessful such that constant rewiring creates a random network with a single large
component. In our model, the number of social links is not kept constant and therefore does allow the population to segregate socially even in the case of large initial diversities.

With larger absolute population sizes (determined by lattice size and population density), there is a larger probability for a given agent to have a culturally compatible partner in the population. This facilitates the merging of cultures. Therefore, there is a larger range of initial diversities at which a single community or culture emerges and hence the critical value of initial diversity is larger. The only exception to this tendency is posed by high-density populations. In this case, migration is impaired such that a single culture can be maintained for a larger range of initial diversities. This leads to a critical value of initial diversity that is disproportionately larger than the ones for other comparable population sizes where migration is more flexible. In effect, the restriction of migration is conducive of cultural and social homogeneity. These results suggest that the phase transition disappears in the limit of absolute population size with a single culture emerging for all levels of initial diversity.

The decrease in the size of the largest community with increasing initial diversity is less rapid for small population densities than for large ones. Therefore, smaller population densities show a much larger diversity in community sizes than larger population densities around the $q$-values at which both monocultural absorbing states and multicultural absorbing classes are observed. Larger population densities, in contrast, are more likely to maintain a dominant largest community, which manifests itself in a more pronounced shape of the bimodal distribution of community sizes in this area.

(C) Our observation that the number of communities increases monotonically with increasing initial diversity for all lattice sizes and population densities partly contrasts with findings by Gracia-Lázaro et al. (2009). In their model, migration destinations are chosen randomly from empty cells on the lattice. For small population densities and small levels of initial diversity, multiple culturally homogeneous components evolve. For medium levels of initial diversity, mobility is increased and random migration promotes cultural exchange, which leads to the emergence of a mono-culture and a single component. Large initial diversities lead to infinite migration and the constant erosion of cultural regions. In our model, migration is not random but directed towards social neighbors. Hence there is no arbitrary migration that could circumvent cultural borders. Instead, migration is actually contributing to social and cultural segregation because social contacts, which are typically culturally similar, tend to migrate closer to each other. This also offers an explanation for the paradox of high-density populations: These populations are restricted in their migration and therefore are not subject to the segregating effect of migration. Consequently, these populations are unexpectedly more prone to evolving a single culture compared to some populations of larger absolute size with more possibility for migration.
4.5 Discussion and Conclusions

This chapter has presented a novel agent-based model of the co-evolution of cultural, social, and physical space that draws on some of the premises and implications of the grounding model of cultural transmission. The formulation of this model based on the grounding model of cultural transmission is this chapter’s contribution to RQ1. In contrast to previous studies based on Axelrod’s model, our model demonstrates how the interaction between the three spaces can be considered through a rigorous analysis of relevant social scientific work. Our analysis of the model’s dynamics contributes to the understanding of how cultural evolution interacts with geographical and social change.

The presented model is located around the micro-macro area of the nano-to-macro axis and not a strictly faithful representation of the grounding model of cultural transmission. Hence, this model falls into the category of abstract theoretical models as discussed in Section 3.2.2. Because of its abstract nature, the construction of the model has not contributed to the refinement of the grounding model of cultural transmission, and we do not claim that the model aligns with the real-world in any strong sense. Due to the model’s simplicity, we do not see any possibility of validating it in a rigorous way, e.g. by laboratory experiments. The only way in which this model is grounded in the real-world is in its adoption of existing models and theories and in the generic predictions that it generates. These predictions are obtained from simulations with larger agent populations, which are possible just because the model’s simplicity yields a moderate computational complexity of agent interactions.

The focus of our investigations has been to use simulations to examine the configuration of cultures and social networks, and in particular to describe cultural diversity within and between different communities in the emergent network. This section starts with a summary of the results and then proceeds with a discussion on the relevance of these results to research on the effect of globalization on cultural diversity. We thereby contribute to RQ2. We close with a discussion of future work, with a focus on possible extensions of this model.

4.5.1 Summary of Results

We found that agents ultimately merge into one or more culturally homogeneous network communities. However, communities pass through an initial phase of cultural diversity when agents of different cultural backgrounds are enabled to join the same community. A larger probability of interaction over longer distances accelerates the creation of social links and hence facilitates the initial peak of cultural diversity in communities. We found that larger initial diversities promote heterogenization, i.e. the survival of multiple communities and hence cultures at equilibrium. In contrast, larger lattice sizes, population densities, and in general absolute population sizes promote homogenization, i.e. the emergence of a
single community and culture. The restriction of migration by high population densities in particular promotes homogenization, suggesting in turn that migration that is directed towards existing social contacts serves heterogenization and cultural and social segregation. It is important to note that the convergence of the system to equilibria does not reflect what has happened in the real world so far, where we observe the continuous evolution of cultures, networks, and locations.

4.5.2 The Effect of Globalization on Cultural Diversity

There appears to be disagreement on the effect of globalization on cultural diversity. Some believe that globalization will cause the homogenization of human culture (Robinson, 2007). Others emphasize, for example, that immigrant diasporas that do resist cultural assimilation by their host country can only exist because of more efficient communication with their home countries (Castles, 2002).

It appears reasonable to equate the extent of social mobility in our model with the quality of available communication and transportation technologies. Our results suggest that improved communication and transportation technologies as reflected by a larger social mobility increase the connectivity between people from different cultural backgrounds. This leads to an increased cultural diversity within network communities, which can cause a feeling of increased diversity from a local perspective when social connections are created between people from different cultural backgrounds. At the same time, however, the overall number of cultures and communities decreases steadily. Ultimately, the diversity within communities vanishes and the diversity between communities reaches its maximum, which corresponds to a situation with dispersed cultures. In effect, global homogenization is accompanied by initial local heterogenization followed by local homogenization. For research on globalization and the effect of improved technology on the exchange of cultural information this implies that a distinction between locally perceived and global, objective diversity is expedient.

In many developing countries, globalization has recently led to a rapid growth of cities, both in size and inhabitants. Our model suggests that larger population densities and absolute population sizes promote homogenization and the merging of cultures. This is a hypothesis that could be tested with data on the relationship between area sizes and the number of individuals and cultures on these areas.

On the other hand, globalization has likely changed life for emigrants as well. Given improved communication and transportation technologies, it is easier to hold contact with friends and family at home and to organize for friends and family to migrate to nearby locations. We found evidence in our results that this type of directed migration facilitates cultural heterogenization. It would be instructive to compare the extent of cultural assimilation of
migrants from different source cultures with a different tendency to migrate close to existing 

4.5.3 Future Work

The model described in this chapter extends Axelrod’s model of cultural dissemination to 
account for a few main aspects of the grounding model of cultural transmission: actual 
common ground, the grounding process and its effect on communal common ground and on 
the regulation of social relationships. In the following, we explore which other aspects could 
be represented in the model.

In this model, the agents have direct access to their actual common ground. They do 
not have a representation for what they assume to be the common ground with a particular 
interaction partner (PREM2). Take as an example three agents a, b, c. Let us assume that 
agents a and b interact and agent a changes its culture. If agent a then interacts with agent 
c and changes its culture again, agent b should not be aware of this and actually assume a 
common ground that is different from the actual one. The lack of the distinction between 
actual and perceived common ground rules out the possibility of investigating under which 
conditions actual common ground within a community might come to deviate from the 
perceived one.

This leads to the next point. Agents do not develop any awareness of their community 
membership, i.e. with respect to network or geographical communities. They do not have any 
representation of what they assume communal common ground to be (PREM3). Therefore, 
cultural transmission depends on properties of the interacting dyad only and second-order 
emergence, which is so characteristic of social systems, is not possible.

However, any such extension would largely increase the complexity of the model. Note 
how we integrated the notion of common ground and some of its ramifications into Axelrod’s 
model without any major changes. This enables results to be comparable with previous work 
and it facilitates the analysis of the model.

While we have considered the quality of social links in cultural transmission, we have 
not considered their purpose. There is no distinction between social links with regard to 
the distribution of epistemic and relational goals that are involved in the interactions over 
these links (PREM8). However, it is not clear how to generate such a distribution for 
this very abstract model, especially considering that the network is actually evolving. One 
could, however, consider a static network generated from richer network models that include 
information about these link attributes.

Additional work is necessary to complete the exploration of the parameter space. For 
example, the parameter that governs the probability for migration over particular distances is
to be investigated, which is complementary to the parameter regulating the probability of interaction over particular distances considered here.

Another possible avenue for further work is the investigation of alternative mechanisms for migration and for the change of social links. For example, one can conceive a more detailed mobility model that accounts for the costs and benefits of an agent making contact with another agent. Agents could deliberate over the costs and benefits of traveling to/interacting with other agents. In such a model, distance and social relationships could be but two of the many factors involved in deciding with whom to interact. Such aspects have already been represented in agent-based models of traveling activities (see for example Ronald et al., 2012).

We also suggest to analyze the model under conditions in which cultural space is only coupled with social space or only with physical space. We expect this to yield dynamics different from the model studied here because of the distinct rules of interaction between each pair of spaces. Such an analysis can shed light on the role of each of the spaces in our results.
Chapter 5

An Agent-based Model of Stereotype Communication

In line with our discussion in Section 3.4, this chapter addresses the question of what is a suitable model for studying the communication of stereotype-relevant information. In comparison with the previous chapter, the model we develop here assumes a more detailed representation of the internal mechanisms governing an agent’s behavior.

The concept of a stereotype has been redefined in a variety of ways since Walther Lippmann (1922) introduced his “pictures in the head” metaphor of stereotypes (Stangor and Lange, 1994). However, for the purpose of this chapter it is sufficient to understand a stereotype as a possibly commonly held generalized belief about a social group. As stereotype-relevant information we consider information that has a relationship with this belief, either because it supports or contradicts it. From the perspective of the grounding model of cultural transmission, stereotypes and stereotype-relevant information can be part of common ground and thus play a role in the grounding process. For example, information that is consistent with commonly held stereotypes is preferred to be communicated over information inconsistent with stereotypes (Lyons and Kashima, 2003). In this chapter, we construct an agent-based model of the communication of stereotype-relevant information. In doing so, we build on the grounding model of cultural transmission and on an empirical study that can be seen as one of the first empirical tests of that model.

Stereotypes are frequently employed in the judgment of others and in the prediction of their behavior if more accurate information is unavailable. Hence stereotypes are an essential cognitive tool that supports us in the navigation of our complex and uncertain social environment. However, stereotypes are also related to generally undesirable effects such as prejudicing and the derogation and discrimination of out-groups (Stephan et al., 2009).
Chapter 5. An Agent-based Model of Stereotype Communication

Therefore, stereotypes can threaten social sustainability. Social psychology—the science that studies how we make sense of and act in our social environment—has long been interested in understanding and controlling the application of stereotypes. Most attention has been paid to the intra-personal cognitive processes of stereotyping (Stangor, 2009).

However, information about members of other social groups is often not acquired first-hand through direct contact but more often second-hand through inter-personal communication. In the wider context of a social community, inter-personal communication of stereotype-relevant information manipulates the state of stereotypes within that community. In fact, this communication typically serves the maintenance of stereotypes, which ensures that they are rather stable over time (Lyons et al., 2007).

In line with our discussion in Chapter 2, this calls for a two-fold research program: First, one needs to gain an understanding of the mechanisms that govern the exchange of stereotype-relevant information during inter-personal communication. Second, given this understanding, one needs to study how on the larger scale inter-personal communication adds up to the formation and maintenance of stereotypes. Recently, research has started to focus on the first part of this agenda (Brauer et al., 2001; Klein et al., 2003; Kurz and Lyons, 2009; Lyons and Kashima, 2003; Wigboldus et al., 2005) and some discussion has been initiated on the second part (Lyons et al., 2007). We see potential for agent-based modeling to contribute to this second part: Agent-based modeling allows the direct representation and simulation of actors and their interactions with the possibility of considering a large population of actors (Gilbert, 2008). Hence, agent-based modeling is predestined for drawing on the results of the first part of this agenda to enable an analysis targeting the second part.

We introduce an agent-based model of the communication of stereotype-relevant information. The model is based on the grounding model of cultural transmission and on an empirical study by Lyons and Kashima (2003) on the communication of stereotype-consistent (SC) and stereotype-inconsistent (SI) information through a chain of communicators. As an example, consider a stereotype “All members of group G are smart”. SC information would then be, for example, “A, a member of G, is smart/clever/intelligent”. SI information would be, for example, “A, a member of G, is stupid/unintelligent”. Lyons and Kashima found that the extent to which either SC or SI information is communicated depends on the actual and perceived sharedness of relevant stereotypes among the communicators. Given that stereotypes are typically perceived to be shared with close friends or family but not with others, these results suggest that the exchange of stereotype-relevant information along strong ties in the social network differs from the exchange along weak ties (Lyons et al., 2007). This highlights the role of the social network in the diffusion process and motivates the application of agent-based models that account for the underlying social network.
Taking the results of Lyons and Kashima as well as some of the premises and implications of the grounding model of cultural transmission (see Section 2.2.3) into account, our model enables agents to create an internal representation of the information they as well as third parties associate with other individuals or social groups. Following the recent development of computational modeling in social psychology (Smith, 2009), we rely on a distributed connectionist model for the representation of memory. This architecture allows for the integrated storage and retrieval of information about individuals and groups, which plays a crucial role in stereotyping processes. When communicating stereotype-relevant information, agents take into consideration the information they perceive to share about the respective stereotype with their communication partners. This model is this chapter’s contribution to RQ1. On the nano-to-macro axis, our model covers an area that stretches from the nano level to the macro level and hence covers a wider range on this axis than the model presented in the previous chapter.

Our model provides a perspective for studying how the communication of stereotype-relevant information might play out in larger societies, also as a function of the composition and configuration of the underlying social network. The model takes into consideration that individuals obtain information about others first- as well as second-hand and make varying assumptions about the stereotypes held by their communication partners. In this process, the common ground associated with the individuals in the network can be accounted for, as well as their group membership and the group membership of the third party that is communicated about. Existing computational models of stereotyping do not represent these details as rigorously. They are either restricted to intra-personal cognitive processes (Queller and Smith, 2002; Smith and DeCoster, 1998; Van Rooy et al., 2003) or represent communication and the storage of information about others in limiting and unintuitive ways (Van Overwalle and Heylighen, 2006; Van Rooy, 2009).

Our analysis by simulation shows that the model can be calibrated to and validated against the data collected by Lyons and Kashima. Furthermore, the model can reproduce patterns of the communication of stereotype-relevant information in larger societies. This provides evidence in favor of the mechanisms of the grounding model of cultural transmission realized in our model. This is the contribution of this chapter towards RQ2.

We begin this chapter in Section 5.1 by introducing the experiments and results of Lyons and Kashima. In Section 5.2, we describe the connectionist modeling approach in general and then discuss those existing agent-based models of stereotype communication that rely on connectionist models. We follow with a description of our model in Section 5.3. Then we describe in Section 5.4 how we estimated the parameters of this model based on one set of data, and in Section 5.5 we describe how we validated the model on another data set. In Section 5.6, we present a sensitivity analysis for the estimated model. In Section 5.7, we
5.1 Experiments by Lyons and Kashima

In this section, we describe the experiments of Lyons and Kashima (2003) and briefly present their results. Their study is based on the observation that the communication of stereotype-consistent (SC) information is generally favored over the communication of stereotype-inconsistent (SI) information. Hence, the maintenance of stereotypes appears to be naturally ensured by communication. A major part of their study addresses the question of how the actual sharedness and perceived sharedness of stereotypes influence this SC bias. Actual and perceived sharedness are equated with actual and perceived common ground.

Lyons and Kashima rely on an experimental paradigm known as the serial reproduction chain to study the communication of stereotype-relevant information (Bartlett, 1932). Participants are grouped into chains of four and are individually primed with background (stereotype) information about a fictional group, the Jamayans. Subsequently, a story about Jay, a particular Jamayan, is communicated through the chains. Every participant reads and memorizes a version of the story reproduced by the previous participant and then reproduces this story for the next participant in the chain. The clue is that the original story about Jay consists to equal parts of SC and SI information with regard to the stereotype about the Jamayans previously learnt by the participants. At every position in the chain, the proportion of SC and SI content retained from the original story is measured, which indicates to which extent communication favors SC over SI information. Figure 5.1 shows this setup.

To study the influence of actual and perceived stereotype sharedness, a $2 \times 2$ factorial experimental design is constructed with the following two factors and levels:
5.1. Experiments by Lyons and Kashima

**Actual sharedness** All participants in the chain are either primed with the same information about the Jamayans (*shared condition*) or the second and fourth participant in the chain are primed with information that is contradictory to the one that the first and third participant are primed with (*unshared condition*).

**Perceived sharedness** All participants in the chain are either told (*knowledge condition*) or not (*ignorance condition*) that the next person in the chain (the audience) has learnt the same information about the Jamayans.

This yields four experimental conditions, which we call *shared-ignorance, shared-knowledge, unshared-ignorance,* and *unshared-knowledge*. Each of these is represented by 8 serial reproduction chains of 4 participants each. The following summary data is reported, summing up to 30 data points:

- Overall proportion of *SC* and *SI* content reproduced with respect to the original story.
- Proportion of *SC* and *SI* content reproduced in the *knowledge* condition.
- Proportion of *SC* and *SI* content reproduced in the *ignorance* condition.
- Proportion of *SC* and *SI* content reproduced in the *shared* condition.
- Proportion of *SC* and *SI* content reproduced in the *unshared* condition.
- Proportion of content reproduced at every position in the chains.
- Proportion of *SC* and *SI* content reproduced in the *shared* condition at every position in the chain.
- Proportion of *SC* and *SI* content reproduced in the *unshared* condition at every position in the chain.

Lyons and Kashima find that an SC bias emerges only when the stereotype is shared by all participants in the chain and participants believe that their audience is ignorant of the stereotype (shared-ignorance condition). This result is explained in terms of Grice’s *maxims of quality and quantity* (Grice, 1975). In general, participants rely on their stereotype as truthful information (maxim of quality) when assessing and interpreting perceived information. Hence, SC information is favored over SI information during perception. When the audience is deemed unaware of the stereotype (ignorance condition), SC information is also assumed to be informative (maxim of quantity), which contributes further to the SC bias. However, when the audience is assumed to have knowledge of the stereotype (knowledge condition), SI information is more informative and an SC bias fails to emerge.
5.2 Connectionist Models of Stereotype Communication

This section gives a brief introduction into connectionist modeling and existing connectionist agent-based models of the communication of stereotype-relevant information. Recall that we mentioned connectionist modeling briefly in Chapter 3 as one of the primary types of models employed by cognitive scientists. Connectionist modeling is a method of computation that borrows from neuroscience: The basic elements of a connectionist model are a set of processing units connected by weighted links reminiscent of neurons and synapses respectively. The state of every unit is described by its activation level. The activation level is determined by the net input received from other units via incoming links, possible external input, and an activation function applied to this input. Computation proceeds by propagating this activation through the network.

The purpose of such a network is generally the storage and reproduction of associations between external inputs and associated target activation patterns. To accomplish this, the weights in the network are adjusted incrementally according to some learning algorithm such that the network configuration becomes a representation of the patterns to be learnt. Provided with an external input, the network is then able to reproduce the associated target pattern. The weights of the network are often equated with long-term memory and the current activation levels within the network with short-term memory. Thus connectionism lends itself to the modeling of human memory and cognition, essentially blurring the distinction between storage and processing. This approach was pioneered by Rumelhart et al. (1986), who coined the term “Parallel Distributed Processing” (PDP), emphasizing that computation in a connectionist model is distributed between units and proceeds in parallel.

A distinction is made between so called localist representations, in which each unit corresponds to a particular concept, and distributed representations, in which no particular meaning is given to individual units but to patterns of activation over these units. A drawback of localist representations is that they need to rely on an external process to create a new unit in the network when a new concept is to be learnt. In a distributed representation, a new concept is simply represented by a new activation pattern. Another advantage of distributed representations is that the degree of similarity between two concepts is encoded in the similarity of the activation patterns they are associated with. This allows the network to learn prototype representations from exemplars with similar activation patterns and to process new exemplars according to their similarity with the previously extracted prototypes (Smith, 1996). Because different representations are superimposed within network weights, recall of particular patterns is inherently context-sensitive in that reproduction is influenced by any other information available to the network at that time (Smith, 2009). All these characteristics make distributed connectionist models well suited for the modeling of stereotyping processes,
which are concerned with the context-sensitive storage and recall of information about exemplars and the extraction and exploitation of prototypes.

Types of connectionist networks can be distinguished by their arrangement of units and links. For the modeling of stereotyping, so called fully recurrent networks have proven useful (Queller and Smith, 2002; Smith and DeCoster, 1998; Van Rooy et al., 2003). Refer to Figure 5.2 for a visualization of this network type. In these networks, all units provide input to all other units. Processing essentially happens in two phases: First, a pattern of external input is applied to the units of the network and activation is allowed to propagate through the network for a number of iterations. The emerging pattern the network settles into can be understood as an interpretation of the external input. Second, the weights of the network are changed according to the difference between the internal input and the target at each unit. Accordingly, the network becomes better able to reproduce this target pattern provided the external input. Models of stereotype communication that are based on this particular connectionist architecture were introduced by Van Overwalle and Heylighen (2006) as well as Van Rooy (2009), whereby the latter work is based on the former. Because Van Overwalle and Heylighen show how their model can reproduce the experiments of Lyons and Kashima, we focus our discussion on their work here.

In the experiment of Van Overwalle and Heylighen for reproducing the results of Lyons and Kashima, every agent consists of a fully recurrent network with a localist representation. The 5 units of each network represent the concepts “Jamayans”, “honest” and “smart” (SC), “dishonest” and “stupid” (SI). Each agent is hence able to learn and reproduce the association between the Jayamans and SC and SI information.

Communication between different agents is represented by the flow of activation along connections that link units of the agents that represent the same concepts. Apart from

---

1 In so called settling or auto-associative networks, the external input corresponds to the target.
the propagation of activation, these connections play also another role: Connections are annotated with so-called trust weights in each direction reflecting how much the receiving agent trusts in the sending agent with regard to the particular concept and vice versa.

The maxims of quality and quantity are implemented based on these trust weights. A receiving agent quantifies incoming information according to the receiving trust weight associated with that connection (maxim of quality). Hence, information assumed not to be trustworthy is given less importance. A sending agent attenuates the transmission of information along connections for which the sending trust weight is high and boosts transmission along connections for which the sending trust weight is low (maxim of quantity). Hence, transmission favors information for which the receiving agent is assumed to have not much knowledge about. Trust weights are adjusted over time according to the similarity between the transmitted activation levels and the internal activation generated by the agents.

The model of Van Overwalle and Heylighen is able to reproduce the results of Lyons and Kashima to a good degree. In an initial learning phase, agents are primed with the association between the Jamayans and the SC information. In the second phase, agents communicate an association between the Jamayans and SC as well as SI information through the chain, reflecting the content of the story about Jay. Key components are the implementations of the maxims of quality and quantity, which have been identified as essential by Lyons and Kashima as well. However, we see room for improvement, both with regard to the general architecture and to the way in which the experiments of Lyons and Kashima are implemented:

- The simulations of Van Overwalle and Heylighen do not properly represent the $2 \times 2$ experimental design of Lyons and Kashima. Their shared-ignorance and unshared-ignorance simulations are inappropriately compared to the data of the shared and unshared conditions of Lyons and Kashima, which actually also include the shared-knowledge and unshared-knowledge conditions respectively. Their shared-ignorance and shared-knowledge simulations are then inappropriately compared to the data of the knowledge and ignorance conditions, which actually also include the unshared-ignorance and unshared-knowledge conditions respectively.

- There is no distinction made between the Jamayans and Jay, and the story about Jay has essentially the meaning of a story about the Jamayans. Yet an important capability of human memory and cognition is the integration of information about individuals and prototypes or stereotypes.

- The representation of what an agent believes another agent to believe in terms of trust weights is neat to some extent but unintuitive. A more direct representation would consider this information to be part of the agent’s memory and hence stored in the network as one of the learnt patterns.
5.3 Model

This section describes an agent-based model that enables the simulation of the communication of stereotype-relevant information. Before going into the details of the model, we give an overview of our approach.

5.3.1 Overview

The previous discussion provides the requirements for the model: Agents need to be able to learn who associates which information with whom. In the experiments of Lyons and Kashima, every participant stores and retrieves what they themselves as well as the next person in the chain (audience) believe about the Jamayans and Jay. Hence these associations consist of three elements: a subject, a target, and the content that the subject associates with the target. Subject and target can correspond to any arbitrary individual or social group, e.g. the self or the audience in case of the subject, and the Jamayans or Jay in the case of the target. Both for the subject and the target, the ability to represent similarity between different individuals or groups is essential. For example, the agent can perceive itself as a member of a particular social group. Likewise, Jay is not an isolated individual but perceived as a member of the Jamayans. The content is what is communicated, subject to adjustment according to Grice’s maxims of quality and quantity. Here, we consider content to consist of the extent to which it carries SC and SI information.

These assumptions relate to the premises and implications of the grounding model of cultural transmission identified in Section 2.2.3 as follows: Two agents’ actual personal common ground is represented by the similarity of content that they attribute to particular targets ($\text{PREM1, PREM3}$). The actual communal common ground of a social group is represented by the similarity of content that agents in that group attribute to particular targets ($\text{PREM3}$). A kind of perceived personal common ground between two agents is represented by the content that each agent attributes to a particular target and by the target and content information that each agent associates the other agent with ($\text{PREM2}$). This is enabled by agents considering the subject in their associations, allowing them to represent which target-content associations are assumed to be held by different subjects. In that, it is perceived common ground that is represented by agents and that they act on, not actual common ground.

An important aspect is that the storage and recall of a target-content association is mediated by the representation of a particular subject. This subject can represent the agent itself or also a social group. In the former case, the information represents one that the agent holds itself. In the latter case, the information represents one that the agent assumes the social group to hold. Group membership is represented by the overlap between the
representation of an agent and the representation of the group. Therefore, storage and recall of target-content associations for a subject are always intertwined with representations of groups this subject is a member of or is related to. This has two effects. First, common ground is never only determined by the information that two agents share because they themselves hold this information (personal common ground) but also by the information that they share because of their group memberships (communal common ground, PREM3). Second, the target-content information associated with social groups represents a kind of perceived communal common ground (PREM2, PREM3). Together, actual and perceived communal common ground form part of the culture of a community (PREM4).

The grounding process is represented by the exchange of content about particular targets (PREM5). Grounding leads to the acceptance of particular content by an agent and hence contributes to the change or maintenance of (perceived) communal common ground and therefore culture (IMPL3). In the experiments of Lyons and Kashima, the interaction that participants are involved in has the explicit purpose of communicating some information. Lyons and Kashima conclude from their experiments that epistemic goals in these interactions are managed by Grice’s maxims of quality and quantity. Representing these maxims of communication therefore is a first step towards considering epistemic goals in the grounding process (PREM8). To obey the maxims of quality and quantity, agents need to rely on their perceived common ground (IMPL1) to assess the compatibility of exchanged information with common ground (IMPL2).

Because of the integration of different subject patterns for individuals and groups, grounding is always also affected by perceived communal common ground, i.e. culture at the macro level (IMPL5). Because different pairs of agents have a different common ground, their arrangement in particular network structures at the macro level affects how information is transmitted in the population at the micro level (IMPL6).

5.3.2 Memory Storage and Reproduction

Following previous models of stereotyping processes, we rely on a fully recurrent network as the representation of memory. However, our model is distinguished by two features: We enable the network to encode the triadic associations of subject, target, and content described above; and we adopt a network model that allows for hidden units. Hidden units are not part of the patterns to be learnt and reproduced but allow the network to create an internal representation of the associations to be learnt. This improves the expressiveness of the network. Figure 5.3 illustrates this architecture. The network consists of four pools of units for subject, target, content, and hidden units. To enable the simultaneous integration of exemplars and prototypes, we assume that subject and target concepts are described by a distributed representation. SC and SI content units assume a localist representation. The
5.3. Model

Figure 5.3: An illustration of the network architecture of the model. The units of the network are grouped into four pools: subject, target, content, and hidden units. The network is fully recurrent, i.e. every unit provides input to every other unit. While the units in the subject and target pools adhere to a distributed representation, the units in the content pool take a localist representation.

extent to which these two units are activated when a particular subject-target pattern is provided represents how much the subject associates the target with SC and SI information. However, SC and SI are only labels and their meaning could be replaced by anything else. Activation values of all units are in the range [0, 1].

5.3.3 Network Processing

We base network processing—propagation of activation and learning of network weights—on the recurrent back-propagation network and learning algorithm described by McClelland (2011). This architecture has been used, for example, in modeling the processing of semantic memory (Rogers et al., 2004). Refer to Williams and Zipser (1995) for a comprehensive review on recurrent back-propagation. In contrast to networks and learning algorithms used in previous connectionist models of stereotyping processes, this architecture enables learning for networks with hidden units. We first introduce the necessary theoretical background and then describe the particular algorithm we used here.

Theory

Let \( \mathcal{N} = \mathcal{I} \cup \mathcal{H} \cup \{\text{bias}\} \) be the set of units in the network consisting of input-output units \( \mathcal{I} \), hidden units \( \mathcal{H} \), and the bias unit. In later sections we assume that \( \mathcal{I} \) consists of individual
sets for subject, target, and content units: $\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_T \cup \mathcal{I}_C$ with $\mathcal{I}_C = \{0, 1\}$. However, this is not relevant to the theoretical background in the rest of this section. Let $w_{ij}$ the weight of the link from unit $j$ to unit $i$. We consider discrete-time recurrent networks here with time progressing in ticks, which we denote by $t$, sometimes with subscript, e.g. $t_0$. The activation $a_i(t)$ of a unit $i$ at tick $t$ is a function of its net input $s_i(t)$ at that tick:

$$a_i(t) = f_i(s_i(t)).$$

The function $f_i$ is the activation function of unit $i$ and assumed to be differentiable, due to a reason becoming obvious later. Following McClelland (2011), the net input $s_i(t)$ is either a function of the input received from other units in the network or based on an external input $x_i(t)$ (what McClelland calls a hard-clamped input). Let $X(t) \subseteq \mathcal{I}$ be the subset of input-output units that have an external input at tick $t$. As McClelland, we assume that net inputs are time-averaged between the current net input and the one at the last time weighted by a factor $\gamma$, unless an input is provided:

$$s_i(t) = \begin{cases} x_i(t) & \text{if } i \in X(t) \\ \gamma \sum_{j \in \mathcal{N}} w_{ij} a_j(t-1) + (1-\gamma) s_i(t-1) & \text{otherwise.} \end{cases}$$

An exception is that $a_{\text{bias}}(t) = 1$ for all $t$.

Activation is propagated through the network according to Equations 5.1 and 5.2 over an interval $[t_0, t_1]$. Let $n = t_1 - (t_0 + 1)$ be the number of ticks after tick $t_0$. An external input $x_i(t)$ can be applied to any unit $i \in \mathcal{I}$ at any tick $t$. Similarly, a target $g_i(t)$ can be supplied to any unit $i \in \mathcal{I}$ at any tick $t$. Let $T(t) \subseteq \mathcal{I}$ be the subset of input-output units that have a target at tick $t$. Then the error at unit $i$ at tick $t$ is:

$$e_i(t) = \begin{cases} g_i(t) - a_i(t) & \text{if } i \in T(t) \text{ and } i \notin X(t) \\ 0 & \text{otherwise.} \end{cases}$$

The goal of learning is to minimize the overall error at all units and ticks by adjusting the $w_{ij}$ accordingly. A learning objective can be to maximize the negative overall sum squared error within the interval $(t_0, t_1]$ of $n$ ticks:

$$E(t_0, t_1) = -\frac{1}{2} \sum_{t=t_0+1}^{t_1} \sum_{i \in \mathcal{I}} [e_i(t)]^2$$
The partial derivatives of $E(t_0, t_1)$ with regard to each weight $w_{ij}$ are
\[ \frac{\partial E(t_0, t_1)}{\partial w_{ij}}. \]
Gradient descent learning can be used to change each weight in proportion to the respective derivative:
\[ \Delta w_{ij} = \eta \frac{\partial E(t_0, t_1)}{\partial w_{ij}}, \quad (5.5) \]
where $\eta$ is a learning rate.

The following theorem can be proven (see below):

**Theorem 5.1.**
\[ \frac{\partial E(t_0, t_1)}{\partial w_{ij}} = \sum_{t=t_0+1}^{t_1} \delta_i(t) \rho_{ij}(t) \quad (5.6) \]
with
\[ e_i(t) = \begin{cases} e_i(t) & \text{if } t = t_1 \\ e_i(t) + \sum_{j \in N \land j \notin X(t+1)} w_{ji} \delta_j(t+1) & \text{if } t < t_1 \end{cases} \quad (5.7) \]
\[ \delta_i(t) = \begin{cases} \gamma f'_i(s_i(t)) e_i(t) & \text{if } t = t_1 \\ \gamma f'_i(s_i(t)) e_i(t) + (1 - \gamma) \delta_i(t+1) & \text{if } t < t_1 \end{cases} \quad (5.8) \]
\[ \rho_{ij}(t) = \begin{cases} 0 & \text{if } i \in X(t) \\ a_j(t-1) & \text{else} \end{cases} \quad (5.9) \]
for $i, j \in N$, and appropriate $t$.

Theorem 5.1 provides a mechanism for calculating the gradient and determining weight updates. This mechanism actually corresponds to back-propagation in feed-forward networks. Essentially, the recurrent network is translated into an equivalent feed-forward one. Figure 5.4 illustrates this idea. All units of the network are copied into $n+1$ layers and a link from unit $i$ to unit $j$ becomes a link from unit $i$ at layer $z$ to unit $j$ at layer $z+1$. First, activations are propagated through the network from tick $t_0$ to tick $t_1$ and the activation and error at each unit at each tick is stored. This process will be called the *forward pass* in the following. Then, starting with Equation 5.7 error terms $\delta_i(t)$ are “back-propagated through time” for all units from tick $t_1$ down to tick $t_0 + 1$ using Equations 5.7 and 5.8. This is called the *backward pass*. Once the $\delta_i(t)$ are calculated, $\frac{\partial E(t_0, t_1)}{\partial w_{ij}}$ can be determined according to Equation 5.6 and weight updates can be performed according to Equation 5.5. It follows from Equation 5.8 that the $f_i$ need to be differentiable.

Williams and Zipser (1995) have proven the backward pass calculations for back-propagation through time. We now adapt their proof to show that Theorem 5.1 holds. In contrast to Williams and Zipser we also take hard clamped inputs and time-averaged net
inputs into account. The main assumption is that to determine \( \frac{\partial E(t_0, t_1)}{\partial w_{ij}} \) one simply sums up all the partial derivatives for that weight with regard to each of the \( n + 1 \) layers of the unrolled network.

**Proof of Theorem 5.1.** Let us introduce variables \( a_i^*(t) \) with \( a_i^*(t) = a_i(t) \) for all \( i \) and \( t \) such that \( E(t_0, t_1) \) is expressed in terms of these variables. Thereby we can separate the direct and indirect influence of changes in the \( a_i(t) \) on the change in \( E(t_0, t_1) \). Indirect influences obtain because later values of \( a_i(t) \) depend on earlier ones. We also consider by \( w_{ij}(t) = w_{ij} \) the weight of the link from unit \( j \) to unit \( i \) at time \( t \). We introduce a few short-hand notations in addition to the ones above:

\[
\varepsilon_i(t) = \frac{\partial E(t_0, t_1)}{\partial a_i(t)} \quad (5.10)
\]

\[
\delta_i(t) = \gamma \frac{\partial E(t_0, t_1)}{\partial s_i(t)} \quad (5.11)
\]

\[
e_i(t) = \frac{\partial E(t_0, t_1)}{\partial a_i^*(t)} = \begin{cases} g_i(t) - a_i(t) & \text{if } i \in T(t) \text{ and } i \notin X(t) \\ 0 & \text{otherwise} \end{cases} \quad (5.12)
\]

\[
v_{ij}(t) = \begin{cases} 0 & \text{if } i \in X(t) \\ w_{ij} & \text{else} \end{cases} \quad (5.13)
\]

Considering that \( E(t_0, t_1) \) is influenced by \( a_i(t) \) only through \( a_i^*(t) \) and the \( s_j(t + 1) \):

\[
\frac{\partial E(t_0, t_1)}{\partial a_i(t)} = \frac{\partial E(t_0, t_1)}{\partial a_i^*(t)} \frac{\partial a_i^*(t)}{\partial a_i(t)} + \sum_{j \in N} \frac{\partial s_j(t + 1)}{\partial a_i(t)} \frac{\partial E(t_0, t_1)}{\partial s_j(t + 1)} \quad (5.14)
\]

\[
= e_i(t) + \sum_{j \in N} v_{ji}(t + 1) \gamma \frac{\partial E(t_0, t_1)}{\partial s_j(t + 1)} \quad (5.15)
\]

\[
= e_i(t) + \sum_{j \in N \setminus j \notin X(t + 1)} w_{ji} \delta_j(t + 1) \quad (5.16)
\]
From this follows that:

\[ \varepsilon_i(t) = \begin{cases} 
\varepsilon_i(t) & \text{if } t = t_1 \\
\varepsilon_i(t) + \sum_{j \in N \setminus j \notin X(t+1)} w_{ij} \delta_j(t+1) & \text{if } t < t_1 
\end{cases} \tag{5.17} \]

Note that errors are back-propagated only from units without external input at some time \( t \).

Since \( E(t_0, t_1) \) depends on \( s_i(t) \) only through \( a_i(t) \) and \( s_i(t+1) \):

\[ \gamma \frac{\partial E(t_0, t_1)}{\partial s_i(t)} = \gamma \left( \frac{\partial a_i(t)}{\partial s_i(t)} \frac{\partial E(t_0, t_1)}{\partial a_i(t)} + \frac{\partial s_i(t+1)}{\partial s_i(t)} \frac{\partial E(t_0, t_1)}{\partial s_i(t+1)} \right) \tag{5.18} \]

\[ = \gamma f'_i(s_i(t)) \varepsilon_i(t) + (1 - \gamma) \delta_i(t+1) \tag{5.19} \]

From this follows:

\[ \delta_i(t) = \begin{cases} 
\gamma f'_i(s_i(t)) \varepsilon_i(t) & \text{if } t = t_1 \\
\gamma f'_i(s_i(t)) \varepsilon_i(t) + (1 - \gamma) \delta_i(t+1) & \text{if } t < t_1 
\end{cases} \tag{5.20} \]

We can then write:

\[ \frac{\partial E(t_0, t_1)}{\partial w_{ij}} = \sum_{t = t_0 + 1}^{n} \frac{\partial E(t_0, t_1)}{\partial w_{ij}(t)} \frac{d w_{ij}(t)}{d w_{ij}} = \sum_{t = t_0 + 1}^{n} \frac{\partial E(t_0, t_1)}{\partial w_{ij}(t)} \tag{5.21} \]

Because \( E(t_0, t_1) \) depends on \( w_{ij}(t) \) only through \( s_i(t) \), we can write:

\[ \sum_{t = t_0 + 1}^{n} \frac{\partial E(t_0, t_1)}{\partial w_{ij}(t)} = \sum_{t = t_0 + 1}^{n} \frac{\partial E(t_0, t_1)}{\partial s_i(t)} \frac{\partial s_i(t)}{\partial w_{ij}(t)} \tag{5.22} \]

\[ = \sum_{t = t_0 + 1}^{n} \frac{\partial E(t_0, t_1)}{\partial s_i(t)} \gamma p_{ij}(t) \tag{5.23} \]

\[ = \sum_{t = t_0 + 1}^{n} \delta_i(t) p_{ij}(t) \tag{5.24} \]

\[ \square \]

Note that \( \delta_i(t) p_{ij}(t) = 0 \) if \( i \in X(t) \). Hence if an input is provided to a unit at some time \( t \), no contribution to the overall error derivative is made. Together with the previous result from Equation 5.17 that no error is back-propagated from units that have an input, this implies that units do not have any impact on \( \frac{\partial E(t_0, t_1)}{\partial w_{ij}} \) for the ticks that they have inputs.
Algorithm
Based on the discussion above, we implement the back-propagation through time algorithm described by McClelland (2011). Forward pass, backward pass, and weight updates are undertaken for every input-target pattern to be learnt, whereby a pattern consists of sets \( X(t) \subseteq \mathcal{I} \) for all \( t \in [t_0, t_1] \), sets \( T(t) \subseteq \mathcal{I} \) for all \( t \in (t_0, t_1] \), inputs \( x_i(t) \) for all \( t \in [t_0, t_1] \) and targets \( g_i(t) \) for all \( t \in (t_0, t_1] \) and all \( i \in \mathcal{I} \). Typically, an input is provided during the first ticks and a target during the last ticks. The \( n \) ticks with \( t \in (t_0, t_1] \) are partitioned into \( m \) intervals, each interval consisting of \( l = n/m \) ticks. For convenience, we will later refer to intervals and not ticks when describing at which tick a particular input or target pattern is supplied to the network. The forward pass alone is applied to recall information from an incomplete pattern.

Forward Pass The forward pass, which follows Equations 5.1 to 5.3, is shown in Algorithm 5.1. In each of the \( n + 1 \) ticks in the interval \([t_0, t_1] \), net inputs and activation values for all input-output and all hidden units are calculated (the activation value of the bias unit is always 1, see line 3). We distinguish the case of input-output units that do not have an input in the current tick (from line 5) and those that do (from line 12). If the unit does not have an input, its net input \( s_i(t_0) \) at tick \( t_0 \) is simply the weighted input received from the bias node (line 7). For every other tick \( t \in (t_0, t_1] \), a unit’s net input is time-averaged between the last net input \( s_i(t - 1) \) and the current net input \( s_i(t) \) (lines 9 and 10). The unit’s activation is determined by the activation function (here the logistic function, line 11). Using the logistic function as the activation function restricts all activation values to the interval \([0, 1]\). If the unit has an input provided at the moment, its activation value is set to that input (line 13) and the net input is calculated by the inverse of the activation function (line 14). This represents a hard-clamped input as mentioned earlier. Afterwards, error terms are calculated (line 16) for all units that do have a target but no input (line 15).

Backward Pass and Weight Updates The backward pass is shown in Algorithm 5.2. Error terms \( \delta_i(t) \) are back-propagated through the network from tick \( t_1 \) down to tick \( t_0 + 1 \) according to Equations 5.7 and 5.8. Again we distinguish between units that do not have an input at the current tick (from line 5) and units that do (from line 11). As the \( s_i(t) \) in the forward pass, the \( \delta_i(t) \) are time-averaged between between tick \( t \) and tick \( t + 1 \) in the general case (lines 9 and 10). After the calculation of delta terms, weight updates can be performed according to Equations 5.5 and 5.6 (line 15).

---

2We assume \( X(t_0) = X(t_0 + 1) \) and \( x_i(t_0) = x_i(t_0 + 1) \) for all \( i \in \mathcal{I} \), i.e. an input provided at tick \( t_0 + 1 \) is also considered at tick \( t_0 \).
5.3. Model

Algorithm 5.1: ForwardPass(N, w, [t₀, t₁], γ, X, x, T, g)

Data:
- Units in the network: N = I ∪ H ∪ {bias} with I ∩ H = ∅ and bias ∉ (I ∪ H).
- Weight matrix w of size |N| × |N|.
- Tick interval [t₀, t₁] with t₀, t₁ ∈ N, t₀ < t₁, and t₁ − t₀ = n + 1 ticks.
- Proportion γ with which net inputs are determined by the current net input compared to the net input in the last tick.
- Sets X(t) ⊆ I for all t ∈ [t₀, t₁].
- Input values xᵢ(t) for all i ∈ I and t ∈ [t₀, t₁].
- Sets T(t) ⊆ I for all t ∈ (t₀, t₁).
- Target values gᵢ(t) for all i ∈ I and t ∈ (t₀, t₁).
- Logistic activation functions fᵢ(x) = 1/(1+e⁻ˣ) for all i ∈ (I ∪ H).

Result:
- Activation values aᵢ(t) for all i ∈ (I ∪ H) and t ∈ [t₀, t₁].
- Error values eᵢ(t) for all i ∈ I and t ∈ (t₀, t₁).

begin
for t = t₀ to t₁ do
    a_bias(t) ← 1
    for i ∈ (I ∪ H) do
        if i ∉ X(t) then
            if t = t₀ then
                sᵢ(t) ← wᵢ bias a_bias(t) = wᵢ bias
            else
                sᵢ(t) ← ∑ᵢ∈N wᵢj aⱼ(t − 1)
                sᵢ(t) ← γsᵢ(t) + (1 − γ)sᵢ(t − 1)
                aᵢ(t) ← fᵢ(sᵢ(t))
            else
                aᵢ(t) ← xᵢ(t)
                sᵢ(t) ← fᵢ⁻¹(aᵢ(t))
        if t > t₀ and i ∈ (T(t) \ X(t)) then
            eᵢ(t) ← gᵢ(t) − aᵢ(t)
    return a, e
end

5.3.4 Communication

Having described how agents are able to store and retrieve associations between subject, target, and content (SC and SI information), we now establish how they actually communicate. Communication knows two sides: the one of the sending agent and the one of the receiving agent. Within the context of this chapter, we assume that it is (SC, SI) tuples with SC, SI ∈ [0, 1] that are communicated between agents. The value of such a tuple describes to which extent the communication contains SC and SI information. This maps onto the data measured by Lyons and Kashima. Next, we need to describe how an agent determines the (SC, SI) tuple to communicate to another agent, and how a receiving agent integrates such a tuple into its memory.
Algorithm 5.2: BackwardPass(N, w, [t₀, t₁], γ, η, X, a, e)

Data:
- Units in the network: \( N = I \cup H \cup \{bias\} \) with \( I \cap H = \emptyset \) and bias \( \notin (I \cup H) \).
- Weight matrix \( w \) of size \( |N| \times |N| \).
- Tick interval \([t₀, t₁]\) with \( t₀, t₁ \in \mathbb{N} \), \( t₀ < t₁ \), and \( t₁ - t₀ = n + 1 \) ticks.
- Proportion \( \gamma \) with which net inputs are determined by the current net input compared to the net input in the last tick.
- Learning rate \( \eta \).
- Sets \( X(t) \subseteq I \) for all \( t \in [t₀, t₁] \).
- Activation values \( a_i(t) \) for all \( i \in (I \cup H) \) and \( t \in [t₀, t₁] \).
- Error values \( e_i(t) \) for all \( i \in I \) and \( t \in [t₀, t₁] \).
- Logistic activation functions \( f_i(x) = \frac{1}{1 + e^{-x}} \) for all \( i \in (I \cup H) \).

Result: Updated weight matrix \( w \).

begin
  for \( v = 0 \) to \( n \) do
    \( t \leftarrow t₁ - v \)
    for \( i \in N \) do
      if \( i \notin X(t) \) then
        if \( t = t₁ \) then
          \( \delta_i(t) \leftarrow \gamma f_i′(s_i(t))e_i(t) = \gamma a_i(t)[1 - a_i(t)]e_i(t) \).
        else
          \( \delta_i(t) \leftarrow f_i′(s_i(t))e_i(t) = a_i(t)[1 - a_i(t)][e_i(t) + \sum_{j \in N} w_{ij}\delta_j(t + 1)] \).
        end
        \( \delta_i(t) \leftarrow \gamma \delta_i(t) + (1 - \gamma)\delta_i(t + 1) \).
      end
      else
        \( \delta_i(t) \leftarrow 0 \).
      end
      for \( i \in N \setminus \{i\} \) do
        \( w_{ij} \leftarrow w_{ij} + \eta \sum_{\ell=t₀}^{t₁} \delta_i(t)a_j(t - 1) \).
      end
  end
return \( w \).

Sending

A sending agent recalls how much it associates itself as the subject with the target (e.g. the Jamayans or Jay) and the content (SC and SI). Then the agent performs the same procedure with the audience as the subject. The recalled content is then adjusted by the maxim of A sending agent recalls how much it associates itself as the subject with the target (e.g. the Jamayans or Jay) and the content (SC and SI). Then the agent performs the same procedure with the audience as the subject. The recalled content is then adjusted by the maxim of
Algorithm 5.3: \( \text{Send}(N, w, l, m, u, x^A, x^T, \beta_{\text{quantity}}) \)

**Data:**
- Units in the network: \( N = I \cup H \cup \{ \text{bias} \} \) with \( I \cap H = \emptyset, \text{bias} \notin (I \cup H) \),
  \( I = I_S \cup I_T \cup I_C \), \( I_C = \{ 0, 1 \} \), \( I_S \cap I_T = \emptyset, 0 \notin (I_S \cup I_T) \), \( 1 \notin (I_S \cup I_T) \).
- Weight matrix \( w \) of size \( |N| \times |N| \).
- Number of ticks per interval \( l \).
- Number of intervals \( m \).
- Number of intervals \( u \) during which input is provided, \( u \leq m \).
- Input value vectors \( x^A \in [0, 1]^{\|I_S\|}, x^A \in [0, 1]^{\|I_T\|}, x^A \in [0, 1]^{\|I_T\|} \) representing the agent itself, the audience, and a particular target respectively.
- The strength of the maxim of quantity \( \beta_{\text{quantity}} \in [0, 1] \subset \mathbb{R} \).

**Result:** \( SC, SI \in [0, 1] \subset \mathbb{R} \)

```
begin
\gamma \leftarrow 1/l 
\gamma \leftarrow 0 
\gamma \leftarrow \gamma + 1 
\gamma \leftarrow \gamma \text{ for all } t \in (t_0, t_1]
\gamma \leftarrow 0 \text{ for all } i \in I, t \in (t_0, t_1]
\gamma \leftarrow \gamma \text{ for all } i \in I, t \in [t_0, lu]
\gamma \leftarrow \gamma \text{ for all } i \in I, t \in \{lu, t_1\}
\gamma \leftarrow \gamma \text{ for all } i \in I, t \in [t_0, lu]
\gamma \leftarrow \gamma \text{ for all } i \in I, t \in [t_0, t_1]
\beta_{\text{quantity}} \leftarrow \beta_{\text{quantity}} \text{ for all } i \in I, t \in (lu, t_1]
a, e \leftarrow \text{ForwardPass}(N, w, [t_0, t_1], \gamma, X, x, T, g)
SC_{\text{self}}, SI_{\text{self}} \leftarrow a_0(t_1), a_1(t_1)
\gamma \leftarrow \gamma \text{ for all } i \in I, t \in [t_0, lu]
\beta_{\text{quantity}} \leftarrow \beta_{\text{quantity}} \text{ for all } i \in I, t \in [t_0, t_1]
a, e \leftarrow \text{ForwardPass}(N, w, [t_0, t_1], \gamma, X, x, T, g)
SC_{\text{audience}}, SI_{\text{audience}} \leftarrow a_0(t_1), a_1(t_1)
SC \leftarrow SC_{\text{self}}, SI_{\text{self}} \ast (1 - SC_{\text{audience}}) \ast SC_{\text{self}} \ast \beta_{\text{quantity}}
SI \leftarrow SI_{\text{self}} \ast (1 - SC_{\text{audience}}) \ast SC_{\text{self}} \ast \beta_{\text{quantity}}
return SC, SI
```

itself associates the target with SC and SI information (line 14). We denote this tuple by \((SC_{\text{self}}, SI_{\text{self}})\).

We highlighted previously that people adjust the information they communicate to their audience in line with the maxim of quantity. Therefore, the agent also reproduces the SC and SI information that it assumes the audience to associate with the target. The procedure is essentially the one above, yet the subject units are activated with a pattern that represents the audience (lines 15 to 17). We denote this tuple by \((SC_{\text{audience}}, SI_{\text{audience}})\).

The \((SC, SI)\) tuple the agent communicates is then determined by an interpretation of the maxim of quantity (lines 18 and 19). Both equations in lines 18 and 19 are symmetrical. The communication of each type of information (SC and SI) is reduced inversely proportionally to what the audience is assumed to know about the other type of information and proportionally
We discussed previously that people turn to their stereotypes or in fact their prior knowledge when assessing information that they receive, mediated by the maxim of quality. So an

Chapter 5. An Agent-based Model of Stereotype Communication

Algorithm 5.4: Learn($\mathcal{N}, w, l, m, u, v, \eta, x^S, x^T, x^C_0, x^C_1$)

Data:
- Units in the network: $\mathcal{N} = \mathcal{I} \cup \mathcal{H} \cup \{\text{bias}\}$ with $\mathcal{I} \cap \mathcal{H} = \emptyset$, $\text{bias} \notin (\mathcal{I} \cup \mathcal{H})$,
- $\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_T \cup \mathcal{I}_C$, $\mathcal{I}_C = \{0, 1\}$, $\mathcal{I}_S \cap \mathcal{I}_T = \emptyset$, $0 \notin (\mathcal{I}_S \cup \mathcal{I}_T)$, $1 \notin (\mathcal{I}_S \cup \mathcal{I}_T)$.
- Weight matrix $w$ of size $|\mathcal{N}| \times |\mathcal{N}|$.
- Number of ticks per interval $l$.
- Number of intervals $m$.
- Number of intervals $u$ during which input is provided, $u \leq m$.
- Number of intervals $v$ during which a target is provided, $v \leq m$.
- Learning rate $\eta$.
- Input value vectors $x^S \in [0, 1]|\mathcal{S}|$, $x^T \in [0, 1]|\mathcal{T}|$ representing a subject and target respectively.
- $(x^C_0, x^C_1)$ input with $x^C_0, x^C_1 \in [0, 1]$.

1. $\gamma \leftarrow 1/l$
2. $t_0 \leftarrow 0$
3. $t_1 \leftarrow lm + 1$
4. $X(t) \leftarrow \mathcal{I}_S \cup \mathcal{I}_T$ for all $t \in [t_0, t_1]$
5. $X(t) \leftarrow \emptyset$ for all $t \in (lu, t_1]$
6. $x_i(t) \leftarrow x^S_i$ for all $i \in \mathcal{I}_S, t \in [lu, t_1]$
7. $x_i(t) \leftarrow x^T_i$ for all $i \in \mathcal{I}_T, t \in [lu, t_1]$
8. $x_i(t) \leftarrow 0$ for all $i \in \mathcal{I}_C, t \in [t_0, t_1]$
9. $x_i(t) \leftarrow 0$ for all $i \in \mathcal{I}, t \in (lu, t_1]$
10. $T(t) \leftarrow \emptyset$ for all $t \in (t_0, t_1 - lv)$
11. $T(t) \leftarrow \mathcal{I}$ for all $t \in [t_1 - lv, t_1]$
12. $g_i(t) \leftarrow 0$ for all $i \in \mathcal{I}, t \in (t_0, t_1 - lv)$
13. $g_i(t) \leftarrow x^S_i$ for all $i \in \mathcal{I}_S, t \in [t_1 - lv, t_1]$
14. $g_i(t) \leftarrow x^T_i$ for all $i \in \mathcal{I}_T, t \in [t_1 - lv, t_1]$
15. $g_0(t) \leftarrow x^C_0$ for all $t \in [t_1 - lv, t_1]$
16. $g_1(t) \leftarrow x^C_1$ for all $t \in [t_1 - lv, t_1]$
17. $a, e \leftarrow \text{ForwardPass}(\mathcal{N}, w, [t_0, t_1], \gamma, X, x, T, g)$
18. $\text{BackwardPass}(\mathcal{N}, w, [t_0, t_1], \gamma, \eta, X, a, e)$

Learning to how much the agent associates the target with the other type of information. The reduction is adjusted by a parameter $\beta_{\text{quantity}}$. This means that the more the audience is assumed to know about one type of information, the more the agent talks about the other, providing information novel to the audience. Also, the less an agent is able to talk about one type of information, the more it talks about the other (proportionally).

Receiving

We discussed previously that people turn to their stereotypes or in fact their prior knowledge when assessing information that they receive, mediated by the maxim of quality. So an agent receiving an $(SC, SI)$ tuple about a particular target adjusts this information towards its current association between target and content before memorizing it.
We describe first the procedure of learning information because we will need to refer back to this one repeatedly. This procedure is illustrated in Algorithm 5.4. When learning to associate subject, target, and content, the agent first constructs an input pattern consisting of a representation of the subject and the target (lines 5 to 10). Then, a target pattern is constructed from a representation of the subject, target, and the content (lines 11 to 17). During the forward pass (line 18), the input pattern is supplied to the network in the first $u$ intervals. The target pattern is supplied to the network during the last $v$ intervals. Then, errors are back-propagated and weight updates are performed (line 19). While not shown in the algorithm, normally distributed noise with $\mathcal{N}(0, 0.01)$ is added to all entries in the input vectors before forward and backward pass are executed. However, as with all other operations on input values in this chapter, inputs are always restricted to the range $[0, 1]$.

The details of the procedure of receiving content are shown in Algorithm 5.5. The first part of recalling the SC and SI information associated with a given subject and content from line 3 to 14 is the same as in Algorithm 5.3. The maxim of quality is represented in lines 15 and 16, in which the received information is made more similar to the one reproduced from memory. After this, the agent learns the adjusted information according to Algorithm 5.4 (line 18).

5.4 Parameter Estimation

In this section, we describe how we estimated the parameters of the model described above based on the data for the experiment of Lyons and Kashima described in Section 5.1. We first describe evolutionary algorithms as a tool for parameter optimization. Then we introduce one such algorithm that optimizes the parameter values of our model to yield simulations with an outcome close to the empirical data. Finally, we present our results in form of a visualization of selected parameter values and goodness of fit estimates.

5.4.1 Evolutionary Algorithms

The parameter estimation of an agent-based model can be framed in terms of an optimization problem: Find those values $\hat{\theta}$ for the parameters $\theta$ that minimize a distance function $d(D_{obs}, D_{\theta})$ between the observed empirical data $D_{obs}$ and the data $D_{\theta}$ the model generates under these parameters. As described in Section 5.1, the data $D_{obs}$ and $D_{\theta}$ in our case comprises the average SC and SI content transmitted in different experimental conditions and at different positions in the serial reproduction chains.

In the case of complex stochastic agent-based models, which have nonlinear relationships between parameters as well as between parameters and variables, an analytical form of the model and hence of $d(D_{obs}, D_{\theta})$ is unavailable. Therefore, $\hat{\theta}$ cannot be found analytically.
Yet in our case we can reasonably expect that values according to some local rules, e.g. following the gradient of \( d \), there exist more patterns in the parameter space and multiple local optima than local search algorithms do. They are easy to implement, but typically depend severely on the initial parameter value and are prone to getting caught in local optima. Similarly, the parameter space is usually high-dimensional and continuous, so that an exhaustive enumeration of all possible parameter values to find \( \hat{\theta} \) is unachievable. Local search methods such as hill climbing or simulated annealing promise relief. A local search starts with some initial parameter value \( \theta_0 \) and iteratively moves to neighboring parameter values according to some local rules, e.g. following the gradient of \( d(D_{obs}, D_0) \). Figure 5.5 illustrates a local search on a two-dimensional parameter space. However, these methods depend severely on the initial parameter value and are prone to getting caught in local optima. Yet in our case we can reasonably expect that \( d(D_{obs}, D_0) \) has multiple such local optima.

As we will see, evolutionary algorithms can cope more successfully with large parameter spaces and multiple local optima than local search algorithms do. They are easy to implement,
do not make any assumptions about $d(D_{\text{obs}}, D_{\theta})$, and can deal with parameters that comprise variables of different types.

The idea behind evolutionary algorithms is based on the theory of natural selection, first advanced by Charles Darwin (1859). According to this theory, organisms are frequently challenged by their environment. Only those that are adapted best to their environment survive and reproduce, thereby having the chance to pass on their genetic characteristics to their offspring. In the long run, this process selects those organisms and characteristics that are best adapted to the environment.

Evolutionary algorithms for parameter optimization mimic nature’s behavior in the following way (Eiben and Smith, 2003): The individuals of the population of organisms are a set of $\mu$ possibly multi-dimensional parameter values $(\theta_1, \ldots, \theta_\mu)$ with $\theta_i \in \Theta$ where $\Theta$ is the set of possible parameter values. A parameter value $\theta$ gives rise to a model instance $M(\theta) \in \mathcal{M}$. We call the parameter $\theta$ the genotype of the individual in the population, and $M(\theta)$ its phenotype. The function $M : \Theta \rightarrow \mathcal{M}$ maps every parameter to its associated model. It is the phenotype that is subject to the selection by the environment. Selection relies on a fitness function $F : \mathcal{M} \rightarrow \mathbb{R}$. The fitness function maps a model to a real-valued fitness value. The higher (or lower) the fitness value of a model, the better the model performs and the more likely the survival of its parameter.

As depicted in Figure 5.6, an evolutionary algorithm proceeds over multiple iterations or in fact generations. Initially, a population of parameters is drawn randomly from $\Theta$ (1). Until a termination criterion is reached (5), individuals from the population are chosen for reproduction (2), the genotypes of their offsprings undergo mutation (3), and the best individuals with regard to their fitness value are selected to be carried over into the next generation (4). Over generations, this process optimizes the fitness function in that the
Chapter 5. An Agent-based Model of Stereotype Communication

Initialization:
Create random initial population \((\theta_1, \ldots, \theta_\mu)\) from \(\Theta\).

Reproduction:
Create \(\lambda\) offsprings \((\theta'_1, \ldots, \theta'_\lambda)\) from parents \((\theta_1, \ldots, \theta_\mu)\).

Mutation:
Change the offsprings \((\theta'_1, \ldots, \theta'_\lambda)\) randomly.

Selection:
Evaluate the fitness \(F(M(\theta))\) for the model of every offspring (and possibly every parent) \(\theta\) and retain the \(\mu\) best ones for the next generation.

Termination:
Check whether any of the termination criteria are fulfilled.

Figure 5.6: The steps taken by an evolutionary algorithm.

genotypes of the population give rise to models with increasingly higher fitness values. We will discuss briefly some basic variations in which reproduction, mutation, and selection have been implemented.

Reproduction During reproduction, \(\lambda\) offsprings are created from the \(\mu\) parents. To create an offspring, either a copy of a single, randomly selected parent is created, or \(\rho\) randomly selected parents are recombined to yield a new individual. Recombination can happen in different ways: a) Each entry \(i\) in the parameter vector \(\theta\) of the offspring is copied from the entry \(i\) of a randomly selected parent; b) predefined crossover points in the parameter vectors specify which entries are copied from which parent; c) the value of an entry \(i\) in the parameter vector of the offspring is obtained by averaging the values of all the entries \(i\) in the parents’ parameter vectors, in case the entry is a continuous value.

Mutation Mutation changes the entries of the offsprings’ parameter vectors randomly, typically by adding some normally distributed random noise. The variance of the noise distribution determines how severe the effect of mutation is. It is common to make noise variances themselves entries in the parameter vectors, so that the precision of the algorithm can adapt autonomously over time as well. While initially large variances support the rough exploration of the parameter space, the eventual approach to an optimum is likely to benefit from the precision of smaller variances.
5.4. Parameter Estimation

Figure 5.7: An evolutionary algorithm with a population size of 5 searching for a value $\hat{\theta}$ of $\theta$ that minimizes $d(D_{ob}, D_{\theta})$ on a two-dimensional parameter space. Note how one individual finds a local optimum in generation $g_1$ but the population escapes this local optimum until generation $g_2$.

Selection  Selection evaluates all individuals in the population based on their fitness and retains the $\mu$ best ones for the next generation. If $\lambda > \mu$, an option is to discard all $\mu$ parents and retain the $\mu$ best among the $\lambda$ offsprings. This strategy is denoted as $(\mu, \lambda)$. Otherwise, selection can consider both parents and offsprings for survival. This strategy is denoted as $(\mu + \lambda)$. While the first selection strategy runs the risk of “forgetting” good parameter values, the latter runs the risk of being dominated by a few parameter values.

In summary, the following entities of an evolutionary algorithm for parameter optimization need to be specified:

- The set of possible parameters $\Theta$.
- The size $\mu$ of the population, the number $\lambda$ of offsprings in every generation, and the number $\rho$ of parents for every offspring.
- The function $M$ that maps parameter values onto models.
- The fitness function $F$.
- A termination criterion.
- Recombination, mutation, and selection operators.

Figure 5.7 illustrates how a population driven by an evolutionary algorithm closes in onto the global optimum of a fitness function.

Evolutionary algorithms actually comprise a host of different techniques such as evolutionary strategies or genetic algorithms, which implement the process described above to
Chapter 5. An Agent-based Model of Stereotype Communication

varying degrees (Eiben and Smith, 2003). For example, evolutionary strategies originally did not consider recombination during reproduction.

The advantages that evolutionary algorithms have over conventional local search methods should be clear from this description. Because evolutionary algorithms perform as many searches in parallel as there are members in the population, they are both less likely to get stuck in a local optimum and their computation is easily executed in parallel. Evolutionary algorithms also benefit from the cooperation between different candidate solutions during recombination.

5.4.2 Method

We describe first how our model is set up to reflect the experimental conditions of Lyons and Kashima. Then we provide the details of the evolutionary algorithm used to conduct the parameter estimation.

Experimental Setup

We represent each serial reproduction chain in the experiments of Lyons and Kashima by 4 agents arranged in a row. All agents undergo an initial learning phase, in which they learn to associate the self with the Jamayans and a particular \((SC, SI)\) tuple as well as the audience with the Jamayans and a particular \((SC, SI)\) tuple. This procedure is shown in Algorithm 5.6. For every agent, network weights are initialized randomly (line 3) and patterns representing the self (line 4), the audience (line 5), the group of Jamayans (line 6), and Jay (line 7) are created randomly. Every agent makes use of its own particular patterns for the self, the audience, the Jamayans, and Jay but this is not of importance here.

After patterns have been created, agents learn particular associations between subject, target, and content depending on the experimental condition of the chain the agent belongs to and on its position within that chain (lines 8 to 21). Table 5.1 summarizes the patterns learnt by the agents in the different conditions and positions.

For each of the 4 experimental conditions we consider 8 such chains. The chains differ by the random initialization of network weights and patterns of the agents. The procedure for propagating the story about Jay through a serial reproduction chain of agents is shown in Algorithm 5.7. First, network units are determined (lines 2 to 10). Then agents are initialized as described in Algorithm 5.6 (line 11). After agents have been initialized, the first agent receives an \((SC, SI)\) tuple of \((1, 1)\) to reflect that the story consists of an equal proportion of SC and SI information (line 12). The number of learning iterations for the story information is \(\frac{2}{3} k\)—the number of learning iterations used for learning the background information about the Jamayans. This is the proportion between the relevant content in the report about
the Jamayans and the content in the story about Jay used in the experiments of Lyons and Kashima. The agent then reproduces the information as described in Section 5.3, yielding an (SC, SI) tuple (line 14). The next agent in the chain receives this information (line 16), stores and reproduces it, and so forth. We record the communicated (SC, SI) tuples at all positions in the chains as representative of the reproduced SC and SI content (line 17).

The computational complexity of each interaction (send and receive) in this model is dominated by the complexity of the Receive procedure. The complexity of this procedure, in

5.4. Parameter Estimation

Algorithm 5.6: Setup($\mathcal{N}$, actualSharedness, perceivedSharedness, $k,l,m,o,u,v,\eta$)

Data:
- Units in the network: $\mathcal{N} = \mathcal{I} \cup \mathcal{H} \cup \{\text{bias}\}$ with $\mathcal{I} \cap \mathcal{H} = \emptyset$, $\text{bias} \notin (\mathcal{I} \cup \mathcal{H})$,
- $\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_T \cup \mathcal{I}_C$,
- $\mathcal{I}_C = \{0,1\}$,
- $\mathcal{I}_S \cap \mathcal{I}_T = \emptyset$, $\{0,1\} \cap (\mathcal{I}_S \cup \mathcal{I}_T) = \emptyset$.
- actualSharedness: shared or unshared.
- perceivedSharedness: ignorance or knowledge.
- Number of iterations $k$ for which background information is learnt.
- Number of ticks per interval $l$.
- Number of intervals $m$.
- Overlap between Jamayans and Jay pattern $o \in [0,1]$. We have $\mathcal{I}_T = \mathcal{I}_{T,0} \cup \mathcal{I}_{T,1}$.
- Number of intervals $u$ during which input is provided, $u \leq m$.
- Number of intervals $v$ during which a target is provided, $v \leq m$.
- Learning rate $\eta$.

Result: Network weights and subject and target patterns for all agents.

1 for $a \in \{1,2,3,4\}$ do
  2 \begin{align*}
  w^{\text{bias}}_{ij} & \sim U(0,0.1) - 0.05 \text{ for all } i \in (\mathcal{I} \cup \mathcal{H}), j \in \mathcal{N}, i \neq j \\
  x^{\text{S},a}_i & \sim U(0,1) \text{ for all } i \in \mathcal{I}_S \\
  x^{\text{A},a}_i & \sim U(0,1) \text{ for all } i \in \mathcal{I}_S \\
  x^{\text{G},a}_i & \sim U(0,1) \text{ for all } i \in \mathcal{I}_T \\
  x^{\text{J},a}_i & \leftarrow x^{\text{G},a}_i(\mathcal{a}) \text{ for all } i \in \mathcal{I}_{T,0}, x^{\text{J},a}_i \sim U(0,1) \text{ for all } i \in \mathcal{I}_{T,1}
  \end{align*}

  3 for $c = 1$ to $k$ do
    4 \begin{align*}
    & \text{if actualSharedness = shared then} \\
    & \quad \text{if perceivedSharedness = ignorance then} \\
    & \quad \quad \text{Learn($\mathcal{N}$, $w^a$, $l$, $m$, $u$, $v$, $\eta$, $x^{\text{S},a}$, $x^{\text{G},a}$, 1, 0)} \\
    & \quad \text{else if perceivedSharedness = knowledge then} \\
    & \quad \quad \text{Learn($\mathcal{N}$, $w^a$, $l$, $m$, $u$, $v$, $\eta$, $x^{\text{A},a}$, $x^{\text{G},a}$, 0, 0)} \\
    & \quad \text{else if perceivedSharedness = unshared then} \\
    & \quad \quad \text{if } a \mod 2 = 1 \text{ then } (\text{sc,si}) \leftarrow (1,0); \text{ else } (\text{sc,si}) \leftarrow (0,1) \\
    & \quad \quad \text{Learn($\mathcal{N}$, $w^a$, $l$, $m$, $u$, $v$, $\eta$, $x^{\text{S},a}$, $x^{\text{G},a}$, $\text{sc,si}$)} \\
    & \quad \text{if perceivedSharedness = ignorance then} \\
    & \quad \quad \text{Learn($\mathcal{N}$, $w^a$, $l$, $m$, $u$, $v$, $\eta$, $x^{\text{A},a}$, $x^{\text{G},a}$, 0, 0)} \\
    & \quad \text{else if perceivedSharedness = knowledge then} \\
    & \quad \quad \text{Learn($\mathcal{N}$, $w^a$, $l$, $m$, $u$, $v$, $\eta$, $x^{\text{A},a}$, $x^{\text{G},a}$, $\text{sc,si}$)}
    \end{align*}

5 return $w^a, x^{\text{S},a}, x^{\text{A},a}, x^{\text{G},a}, x^{\text{J},a}$ for all $a \in \{1,2,3,4\}$
Table 5.1: An overview of the patterns that agents learn in the priming phase of the different experimental conditions (see Algorithm 5.6 and text for details). Every tuple in the cells consists of two tuples: One representing an input pattern over the subject and target units, and one pattern over the content (SC and SI) units. $x^S$ and $x^A$ stand for patterns of activation corresponding to the self and the audience respectively. $x^G$ stands for a pattern of activation corresponding to the Jamayans. In the shared-ignorance condition, for example, all agents learn to associate the self with the Jamayans and SC information but not SI information. They also learn to associate the audience with the Jamayans with no SC or SI information.

<table>
<thead>
<tr>
<th>Ignorance Condition</th>
<th>Shared Condition</th>
<th>Unshared Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agents 1-4:</td>
<td>$[(x^S, x^G), (1,0)]$</td>
<td>$[(x^A, x^G), (0,0)]$</td>
</tr>
<tr>
<td>Agents 1,3:</td>
<td>$[(x^S, x^G), (1,0)]$</td>
<td>$[(x^A, x^G), (0,0)]$</td>
</tr>
<tr>
<td>Agents 2,4:</td>
<td>$[(x^A, x^G), (0,1)]$</td>
<td>$[(x^A, x^G), (0,0)]$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Condition</th>
<th>Shared Condition</th>
<th>Unshared Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agents 1-4:</td>
<td>$[(x^S, x^G), (1,0)]$</td>
<td>$[(x^A, x^G), (1,0)]$</td>
</tr>
<tr>
<td>Agents 1,3:</td>
<td>$[(x^S, x^G), (1,0)]$</td>
<td>$[(x^A, x^G), (1,0)]$</td>
</tr>
<tr>
<td>Agents 2,4:</td>
<td>$[(x^S, x^G), (0,1)]$</td>
<td>$[(x^A, x^G), (0,1)]$</td>
</tr>
</tbody>
</table>

Evolutionary Algorithm Setup

Every parameter $\theta_i \in \Theta$ in the population of $\mu$ parameters is a vector $(\theta_i^1, \theta_i^2, \theta_i^3, \ldots)$ of entries $\theta_i^j$, each standing for one of the model parameters introduced previously. Table 5.2 lists all model parameters $\theta^j$ and their sets of possible values $\Theta^j$. Some model parameters are fixed, the first 5 are not. Hence, for the purpose of parameter optimization each parameter $\theta_i$ is a tuple $(\theta_i^1, \ldots, \theta_i^5)$ and the set of possible parameters $\Theta = \Theta^1 \times \ldots \times \Theta^5$ is the set of all possible tuples. We opt for restricting the range of valid values for $\theta^2$ to $[0, 0.2]$. Larger values of $\eta$ usually enable faster learning but can also cause instabilities that need to be compensated for by adjustments to the learning procedure (Rumelhart et al., 1986). Examples of such adjustments are the decay of weights over time or a momentum term so that weight updates are influenced by earlier updates in addition to the current one. These adjustments, however, would make the model more complicated. The value of parameter $\theta^6$ is fixed to 0.5 reflecting that a reasonable assumption is that Jay is perceived to be in equal proportions an individual and a Jamayan group member (in the story about Jay, half of the
5.4. Parameter Estimation

Algorithm 5.7: Chain(actualSharedness, perceivedSharedness, d, h, k, l, m, o, u, v, $\beta_{\text{quantity}}$, $\beta_{\text{quality}}$)

Data:
- actualSharedness: shared or unshared.
- perceivedSharedness: ignorance or knowledge.
- Number of input-output units per pool $d$.
- Number of hidden units $h$.
- Number of intervals $k$ for which background information is learnt.
- Number of ticks per interval $l$.
- Number of intervals $m$.
- Overlap between Jamayans and Jay pattern $o \in [0, 1]$.
- Number of intervals $u$ during which input is provided, $u \leq m$.
- Number of intervals $v$ during which a target is provided, $v \leq m$.
- Learning rate $\eta$.
- The strength of the maxim of quality $\beta_{\text{quality}} \in [0, 1] \subset \mathbb{R}$.
- The strength of the maxim of quality $\beta_{\text{quantity}} \in [0, 1] \subset \mathbb{R}$.

Result: Measured SC and SI content for all agents in the chain.

begin
1. $I_c \leftarrow \{0, 1\}$
2. $I_s \leftarrow \{2, 3, \ldots, d + 1\}$
3. $I_{t, 0} \leftarrow \{d + 2, d + 3, \ldots, d + od + 1\}$
4. $I_{t, 1} \leftarrow \{d + od + 2, d + od + 3, \ldots, 2d + 1\}$
5. $H \leftarrow \{2d + 2, 2d + 3, \ldots, 2d + h + 1\}$
6. $\text{bias} \leftarrow 2d + h + 2$
7. $I_t \leftarrow I_{t, 0} \cup I_{t, 1}$
8. $I \leftarrow I_s \cup I_t \cup I_c$
9. $N \leftarrow I \cup H \cup \{\text{bias}\}$
10. $w^a, x^{S,a}, x^{A,a}, x^{G,a}, x^{d,a}$ for all $a \in \{1, 2, 3, 4\}$ \(\rightarrow\) Setup($N$, actualSharedness, perceivedSharedness, $k$, $l$, $m$, $o$, $u$, $v$, $\eta$)
11. Receive($N$, $w^a$, $\frac{2}{3}k$, $l$, $m$, $u$, $v$, $\eta$, $x^{S,0}$, $x^{T,0}$, $\beta_{\text{quality}}$, $1$, $1$)
12. for $a = 1$ to 4 do
13. SC$^a$, SI$^a$ \(\leftarrow\) Send($N$, $w^a$, $l$, $m$, $u$, $\eta$, $x^{S,a}$, $x^{A,a}$, $x^{G,a}$, $x^{d,a}$, $\beta_{\text{quantity}}$
14. if $a < 4$ then
15. Receive($N$, $w^{a+1}$, $\frac{2}{3}k$, $l$, $m$, $u$, $v$, $\eta$, $x^{S,a+1}$, $x^{T,a+1}$, $\beta_{\text{quality}}$, SC$^a$, SI$^a$
16. return SC$^a$, SI$^a$ for all $a \in \{1, 2, 3, 4\}$

items are stereotype-consistent, half of the items are stereotype-inconsistent). The value of parameter $\theta_7$ is set arbitrarily to a reasonably large number. Previous work that considered recurrent networks for the representation of stereotype-relevant information has opted for similarly or smaller sized networks (Queller and Smith, 2002; Smith and DeCoster, 1998; Van Overwalle and Heylighen, 2006; Van Rooy et al., 2003; Van Rooy, 2009). The values of parameters $\theta^8$ to $\theta^{11}$ are set based on previous work (Rogers et al., 2004).

We set the size of the population in every generation of the evolutionary algorithm to $\mu = 50$, the number of offsprings to $\lambda = 20$, and the number of parents per offspring to $p = 2$. Preliminary tests have proven this to be sufficient.
Chapter 5. An Agent-based Model of Stereotype Communication

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Set of possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta^1$</td>
<td>Number of iterations $k$ for learning background knowledge about Jamayans</td>
<td>$\Theta^1 = {10, 20, 50, 100, 200, 500, 800, 1000, 1500, 2000}$</td>
</tr>
<tr>
<td>$\theta^2$</td>
<td>Learning rate $\eta$</td>
<td>$\Theta^2 = [0, 0.2] \subset \mathbb{R}$</td>
</tr>
<tr>
<td>$\theta^3$</td>
<td>Strength of maxim of quality $\beta_{quality}$</td>
<td>$\Theta^3 = [0, 1] \subset \mathbb{R}$</td>
</tr>
<tr>
<td>$\theta^4$</td>
<td>Strength of maxim of quantity $\beta_{quantity}$</td>
<td>$\Theta^4 = [0, 1] \subset \mathbb{R}$</td>
</tr>
<tr>
<td>$\theta^5$</td>
<td>Number of hidden units $h$ in the network</td>
<td>$\Theta^5 = {0, 1, 5, 10, 20, 40, 50}$</td>
</tr>
<tr>
<td>$\theta^6$</td>
<td>Proportion $\rho$ of units in the patterns of the Jamayans and Jay that overlap</td>
<td>$\Theta^6 = {0.5}$</td>
</tr>
<tr>
<td>$\theta^7$</td>
<td>Number of units $d$ for the subject and target pools</td>
<td>$\Theta^7 = {20}$</td>
</tr>
<tr>
<td>$\theta^8$</td>
<td>Number of intervals $m$ for recall and learning</td>
<td>$\Theta^8 = {7}$</td>
</tr>
<tr>
<td>$\theta^9$</td>
<td>Number of ticks per interval $l$</td>
<td>$\Theta^9 = {4}$</td>
</tr>
<tr>
<td>$\theta^{10}$</td>
<td>Number of initial intervals $u$ of the $\theta^8$ intervals during which any input pattern is provided</td>
<td>$\Theta^{10} = {3}$</td>
</tr>
<tr>
<td>$\theta^{11}$</td>
<td>Number of last intervals $v$ of the $\theta^8$ intervals during which any target pattern is provided</td>
<td>$\Theta^{11} = {2}$</td>
</tr>
</tbody>
</table>

Table 5.2: An overview of all parameters of the model. Some parameters have fixed values, others are provided with a set of possible values from which the evolutionary algorithm can create parameters.

During recombination, for each entry in the offspring’s parameter vector the parent that provides the corresponding entry is selected randomly. Mutation distinguishes between those entries in the offspring’s parameter vector holding discrete values and those entries holding continuous values. Discrete-valued entries are mutated by randomly retaining the current value or selecting the value with the next lower or higher value for that parameter entry (see Table 5.2). Continuous-valued entries are mutated by adding normally distributed random noise $\mathcal{N}(0, \sigma_g)$ to the current value, where $\sigma_g$ depends on the current generation $g$. We set $\sigma_1 = 1.0$ and $\sigma_g = 0.99 \ast \sigma_{g-1}$. Thus, the precision of the evolutionary algorithm with respect to continuous-valued entries in the parameter vectors increases with generation. The selection strategy applied here is $(\mu + \lambda)$.

The function $M$ that maps parameter values to models is a function that creates an experimental setup as described in the previous section (starting from Algorithm 5.7) from a given parameter $\theta$. We then run the experiments under this setup and create the same 30 summary statistics as Lyons and Kashima, which we described in Section 5.1. The fitness function $F$ is here the squared correlation coefficient $R^2$ between the summary statistics.
obtained by Lyons and Kashima ($D_{obs}$) and the ones obtained by running the simulation model ($D_{θ}$). We are not interested in an exact reproduction of the experimental results but in reproducing the pattern of the data. Therefore, using the correlation coefficient as a distance measure is appropriate. The evolutionary algorithm is terminated after 100 iterations.

### 5.4.3 Results

Figure 5.8 displays the distribution of fitness values for all generations of the evolutionary algorithm. Fitness values and in fact $R^2$ values undergo a rapid initial increase and then level off approximately between generations 70 and 100. This indicates that running the algorithm for 100 generations is reasonable. The best performing parameter found by the optimization method produced an $R^2$ value of 0.96. Repeated executions of this optimization also yielded best fitness values with $0.95 < R^2 < 0.96$. This consistency contributes to the confidence we can have in the optimization method and in the ability of the model to be calibrated to the data used here.

Figure 5.9 shows the distribution of parameter values for those 150 parameters from all generations that yielded $R^2 \geq 0.95$. The values of the best performing parameter are marked with a red line in each plot. The number of learning iterations ($θ^1$) is generally high, which indicates that learning in each iteration is slow. Indeed, the bulk of the learning rate $η (θ^2)$ is at the lower end of the considered range, which is in line with typical manual choices. The strength of the maxim of quality $β_{quality} (θ^3)$ also shows a small spread relative to the considered range. Apparently, the model is more sensitive to $θ^2$ and $θ^3$ than to other parameters, which show a larger spread. The strength of the maxim of quantity $β_{quantity} (θ^4)$ extends over a range of about 0.5. The number of hidden units ($θ^5$) is substantial in most parameters, which suggests a crucial role for hidden units. However, further investigations would be required to assess whether hidden units are definitely required
Chapter 5. An Agent-based Model of Stereotype Communication

Figure 5.9: Distribution of parameter values for all parameters created during the evolutionary algorithm that yielded $R^2$ values $\geq 0.95$. Red lines mark the parameter values of the parameter that yielded the best $R^2$ value.

in this scenario. Regardless of this, hidden units increase the variety of patterns that can be stored in the network, which enables the model to represent more complex scenarios of stereotype communication. This allows agents, for example, to store associations for a set of different subjects and targets.

Figures 5.10 and 5.11 show the regression-adjusted results of the best performing parameter plotted onto the results of the experiment of Lyons and Kashima.\(^3\) These are in fact the plots presented by them apart from an aggregation of the data for different story zones that we conducted. Obviously, the match between simulation and experimental data is within a reasonable error. We take that the model is generally capable of reproducing the results of Lyons and Kashima.

However, a remark is expedient about the plot for the unshared condition in Figure 5.10. We see that at the last position in the simulation more SI content is reproduced than in the previous position despite the overall trend of decrease. Hence in this regime, the model does not reflect the empirical data exactly. In the unshared condition, agents in adjacent positions in the chain learn contradicting information about the Jamayans during the priming phase. The stereotype held by agents at positions 1 and 3 associates SC information with the Jamayans but the stereotype held by agents at positions 2 and 4 associates SI information with the Jamayans. It appears that in this simulation agents do not give the proper weight to

\(^3\)Note that we do not show error bars here and in some of the following plots because each data point is an average over multiple levels with different averages themselves.
the information carried by the story about Jay in proportion to the stereotype learnt previously about the Jamayans. Their communication is affected by their stereotypes a bit too much. In fact, among the 150 parameters that yielded \( R^2 \geq 0.95 \), there are 27 for which no position in the unshared condition shows more reproduced content than the previous position. An adjusted optimization function would be able to yield a set of parameter values that fulfill this requirement.

### 5.5 Validation

We base the validation of the model on a second experiment by Lyons and Kashima (2003). This experiment has the same setup as the previous one but partial knowledge is introduced as a new level of perceived sharedness. In the partial knowledge condition, participants are told that their audience has some knowledge of the Jamayan stereotype. Recall that in the ignorance and knowledge conditions participants were told that their audience has none or complete knowledge of the stereotype respectively. In this second experiment, Lyons and
Kashima measure the SC and SI information transmitted in the shared-partial condition, i.e. the stereotype is actually shared but only partial knowledge of the stereotype is expected from the audience.

For this second experiment, Lyons and Kashima report the following 6 data points:

- Overall proportion of SC and SI content reproduced.
- Proportion of content reproduced at every position in the chains (4 data points).

They find that the stereotype bias in the shared-partial condition is less than in the shared-ignorance condition but more than in the shared-knowledge condition. Interestingly, they also observe that overall less information (SC and SI) is reproduced. A possible explanation is that participants were generally not that motivated as in the first experiment (personal correspondence with the authors). However, this poses a challenge for the validation of our model because the model was implicitly calibrated to the circumstances under which the first experiment was performed. Therefore, for this validation we do not rely on the absolute values obtained by Lyons and Kashima for the shared-partial condition but on the relations between them. We assume these relations to be rather independent of the circumstances under which the experiments are conducted. We calculate the following 4 data points against which we validate our model:

- Overall SC content reproduced divided by overall SI content reproduced.
- Content reproduced at position 1 divided by content reproduced at positions 2,3,4.

For the setup and execution of the model we follow the same procedure as above. Following the notation in Table 5.1, we represent the shared-partial condition by all agents learning the following two patterns during the priming phase: \([x^S, x^G, (1,0)]\) and \([x^A, v^G, (p,0)]\). The parameter \(p\) with \(0 \leq p \leq 1\) determines the extent to which agents assume the SC information to be shared by the audience. Hence \(p < 1\) denotes that partial knowledge of the stereotype is expected from the audience. In Algorithm 5.6 this requires an additional condition to be added to the for-loop:

| 22 | else if actualSharedness = shared and perceivedSharedness = partial then |
| 23 | Learn\((N, w^a, I, m, u, v, \eta, x^S, x^G, 1, 0)\) |
| 24 | Learn\((N, w^a, I, m, u, v, \eta, x^A, v^G, p, 0)\) |

Considering that the quality of all parameter estimates from the previous section are similar (i.e. \(R^2 > 0.95\)), we conduct the validation for all of them and report on the parameter that performs best during validation. This procedure was also proposed by Helbing and
5.5. Validation

Balietti (2011). We measure the overall SC and SI content reproduced as well as the overall content (SC and SI) reproduced at each position in the chain. We apply the regression adjustment to these values according to the regression coefficients determined for that parameter during parameter estimation. Afterwards, we calculate the 4 summary statistics listed above.

We first set $p = 0.5$. For the best performing parameter, we find the mean absolute error (MAE) between the simulation data and the empirical data to be 0.15. Figure 5.12a shows that our model qualitatively reproduces the results of Lyons and Kashima: The stereotype bias in the shared-partial condition is smaller than in the shared-ignorance condition but larger than in the shared-knowledge condition. We find the following overall content reproduced at the four positions in the chain: 0.79, 0.59, 0.50, 0.45. The decrease over the positions of the chain is in line with what is expected.

Figure 5.12: The proportion of SC and SI information reproduced at each position in the shared-ignorance, shared-partial, and shared-knowledge conditions for $p = 0.5$ and $p = 0.25$. 

Balietti (2011). We measure the overall SC and SI content reproduced as well as the overall content (SC and SI) reproduced at each position in the chain. We apply the regression adjustment to these values according to the regression coefficients determined for that parameter during parameter estimation. Afterwards, we calculate the 4 summary statistics listed above.

We first set $p = 0.5$. For the best performing parameter, we find the mean absolute error (MAE) between the simulation data and the empirical data to be 0.15. Figure 5.12a shows that our model qualitatively reproduces the results of Lyons and Kashima: The stereotype bias in the shared-partial condition is smaller than in the shared-ignorance condition but larger than in the shared-knowledge condition. We find the following overall content reproduced at the four positions in the chain: 0.79, 0.59, 0.50, 0.45. The decrease over the positions of the chain is in line with what is expected.
We now set $p = 0.25$. The best performing parameter is found to differ from the empirical data with an MAE of 0.13, which is slightly better than the one above. Again, Figure 5.12b confirms that our results correspond qualitatively to the empirical data. For this setting, the following overall content is reproduced at the four chain positions: 0.76, 0.54, 0.44, 0.40.

In summary, the model could be validated qualitatively and we also find a good quantitative match between simulation results and empirical data.

5.6 Sensitivity Analysis

In this section, we analyze the sensitivity of the validated model to local changes in parameters. Our goal is to identify those parameters to which the model is most sensitive.

We create instances of the model each of which differs from the validated model with $p = 0.5$ only in local changes of one of the parameters $\theta_1$ to $\theta_6$. We create these local changes by adding uniformly distributed random noise within $\pm 10\%$ to the respective parameter. Given a parameter value $\theta$ and a perturbed version $\theta'$, we calculate the relative change in the value of the parameter as $|\theta' - \theta|/\theta$, which falls in the interval $[0, 0.1]$. Then we simulate the validation scenario and record again the mean absolute error MAE between the empirical data and the one obtained by simulation. Given the original MAE of 0.15 as reported above and the MAE obtained from a simulation run with a perturbed parameter, the absolute change in error is MAE$-0.15$. For each of the parameters $\theta_1$ to $\theta_6$ we run 100 such simulations with different randomly drawn changes to the respective parameter. We focus our sensitivity analysis on a limited set of parameters as we feel that all others are properly based on previous work.

Figure 5.13a shows the relationship between changes in the value of each of the parameters and the resulting change in the MAE for all 100 instances of each parameter. As indicated above, the relative change in the value of parameters is within the interval $[0, 0.1]$. We display regression lines with confidence intervals with level 0.95. Only in the case of $\eta$, the change in error grows reliably with increasing relative change of $\eta$. For all other parameters, the change in error with increasing change in the parameter is negligible. However, overall most of the changes of a parameter lead to an increase in the MAE, i.e. to a less close fit with the empirical data. Figure 5.13b confirms this picture. This figure shows the distribution of the change in error obtained from all 100 simulation runs for each of the parameters. We see that for all parameters the bulk of the distribution lies between a change of 0 and 0.05. This is comparably low. However, in particular a change in $\eta$ or in the number of hidden units is able to cause a change in error of up to 0.25. Given the original MAE of 0.15, this means that some of these simulations yield an MAE of up to 0.4.
5.6. Sensitivity Analysis

![Graph](image)

(a) The relationship between changes in the value of each of the parameters and the resulting change in the MAE for all 100 instances of each parameter.

![Graph](image)

(b) The distribution of the change in error obtained from all 100 simulation runs for each of the parameters.

Figure 5.13: Results of the sensitivity analysis.

We conclude that while most local parameter changes do not cause substantial deviations from the fit of the validated model, the model is particularly sensitive to changes in \( \eta \) and in the number of hidden units. This suggests that the network is a crucial factor in the performance of the model.
5.7 The Influence of Network Configuration on Stereotype Communication

In the following, we demonstrate how this model can be used to further explore the effects of communication on the maintenance of stereotypes. Communicators in general assume an unfamiliar audience to have at least partial knowledge of common stereotypes (Lyons and Kashima, 2003). A familiar audience, in contrast, is likely assumed or known to have full knowledge of stereotypes. Following the results described before, interactions with an unfamiliar audience are conducive to an SC bias while interactions with a familiar audience are not. Consider now that individuals interact frequently with familiar others within their social communities but less often with unfamiliar people from other communities. SI information from first-hand experiences with members of a stereotyped group can spread within the community of the observer. However, the diffusion of this information to other communities is hindered by the assumption of partial knowledge of the stereotype.

5.7.1 Model

We explore how the rates of first-hand experiences providing SC as well as SI information, second-hand within-community information exchange, and second-hand between-community information exchange affect the expression of stereotype-relevant information. In doing so, we adopt the agent model from above but adjust the arrangement of agents and their interactions. Agents are allocated to two groups or communities. In a prior learning stage, agents learn to associate themselves (subject) with a stereotyped group (target) and an \((SC, SI)\) tuple of \((1, 0)\) (content). Let us call this stereotyped group the Jamayans again. Each agent also learns to associate a familiar, and an unfamiliar audience with the Jamayans and an \((SC, SI)\) tuple of \((p, 0)\). The familiar audience represents members of the same community, the unfamiliar audience represents members of the other community. For the unfamiliar audience, partial knowledge is assumed \((p = 0.5\) as above), while complete knowledge of the stereotype is assumed for the familiar audience \((p = 1)\). The number of iterations or times these three stereotype patterns are learnt for each agent corresponds to the value of \(k\) found for the best performing parameter from parameter estimation and validation. In fact, all parameter values for the agent model are the same as the ones from calibration and validation unless specified otherwise in the following.

In every iteration of the simulation, an interaction occurs. With probability \(q\), this is a first-hand experience for a randomly selected member from the first community. With probability \(r\), this is a second-hand information exchange within either community 1 or community 2 between two randomly selected members from that community. With probability \(s\), the interaction is a second-hand information exchange between two randomly selected members.
5.7. The Influence of Network Configuration on Stereotype Communication

from each community. We have \( q + r + s = 1 \). We describe these different interaction types in the following.

During a first-hand experience, a member of the first community receives an \((SC, SI)\) tuple of \((1, 1)\), reflecting that this experience with a member of the Jamayans provides both SC and SI information. Then, the agent learns to associate itself (subject) with this member of the Jamayans (target) and the received \((SC, SI)\) tuple (content). The pattern representing the specific member of the Jamayans is constructed from the Jamayans pattern specific to this agent where a proportion of \(1 - o\) entries are reset to random values. Recall that \(o\) above has denoted the overlap between the target pattern representing the Jamayans and the target pattern representing Jay. Other than that, receiving this information proceeds as described for the Receive procedure above.

During a second-hand information exchange, both agents first produce an \((SC, SI)\) tuple, which is then received by the other agent. Production requires the construction of a pattern representing a concrete member of the Jamayans considering overlap \(o\) as described above. Then, the SC and SI information associated with the agent itself and that particular Jamayan is recalled. Likewise, the SC and SI information associated with a familiar or unfamiliar audience (depending on community membership of the communication partner) and that particular Jamayan is recalled. Subsequently, the information to communicate is obtained as in the Send procedure described earlier. Receiving information proceeds as described in the previous paragraph. However, the number of times the pattern is learnt is now \(k/10\) instead of \(2/3k\). This reflects that a typical second-hand information exchange will carry less information than the detailed story about Jay exchanged in the experiments of Lyons and Kashima. Note that agents do not adjust their information regarding the content produced by other agents, i.e. they do not learn about the agents they communicate with. We leave this option for future work.

For each second-hand information exchange we record the value of SC and SI information communicated by both agents. This allows us to track the extent to which SC and SI information is communicated by members of each community during all simulated iterations.

5.7.2 Results

Figure 5.14 shows the communicated SC and SI information between two communities of 10 agents each for varying levels of \(q\), \(r\), and \(s\) over 500 iterations. Each line is constructed by a local polynomial regression from individual data points at each iteration of 15 simulation runs. Confidence intervals with level 0.95 are plotted as well but hardly visible because they are generally tiny.

First, recall that \(q\) is the probability that an interaction is a first-hand experience, \(r\) the probability that it is a second-hand intra-community information exchange, and \(s\) the
probability that it is a second-hand inter-community information exchange. Because we have \( q + r + s = 1 \), \( s \) can be inferred from \( q \) and \( r \) in each cell of the grid of graphs. Now let us look at this figure bit by bit.

For \( q = 0 \), there are no first-hand experiences and hence no SI information is introduced to the population from the outside. Accordingly, any information exchange transmits complete SC but no SI information as can be seen in the first row of the figure. Now consider the last graph of each row, in which \( s = 0 \) and hence no second-hand inter-community information exchange. As expected, information exchange within the second community, which does not receive any first-hand information, transmits complete SC and no SI information. Agents from the first community, in contrast, eventually start transmitting as much SI as SC.
5.7. The Influence of Network Configuration on Stereotype Communication

information. This is to be expected because each member of that community is repeatedly confronted with first-hand experience that exhibits to the same extent SC and SI information.

Let us consider now all graphs for which \( q > 0 \) and \( q + r < 1 \). In these cases, all interaction types occur. We note immediately that for both communities in all these graphs there is a decrease in communicated SC information and an increase in SI information over time. Hence, first-hand information introduced to the population shows an effect in both communities, namely a decrease of the SC bias. The increase in the communication of SI information obviously follows from the introduction of this information into the population by first-hand experiences. The decrease in SC information can be assumed to follow from the representation of Grice’s maxim of quantity as we argue in the next paragraph.

In any column, the decrease in SC information over time is less pronounced for larger values of \( q \), i.e. for more first-hand experiences and less second-hand inter-community exchanges. A reasonable explanation is that first-hand experiences reintroduce larger values of SC to the first community which compete with second-hand information exchanges that cause a decrease in communicated SC information over time. This indicates that the mechanism of communication between two agents is responsible for the decrease in transmission of SC information over time. The main differences between first-hand experiences and second-hand exchanges are memory recall by the transmitting agent and the representation of Grice’s maxims. In fact, we can assume that the maxim of quantity is the only component that makes a difference because memory recall and the maxim of quality do not lead to any decrease in SC information in the case of \( q = 0 \). In the case of \( q > 0 \), however, the maxim of quantity comes into play by reducing the communication of SC information. This is to be expected because according to Algorithm 5.3 less SC information is transmitted by an agent the more SI information this agent is able to recall.

With increasing \( r \) and therefore decreasing \( s \) in any row, the proportion of intra-community information exchanges increases at the cost of inter-community exchanges. For all levels of \( q \) this causes the SI information transmitted by members of the first community to become increasingly larger than the one transmitted by members of the second community. This is expected because a decrease in second-hand information exchange hinders the transmission of this SI information from the first to the second community.

Now consider again an increase of \( q \), i.e. an increase in first-hand experiences and a decrease in second-hand inter-community exchanges within the columns of the figure. The SI information transmitted by members of both communities increases with increasing \( q \). This is expected because an increase in \( q \) corresponds to more first-hand experiences that introduce more SI information to the population. Likewise, the gap between SI information transmitted by members of the first community and by members of the second community increases with increasing \( q \). This is also expected because an increase in \( q \) (the rate of first-hand
experiences) corresponds to a decrease in \( s \) (the rate of inter-community exchanges). Hence, the second community is less affected by SI information introduced to the first community.

For all graphs with \( q > 0 \) and \( q + r < 1 \), more SC (SI) information is transmitted by members of the first community than SC (SI) information by members of the second community. This is to be expected because the first community is confronted with new information with the highest possible value of SC and SI while the second community is not.

Let us examine the support for our hypothesis that the assumption of partial knowledge of stereotypes between communities hinders the diffusion of SI information from one community to the other. We do observe that for \( q > 0 \) and \( q + r < 1 \) more SI information is transmitted by members of the first community than by members of the second, even in the long run. This holds even for high values of \( s \), i.e. a large proportion of inter-community second-hand information exchanges. Thus, our hypothesis is confirmed.

The conclusion we can draw from these results is that the model behaves as expected as far as our analysis goes and also confirms our previously formulated hypothesis. In effect, we have followed the approach of pattern-based modeling to compare the output of our model to multiple expected patterns, first the validation data set and then the qualitative patterns we would expect to observe in a larger population. Pattern-based modeling was briefly mentioned in Section 3.2.2.

## 5.8 Discussion and Conclusions

We have presented an agent-based model for the investigation of the relationship between actual and perceived sharedness of stereotypes and the communication of stereotype-relevant information. Our model has three key features: a) Agents hold a representation not only of the information they (subject) associate with other individuals or groups (target) but also of the information others (subject) associate with third parties (target); b) The representations of subjects and targets allow overlap between different individuals or groups so that learning or recalling information about one individual always happens in light of the information stored about similar others; c) the communication of stereotype-relevant information is affected by the extent to which it is perceived to be shared. We have identified these three features as crucial for models of stereotype communication and their integration distinguishes our model from previous work.

In creating this model, we have relied on empirical work on the communication of stereotype-relevant information by Lyons and Kashima (2003) and on the premises and implications of the grounding model of cultural transmission. Demonstrating how this social scientific knowledge can be translated into a computational model is this chapter’s contribution to answering RQ1. The model assumes a rich representation of the mechanisms
5.8. Discussion and Conclusions

internal to agents and also facilitates larger scale simulations. Hence, this model covers an area on the nano-to-macro axis stretching from the nano to the macro level.

Our agent model relies on a connectionist architecture as a representation of the agents’ memory. A simple communication mechanism allows the exchange of stereotype-relevant information in terms of the degree to which this information is stereotype-consistent or -inconsistent. We have shown our work to stand in the tradition of other approaches to the modeling of stereotyping processes in psychological literature. We have not drawn on any approaches to inter-agent communication between goal-directed agents, such as speech acts as rational action (Cohen and Levesque, 1990) or argumentation dialogues (Bench-Capon and Dunne, 2007). The reason for this is that the information exchange our model represents is one-way and does not involve any complex form of dialogue, and that a complex communication protocol would impede the scaling of this model to larger agent populations.

We employed an evolutionary algorithm to find the set of parameters that yield the best match between simulated and empirical data. We found the match between empirical data and the data generated by our model to be quantitatively close in terms of correlation, which suggests that it is worthwhile to invest further effort in the analysis of this model. A drawback of evolutionary algorithms is that only a point estimate for the “true” parameter is found. While we have provided plots of the distribution of good parameter values, Bayesian approaches would provide a theoretically more rigorous analysis of the uncertainty of parameter estimates (Beaumont, 2010). After parameter estimation, we validated the calibrated model against another, previously unseen, data set. The match between the data produced by the calibrated model and this new data set was shown to be close qualitatively. We have also conducted a local sensitivity analysis for parameter changes around the estimated parameters. We found that the model is more sensitive to parameters relating to the connectionist architecture than to other parameters.

This work enables the development of agent-based models of stereotype communication in which agents learn about each other or third parties not only first- but also second-hand by communication. Such a model lends itself to the investigation of how stereotype-relevant information diffuses through network structures with different composition and configuration. Through the representation of the maxim of quantity, the common ground associated with a social link is accounted for in the diffusion process. Due to the integrated storage of information about individual/group subjects and targets, the group memberships of the communicating agents and the third party communicated about are accounted for. This

\[^{4}\text{We note that a comparison between the performance of our model and the one of Van Overwalle and Heylighen (2006) in reproducing the experiments of Lyons and Kashima (2003) is not meaningful because their experimental setup does not match the one of Lyons and Kashima appropriately as discussed in Section 5.2.}\]
amounts to a consideration of network configuration in terms of the group memberships of
the interacting agents and of the third party communicated about.

The model allows, for example, to investigate under which network configurations
first-hand experience could help overcome the stereotype-maintaining force of second-hand
information exchange. We have employed the model after its parameters were estimated
to shed light on one such question. Our hypothesis was that the assumption of partial
knowledge of stereotypes by unfamiliar others hinders the diffusion of stereotype-inconsistent
information along network ties between different communities. The rationale behind this is
that interactions along these weak links are likely to be between unfamiliar coactors. We
have indeed observed such an effect. We are not aware of any existing studies investigating
the effect of network configuration beyond a chain of agents as considered by Lyons and

Our observation that the model generally behaves as expected supports the conclusions
drawn by Lyons and Kashima (2003) regarding the role of common ground and of Grice’s
maxims of quality and quantity in the transmission of stereotype-relevant information.
Thereby, our results provide support for the mechanisms postulated by the grounding model
of cultural transmission. This is the contribution of this chapter to RQ2. We have to remark
that the empirical data available was sparse, especially the one for validation. More data
is required to increase the confidence we can have in this model. We have refrained from
extensive large scale simulations. However, we have demonstrated that by and large the
model behaves as expected when scaled up to a larger agent population. The computational
complexity of interactions in the model presented here is larger than the one of the model
introduced in the previous chapter. This is because the model here is more faithful to the
grounding model of cultural transmission.
Towards a Formal Model of the Grounding Model of Cultural Transmission

The previous two chapters have built on some of the premises and implications of the grounding model of cultural transmission to create executable simulation models of cultural dynamics. Those models include only simple representations of social interactions but cover a large area on the nano-to-macro axis. Now this thesis turns towards models that take more seriously the role of complex social interactions in cultural transmission by being more faithful to Kashima et al.’s (2008) grounding model of cultural transmission. Hence, we shift our investigations towards models that are located at the nano and micro level only. While the models introduced in the remainder of this thesis are valuable assets in their own right, they can also fulfill the role of auxiliary models as discussed in Chapter 3.

As announced in Section 3.4, we address in this chapter the question of what is an appropriate more formal description of the grounding model of cultural transmission. In addressing this question, we work towards a formal logics account of the grounding model of cultural transmission. In particular, we present a computational-conceptual model (terminology introduced in Section 3.2.2) that represents in more detail the interactions between the three main ingredients of the grounding model of cultural transmission: joint actions, common ground, and the grounding process. Figure 6.1 represents these interactions. Existing common ground provides the epistemic background for joint actions. Joint actions, in turn, imply epistemic goals, which require common ground to be amended through grounding processes. Existing common ground as well as relational goals affect the grounding process.
We focus in this chapter on the analysis of common ground as the central piece of this puzzle. In doing so, we provide a modal logic of common ground, which we call Common Ground Logic. This model pays particular attention to the different types of common ground (context-specific, personal, and communal), to the bottom-up emergence of common ground from the private mental attitudes of agents, and to the top-down determination of the composition of the coactors’ context-specific, personal, and communal common ground in an interaction. We term this composition salient common ground, which is, in fact, what Stalnaker and Kashima et al. simply call “common ground” (more detail provided in Section 6.1.1). We feel that a more distinct term helps our discussion. The bottom-up emergence of communal common ground corresponds to the emergence of culture. The top-down determination of salient common ground represents the way in which common ground and hence culture impacts on cultural transmission in social interactions.

Following from that we sketch how the grounding process could be represented formally and how it could be integrated into a formal model of joint activities. For the representation of joint activities, we build on the computational SharedPlan formalism of joint activities (Grosz and Kraus, 1996). This model stands in the tradition of the BDI folk-psychological model of practical reasoning (Bratman, 1987), which was briefly mentioned in Section 3.2.2. The grounding process, in turn, is represented as an argumentation dialogue, enabling agents to advance their joint activity by agreeing on information that is compatible with their existing beliefs, intentions, and common ground. While we do not provide a complete dialogue protocol, we do present a framework for further elaboration and discussion. On the nano-to-macro axis, the model presented here is located at the nano and micro level.
This chapter contributes to RQ1 by providing a more comprehensive account of the grounding model of cultural transmission. RQ2 is addressed by contributing to the refinement of the grounding model of cultural transmission. For example, the notion of common ground is discussed more extensively. In addition, we suggest new questions that our model evokes. This chapter also provides the basis for some of the discussion in the next two chapters in that the mechanisms of practical reasoning proposed by Bratman (1987) and others are discussed.

Section 6.1 presents the Common Ground Logic. Section 6.2 describes the Shared-Plan formalism and its interface with Common Ground Logic. Section 6.3 provides a computational-conceptual account of the grounding process and explains how grounding is evoked and affected by the joint activity and how it interacts with common ground. The chapter closes with a discussion in Section 6.4.

6.1 Common Ground

In this section, we develop a modal logics account of common ground. First, however, we need to elaborate on the discussion in Section 2.2 about the role of common ground in the grounding model of cultural transmission. Recall the different parts of common ground we discussed there: Context-specific, personal, and communal common ground on the one hand, actual and perceived common ground on the other hand. Salient common ground in the following is that composition of context-specific, personal, and communal common ground that yields the common ground for the participants of a joint action. Note that the composition operator is not necessarily the union operator. The union of context-specific, personal, and communal common ground might be inconsistent and clearly context-specific common ground can override personal or communal common ground in any given context. Also, context-specific common ground is not equivalent to salient common ground. Context-specific common ground is manipulated by grounding processes but salient common ground is what evolves through this manipulation, in composition with personal and communal common ground.

Recall from our discussion in Section 2.2 that common ground is indexed by location, time, and participants. We assume now that these indices can be subsumed by an index for context, whereby a context is either a joint action or a group. This index determines the context in which common ground is valid. A joint action context yields a representation of context-specific common ground, which is valid only within the current context/joint action. A group context yields a representation of the personal or communal common ground of a group or community, which is valid for that group beyond any particular joint action.
context.\(^1\) In that, we do not treat personal and communal common ground differently from a technical perspective. The personal common ground of a small group of people is simply their “communal” common ground, even though strictly speaking personal common ground is not held by a community.

We first discuss philosophical conceptions of epistemic states relevant to the analysis of common ground. We then advance our own understanding and present and analyze a modal logic of common ground. We close this section with a discussion of our model in relation to other attempts at formalizing common ground or related concepts.

### 6.1.1 An Analysis of Common Ground

From earlier discussions in Section 2.2 it follows that the communal common ground of a community and therefore its culture affects individual interactions (top-down). At the same time, individual interactions can give rise to communal common ground and therefore to culture (bottom-up). Both of these mechanisms are crucial to the process of cultural dynamics and therefore need to be represented by our model. The top-down influence of communal common ground is encoded in the composition of salient common ground. This composition determines how personal common ground and different communal common grounds become relevant in interactions. We assume that in a given context only certain group memberships are salient, and therefore only a subset of available communal common grounds are salient. For example, two Texan-Americans could interact with their Texan group membership or their American group membership being salient, or both for that matter. This is in line with the broader ideas of social identity theory (Tajfel and Turner, 1979) and self-categorization theory (Turner, 1982), according to which different social identities are salient during any social interaction (Jenkins, 2008).

The mechanism by which common ground changes reflects the bottom-up mechanism of the emergence of culture in general and of communal common ground in particular. As stated in Chapter 2, we see context-specific, personal, and communal common ground and in fact also salient common ground as a state of collective acceptance. This claim deserves further analysis and in particular a discussion of collective or group beliefs more generally and a discussion of the relationship between common ground and group beliefs.\(^2\)

---

1Recall that in Section 2.2 we have used the term “communal common ground” also to denote that part of common ground which is more generally due to common community memberships of the participants of a joint action. Here, instead, we use the term to denote that information that is taken for granted by the members of a particular community.

2More elaborate discussions on this topic can be found in Hakli (2006) and in Gaudou (2008, Chapter 2).
6.1. Common Ground

Conceptualizations of group belief can be broadly categorized into summative and non-summative accounts (Hakli, 2006). According to summative accounts, if a group believes a proposition, it is logically necessary that all or most group members believe that proposition (Gilbert, 1987). Note that an account that considers group beliefs reducible to individual mental attitudes other than belief is not necessarily a summative account in this sense.

A common summative account is given by Gilbert (1987) based on Quinton (1975): A group is said to have a belief if and only if most of its members have this belief. However, this statement is rather weak. Members of such a group would not be able to act based on the premise that other members of the group have the same belief because there is no awareness of the group belief (Gilbert, 1987). This observation has led to stronger accounts, in particular the one of mutual belief or common belief (see for example Schiffer, 1972, and Lewis, 1969, for a corresponding account of common/mutual knowledge). Schiffer’s now widely accepted account of mutual belief can be stated as follows (Stalnaker, 2002):

**Definition 6.1** (Mutual belief according to Schiffer, 1972). *A proposition φ is a mutual belief of a group if and only if all group members believe that φ, and all of them believe that all believe that φ, and all believe that all believe that all believe that φ, and so on ad infinitum.*

One of the main problems of summative accounts with respect to our discussion, however, is that a summative group belief implies corresponding individual beliefs (Gilbert, 1987). Yet it is frequently (though not always) the case that individual beliefs deviate from group beliefs. Take as an obvious example a government (the group of officials/ministers governing a state): The official position of a government rarely corresponds to the individual beliefs of its members. According to non-summative accounts, group beliefs are independent of the beliefs of group members. A lively debate between proponents of non-summative accounts is concerned with the question whether group beliefs are in fact a type of belief or whether they are a type of acceptance. Proponents of the first claim are commonly called “believers” and the proponents of the second claim are commonly called “rejectionists” (Hakli, 2006).

The most prominent “believer” is certainly Margaret Gilbert. She initially advanced the following definition of non-summative group beliefs (1987):

**Definition 6.2** (Non-summative group belief according to Gilbert, 1987).

1. A group G believes that φ if and only if the members of G jointly accept that φ.

2. Members of a group G jointly accept that φ if it is common knowledge in G that the individual members of G openly expressed a conditional commitment jointly to accept that φ together with the other members of G.

---

3Other problems with summative accounts of group belief are offered by Gilbert (1987); Meijers (2002).
In Gilbert’s definition group belief is not reduced to individual beliefs. In fact, a group belief comes about only through the public behavior of group members in that group belief requires their openly expressed commitment. According to Gilbert, joint acceptance can but does not necessarily involve a verbal agreement. Instead, joint acceptance can also come about by the communication of nonverbal signals (Gilbert, 1987). Gilbert later (1996) concedes that she did not intend to introduce a concept different from belief when she used the term “to accept” (and there is no relationship between this notion of acceptance and the one we mentioned in Section 2.2.1 and discuss later in this section). This has led her to revise her original definition:

**Definition 6.3** (Non-summative group belief according to Gilbert, 1994). *A and B form a plural subject of believing that φ if and only if A and B are jointly committed to believing φ as a body.*

This definition adds the condition of a joint commitment, which in Gilbert’s understanding is not reducible to individual commitments (2009). Among the consequences of the commitment to a group belief that φ is the obligation on group members to act as if φ and not to state that not φ while acting in the role of a group member. Elsewhere, Gilbert (1996) states that she considers her various descriptions of group belief to refer to the same account.

Among those that agree with Gilbert that group belief is generally a non-summative concern but disagree that groups can hold beliefs in the same way that individuals can are Meijers (2002), Tuomela (2000a, 2003), and Wray (2001). Instead, these authors argue that group beliefs are in fact instances of collective acceptance. According to Hakli (2006), an individual’s acceptance that p is “typically taken to be a kind of a mental act, a decision to treat p as true in one’s utterances and actions, or an act of adopting a policy to use p as a premiss in one’s theoretical and practical reasoning”. The following differences between individual acceptances and beliefs have been proclaimed, and different authors typically subscribe to different subsets of these differences (Hakli, 2006):

1. Beliefs are involuntary and acceptances are voluntary.
2. Beliefs aim at truth and acceptances depend on goals (they aim at utility or success).
3. Beliefs are shaped by evidence and acceptances need not be.
4. Beliefs are independent of context and acceptances are context-dependent.
5. Beliefs come in degrees and acceptances are categorical.

From our perspective, goal- and context-dependence are the most important distinguishing features of acceptances. This claim will become more obvious later when we relate common
ground to acceptance. However, refer to Hakli (2006) for an argument that voluntariness is the only feature that is valid to distinguish belief and acceptance.

Tuomela’s account is representative of the rejectionist non-summative perspective on group belief. Tuomela (2000a) distinguishes between pragmatic acceptances of something as good or satisfactory in a context and aiming at utility, and acceptances as true or correctly assertable, which are truth-oriented and not bound to any particular context.

On Tuomela’s analysis, a collective acceptance can be in I-mode (private acceptance) or we-mode (public acceptance as a group member). His definition of I-mode collective acceptance can be phrased as follows:

**Definition 6.4** (I-mode collective acceptance according to Tuomela, 2003). *There is a collective acceptance of a proposition* $\phi$ *in a group if and only if all members of that group accept* $\phi$ *and there is mutual belief that all accept that* $\phi$.\(^4\)

In the we-mode case, the collective acceptance in addition implies a collective commitment to the accepted proposition. In the we-mode case and exceptionally also in the I-mode case, the collective acceptance of a proposition is *for the group*, i.e. the proposition is correctly assertable for the group. We-mode collective acceptances can only be pragmatic acceptances (with acceptance as true or correctly assertable for the group) while I-mode collective acceptances can also be acceptances as true (Tuomela, 2000a).

Acceptance of a proposition for the group entails that this proposition is correctly assertable for group members when they act in group contexts in their capacity as group members (Tuomela, 2003). Likewise, when a proposition is correctly assertable for a group, the proposition is collectively accepted for the group. When a proposition is correctly assertable for the group, it can be used as a premise in any kind of reasoning within group contexts. In case of collective commitment of a proposition, group members are mutually committed to each other to act as if this proposition holds and they are mutually committed to uphold this proposition. The notion of group-contexts suggests that collective acceptances in Tuomela’s understanding can be context-dependent.

The creation of a collective acceptance would require a collective performative speech act (Tuomela, 2003). However, Tuomela admits that this is not practical in real situations. In these cases it suffices that group members think and act in a way as if this collective speech act had been performed. Likewise it suffices that not all but most of the members of the group hold the acceptance and belief attitudes required to establish collective acceptance.

We side with Tuomela in the claim that group beliefs are often (though not always) more suitably understood as collective acceptances. The characteristics of collective acceptance

---

\(^4\)Note that Tuomela (2003) also requires that all group members believe that everyone accepts that proposition. However, this sentence is already implied by mutual belief when once accepts Schiffer’s account of mutual belief (Definition 6.1).
that are of primary interest here are that (i) collective acceptances can be goal- and context-dependent (pragmatic acceptances), (ii) collective acceptances do not imply corresponding private beliefs, and (iii) collective acceptances can imply collective commitments.

**Common Ground**

Having presented an overview on group beliefs and collective acceptances, we turn towards common ground. As indicated in Chapter 2, there is no agreement as to a single definition of common ground. In the following, we consider the most prominent ones.

Clark’s (1996c, p. 93) definition of common ground can be phrased as follows:

**Definition 6.5 (Common Ground according to Clark (1996c)).** A proposition, φ, is common ground for members of a group if and only if: (i) the members of the group have information that φ and information that (i).

The term “have information” includes various epistemic attitudes such as knowledge, belief, or assumptions (Clark, 1996c). Hence, it is not logically necessary that if a proposition is common ground that this proposition is believed by the involved agents. Therefore, common ground as per Clark is not a summative epistemic group attitude in Gilbert’s sense. However, Clark sees common ground to be reducible to individual mental attitudes. As we will see later, Kashima et al. (2008) consider the first part of the definition above to denote actual common ground and the second part to denote perceived common ground. Clark (1996c, p. 121) defines personal and communal common ground as follows:

The common ground between two people divides into two broad types. Communal common ground is information based on the cultural communities a person is believed to belong to – from nationality and occupation to ethnic group and gender. Personal common ground is information based on personal acquaintance: It is lacking in strangers and greatest for intimates.

It is expedient to drill into Clark’s understanding of personal and communal common ground. Clark (1996b) defines the personal common ground between two people as “all the mutual knowledge, beliefs, and assumptions they have inferred from personal experience with each other”. The communal common ground between two people is defined by Clark (1996b) as “all the knowledge, beliefs, and assumptions they take to be universally held in the communities to which they mutually believe they both belong”. The mutual belief of a community membership unlocks that information which coactors “take to be universally held” in that community. It appears that the information taken to be universally held within a community is the communal common ground of this community.
Stalnaker (2002) emphasizes that what is added to common ground during a conversation can deviate from the individual beliefs of the participants of the conversation. Therefore he opts for an account based on the concept of acceptance, which can be phrased as follows under the assumption of Schiffer’s account of mutual belief (see Definition 6.1):

**Definition 6.6 (Common ground according to Stalnaker (2002)).** There is common ground of a proposition $\phi$ in a group if all members accept (for the purpose of the conversation) that $\phi$ and there is mutual belief that all accept that $\phi$.

At first glance, this definition of common ground is equivalent to the essential part of Tuomela’s definition of collective acceptance, namely I-mode collective acceptance (Definition 6.4). On closer inspection, there are three differences: First, Stalnaker considers only the common ground in the context of a particular conversation, i.e. context-specific common ground. Second, he uses the conjunction “if” instead of “if and only if”, implying that common ground of $\phi$ does not logically entail acceptance and mutual belief thereof. However, his later discussion (Stalnaker, 2002, p. 717) indicates that he actually meant “if and only if”. Third, both authors have a different idea of the notion of acceptance. Acceptance is understood by Stalnaker as a broader class of epistemic attitudes that includes, amongst others, beliefs, presumptions, and assumptions. In contrast, Tuomela considers acceptance to be independent of belief. In the we-mode case of collective acceptance, moreover, a collective commitment to the accepted proposition is entailed. This clearly goes beyond Stalnaker’s understanding of common ground.

According to Stalnaker, acceptance of a proposition means to treat that proposition as true for a particular reason and in the simplest case this reason is just that one believes that proposition to hold. Acceptance can be limited to a particular context but does not have to be. Common ground as per Stalnaker is a non-summative epistemic attitude because a proposition being common ground does not necessarily entail that all coactors believe that proposition, which is according to Gilbert (1987) the distinguishing feature of summative group attitudes. However, Stalnaker’s common ground is reduced to individual mental attitudes. Stalnaker does not distinguish between personal and communal common ground or between actual and perceived common ground.

Lee (2001) argues based on an analysis of empirical dialogue data that any kind of summative group belief that requires an infinite regression, such as mutual belief or knowledge, is a psychologically implausible conception of common ground. He conceptualizes context-specific common ground (what he calls established common ground) as shared belief/knowledge. Shared belief is defined as a truncated version, i.e. limited recursion, of mutual belief/knowledge in Schiffer’s sense. Obviously, this implies a summative conceptualization of common ground because individual beliefs are implied by common ground.
Lee identifies assumed common ground as that information which is assumed to be shared due to common community memberships. Assumed common ground consists of background/common knowledge/belief. Hence assumed common ground corresponds to perceived communal common ground. However, Lee does not specify in more detail what the epistemic status of assumed common ground is. He also does not distinguish between actual and perceived common ground and between context-specific, personal, and communal common ground as carefully as our analysis requires.

Kashima et al.’s (2008) definition of common ground, which is based on Clark’s definition (Definition 6.5), can be phrased as follows:

**Definition 6.7** (Common ground according to Kashima et al. (2008)). A proposition $\phi$ is common ground for members of a collective if and only if: (i) the members of the collective have information that $\phi$ is true and information that (i).

The first part of (i) can be said to represent actual common ground, while the second part adds the condition for this to also be perceived common ground (Kashima et al., 2008). The collective in this case can be of any size from a dyad to a larger social group defined, for instance, by shared gender, profession, or religion. Hence, this definition is applied to personal common ground as it is to the communal common ground of a community or any other collective. From the discussion of Kashima et al. (2008) we infer that the same definition applies to context-specific common ground.

Common ground as defined by Kashima et al. is a non-summative attitude because common ground does not necessarily imply the same individual beliefs. Thus it avoids one of the main problems of summative group beliefs. However, as in Clark’s definition common ground is reducible to individual attitudes. The term “have information that” in the above definition stands for a host of attitudes such as belief, assumption, and supposition (Kashima et al., 2008).

From this discussion we extract as the main features of common ground that (i) common ground can be context-dependent and that (ii) the common ground of a group can differ from individual beliefs.

**Synthesis**

In light of the above discussion, we advance our own understanding of common ground.

Considering the overlap between the features of collective acceptance and common ground, we conceptualize all kinds of common ground as collective acceptance: Both collective acceptance and common ground can be context-dependent and be motivated by pragmatic and epistemic reasons. Both collective acceptance and common ground can deviate
from individual beliefs. Moreover, we suggest that not only collective acceptance but also common ground can imply a collective commitment. We comment on the last point below.

We follow in particular Tuomela’s understanding of acceptance as acceptance of something as true or correctly assertable for the group. The most essential part of collective acceptance is then covered by collective acceptance in the I-mode (without collective commitment), presented in Definition 6.4. This is what we seek to model in the following. I-mode collective acceptance is typical of unstructured groups or indeed collectives rather than groups, which we are primarily interested in as those groups that hold common ground. However, we do acknowledge that we-mode collective acceptance with collective commitment also occurs in the case of unstructured groups and also that structured groups can hold common ground. Nevertheless, we focus on the essential bits of collective acceptance here.

Let us analyze these observations with respect to the different types of common ground that the grounding model of cultural transmission considers.

Context-specific common ground is obviously bound to the context of the joint activity to which it belongs. An obvious way in which context-specific common ground can deviate from individual beliefs is when coactors accept a certain proposition just for the purpose of advancing their joint activity. In that, context-specific common ground can clearly be pragmatically motivated. Moreover, the acceptance of a proposition can entail that the coactors are committed to this proposition because in some way or the other they have mutually agreed to go with this proposition. In general, this agreement is a consequence of the grounding process. Hence the grounding process is the mechanism by which context-specific common ground emerges.

Personal and communal common ground are bound to the contexts of groups. In particular, personal and communal common ground are held by associated groups, and they are only available in contexts in which these respective group memberships are salient for those actors involved in the joint activity. Obviously, what is common ground of a group does not imply anything about the beliefs of the group’s members. And finally, the common ground of a group can imply that group members are collectively committed to the propositions implied by this group’s common ground when the relevant group membership is salient. Consider Tuomela’s favorite example (Tuomela, 2003)—the consideration of squirrel fur as money in medieval Finland. Obviously, for this fiscal system to work out it was not sufficient that the associated set of propositions was part of common ground but there also needed to be a collective commitment to treating squirrel fur as money.

We assume there is no technical difference between personal common ground and the communal common ground of groups since both consist of information. However, we do acknowledge that personal common ground consists to a large part of information about common experiences. Communal common ground lacks such information. We require an
appropriate mechanism that specifies how personal and communal common ground emerge. Again, this mechanism consists of the grounding process by which individual mental attitudes are changed and of a specification of how collective acceptance follows from the group members’ individual attitudes.

Understanding context-specific, personal, and communal common ground as collective acceptances, we assume that salient common ground shares crucial features with the former three concepts.

Briefly, we have identified the following requirements for our account of common ground:

- Common ground is best described as a type of I-mode collective acceptance (Definition 6.4, Tuomela, 2003).

- Common ground can be context-dependent and motivated by pragmatic and epistemic reasons. A context can be a joint action or group context. The first case yields context-specific common ground, the second case yields personal or communal common ground.

- Salient common ground is a composition of context-specific, personal, and communal common ground.

- Common ground does not imply corresponding individual beliefs.

- There is no technical difference between personal and communal common ground.

- Salient common ground is also a type of collective acceptance.

6.1.2 Common Ground Logic

In this section, we propose a modal logic of common ground that respects the results of our synthesis above.

Modal logics are a mathematical modeling tool that allows truth-functional sentences to be augmented by modalities in the form of non-truth-functional modal operators. Modal operators include epistemic operators such as “I know” or “I believe” that allow the expression of an agent’s epistemic state with respect to a particular sentence. Because modal operators are not truth-functional, the truth value of a sentence consisting of a modal operator and its argument is not depending on the truth-value of the argument alone. This allows, for example, for the simultaneous truth of the following two sentences: (1) “I believe that the earth is flat”, (2) “I believe that the sky is blue”. Here, the operator is “I believe” and the arguments are “the earth is flat” and “the sky is blue”. Note that the arguments have different truth values. Nevertheless, both composite sentences can have the same truth value because the truth value of a modal operator does not depend on the truth-value of its argument.
6.1. Common Ground

alone. That is, modal operators respect that assertions about epistemic states yield opaque contexts; logically equivalent arguments do not necessarily lead to the same truth value of the compound (Russell and Norvig, 2003).

The semantics of modal logics are usually provided by Kripke semantics, which build on Hintikka’s (1962) concept of possible worlds (Kripke, 1963). In the case of epistemic modal logics, a possible world is one of many epistemic alternatives (Hintikka, 1962). Given the actual world, i.e. the epistemic alternative that is the actual one, binary accessibility relations between this actual world and other worlds determine which epistemic alternatives the agent considers possible. The smaller this set, the more precise the agent’s information. This perspective reflects the contingent nature of epistemic attitudes: An epistemic attitude is evaluated with respect to the epistemic alternatives that are consistent with the agent’s epistemic state.

Because of these characteristics, which are not provided by classical logic, it is common practice in philosophy and computer science to use modal logics and Kripke semantics to represent an agent’s epistemic state (e.g. Cohen and Levesque, 1991; Groz and Kraus, 1996; Stalnaker, 2002). Considering that in our analysis above acceptance and common ground are epistemic states of agents and groups of agents themselves, it is reasonable to express them with modal operators. Accordingly, modal logics of publicly grounded information (Gaudou et al., 2006a) and (collective) acceptance (Lorini et al., 2009) have been proposed. In fact, modal logics are often used to represent other propositional attitudes as well, for example goals or intentions (Shoham and Leyton-Brown, 2008; Wooldridge, 2009). The case of acceptance follows the similar cases of belief or knowledge: An agent can be said to accept a proposition if this proposition is true in all possible worlds (epistemic alternatives) accessible from the current world via the respective accessibility relation. That is, the set of accessible possible worlds represents those epistemic alternatives that are consistent with what is accepted. Based on this idea we formalize in the following common ground as interpreted in the previous subsection. In doing so, we present first the syntax, semantics, and axiomatics of our modal logic. We then prove the validity of our axioms with respect to this logic and analyze its properties.

Syntax

Let \( \mathcal{A} \) be a finite and non-empty set of agents. Let \( \mathcal{C} \) be a finite set of contexts, consisting of elements that are either joint actions or non-empty subsets of \( \mathcal{A} \) (groups) in line with our discussion early on in this section. Let \( \Sigma \) be a set of propositional variables. We define the
modal language $L_{CG}$ (Common Ground Logic) to be the smallest set that is determined by
the following Backus Naur Form:

\[
\phi ::= p | \neg \phi | \phi \land \phi | B_i \phi | EB_c \phi | MB_c \phi | A^i_c \phi | EA_c \phi | CA_c \phi | DCG_\alpha \phi | CG_\alpha \phi
\]

where $p$ is any propositional variable in $\Sigma$, $i$ is any agent in $A$, $c$ is any context in $C$, and $\alpha$ is any joint action. The other usual connectives such as $\lor$, $\rightarrow$, and $\leftrightarrow$ can be defined in terms of $\neg$ and $\land$.

Informally, the intended interpretation of $B_i \phi$ is that agent $i$ believes $\phi$. The intended interpretation of $EB_c \phi$ is that $\phi$ is believed by every agent in the group involved in context $c$, the intended interpretation of $MB_c \phi$ is that $\phi$ is mutual belief for the group of agents involved in context $c$. Up to here, the syntax follows closely existing modal logics of belief and mutual belief (e.g. Fagin et al., 1995). The following non-standard components are required by us in addition.

The intended interpretation of $A^i_c \phi$ is that agent $i$ accepts $\phi$ as true in context $c$ and the intended interpretation of $EA_c \phi$ is that $\phi$ is accepted by every agent involved in context $c$. The intended interpretation of $CA_c \phi$ is that $\phi$ is a collective acceptance for the group of agents involved in context $c$. If the context is a joint action, this denotes context-specific common ground. If the context is a group of agents, this yields personal or communal common ground. The intended interpretation of $DCG_\alpha \phi$ is that $\phi$ is distributed common ground of the agents involved in $\alpha$. The notion of distributed common ground is based on the idea of distributed knowledge (see for example Fagin et al., 1995): A proposition $\phi$ is distributed common ground in the context of a joint action if it can be deduced from the combined information of context-specific common ground and the communal common ground of all salient groups. We elaborate on the notion of distributed common ground in the next subsection. The intended interpretation of $CG_\alpha \phi$ is that $\phi$ is salient common ground in the context of joint action $\alpha$, given that particular groups memberships of the agents involved in $\alpha$ are salient. Table 6.1 summarizes those latter, more intricate, parts of the syntax.

**Semantics**

We define the semantics of Common Ground Logic with an extension of Kripke semantics, which are based on possible worlds.

**Definition 6.8.** An $L_{CG}$ frame is a tuple $\langle A, G, Q, C, X, Y, W, R_B, R_A \rangle$ where

- $A$ is a finite and non-empty set of agents. We denote particular agents by $i$ and $j$.
- $G \subseteq 2^A \setminus \{\emptyset\}$ is a set of non-empty subsets of agents. We denote elements of $G$ by $G$, $H$, and $I$. 


6.1. Common Ground

Table 6.1: Summary of the intended interpretations of operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Intended interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_i^\alpha)</td>
<td>An acceptance of agent (i) in the context of the joint action (\alpha).</td>
</tr>
<tr>
<td>(A_i^G)</td>
<td>An acceptance of agent (i) in the context of group (G).</td>
</tr>
<tr>
<td>(CA_\alpha)</td>
<td>A collective acceptance of the participants in the joint action (\alpha). This denotes context-specific common ground.</td>
</tr>
<tr>
<td>(CA_G)</td>
<td>A collective acceptance of group (G) in the context of that group. This denotes personal or communal common ground.</td>
</tr>
<tr>
<td>(DCG_\alpha)</td>
<td>The distributed common ground of a group of agents in the context of their joint action (\alpha) given particular salient group memberships.</td>
</tr>
<tr>
<td>(CG_\alpha)</td>
<td>The salient common ground of a group of agents in the context of their joint action (\alpha) given particular salient group memberships.</td>
</tr>
</tbody>
</table>

- \(Q\) is a finite set of joint actions. For the purpose of this chapter, a joint action is an indecomposable entity that is associated with a set of agents involved in that joint action. We denote a joint action by \(\alpha\).

- \(C\) is a finite set of contexts with \(C = \mathcal{G} \cup \mathcal{Q}\). We denote a context by \(c\).

- \(X : C \rightarrow \mathcal{G}\) is a function that maps every context to a non-empty subset of agents that are involved in this context. We have \(X(G) = G\) for every group of agents \(G \in \mathcal{G}\).

- \(Y : Q \rightarrow 2^\mathcal{G}\) is a function that maps every joint action context to a set of non-empty subsets of agents that represent the groups whose membership is salient in that joint action. We assume \(X(\alpha) \subseteq G\) for all \(G \in Y(\alpha)\), i.e. only groups of which the participants in the joint action are members of can be salient.

- \(W\) is a non-empty set of possible worlds. We denote a particular world by \(u\), \(v\), or \(w\).

- \(R_B : A \rightarrow 2^{W \times W}\) maps every agent \(i \in A\) to an accessibility relation \(R_B^i\) between possible worlds in \(W\), representing belief. We shall sometimes write \(R_B^i(w)\) to denote the set \(\{v \in W \mid (w,v) \in R_B^i\}\).
• $R_A : A \times C \rightarrow 2^{W \times W}$ maps every agent $i \in A$ and every context $c \in C$ with $i \in X(c)$ to an accessibility relation $R_A^{i,c}$ between possible worlds in $W$, representing acceptance.

We shall sometimes write $R_A^{i,c}(w)$ to denote the set $\{ v \in W \mid (w,v) \in R_A^{i,c} \}$.

**Definition 6.9.** An $L_{CG}$ model $M_{CG}$ is a tuple $\langle A, G, Q, C, X, Y, W, R_B, R_A, \pi \rangle$ where the tuple $\langle A, G, Q, C, X, Y, W, R_B, R_A \rangle$ is an $L_{CG}$ frame and $\pi : W \rightarrow 2^\Sigma$ is a valuation function that maps each possible world to the set of propositional variables that are true in that world.

The satisfiability relation between a model $M_{CG}$, a world $w$, and a sentence $\varphi$ is denoted by $M_{CG}, w \models \varphi$. A sentence $\varphi$ is **true** in a model $M_{CG}$ if and only if $M_{CG}, w \models \varphi$ for all worlds $w$ in $M_{CG}$. A sentence $\varphi$ is **valid**, denoted by $\models_{c_{CG}} \varphi$, if and only if $\varphi$ is true in all models of $L_{CG}$. The following parts of the satisfiability relation are common to existing modal logics of belief and mutual belief (e.g. Fagin et al., 1995):

- $M_{CG}, w \models p$ iff $p \in \pi(w)$.
- $M_{CG}, w \models \neg \varphi$ iff $M_{CG}, w \not\models \varphi$.
- $M_{CG}, w \models \varphi \land \psi$ iff $M_{CG}, w \models \varphi$ and $M_{CG}, w \models \psi$.
- $M_{CG}, w \models B_i \varphi$ iff $\forall v : (w,v) \in R_B^{i,c}$ we have $M_{CG}, v \models \varphi$.
- $M_{CG}, w \models EB_i \varphi$ iff $\forall i \in X(c)$ we have $M_{CG}, w \models B_i \varphi$.
- $M_{CG}, w \models MB_i \varphi$ iff $M_{CG}, w \models EB_i \varphi$ for $k = 1, 2, \ldots$ with $EB_i^0 \varphi \overset{\text{def}}{=} EB_i \varphi$ and $EB_i^{k+1} \varphi \overset{\text{def}}{=} EB_i (EB_i^k \varphi)$.

The following operator $A_i^c$ is a standard modal operator and $EA_c$ denotes those sentences that are accepted by all agents involved in context $c$:

- $M_{CG}, w \models A_i^c \varphi$ iff $\forall v : (w,v) \in R_A^{i,c}$ we have $M_{CG}, v \models \varphi$.
- $M_{CG}, w \models EA_c \varphi$ iff $\forall i \in X(c)$ we have $M_{CG}, w \models A_i^c \varphi$.

We continue by providing a semantic definition of collective acceptance. We do a bit of scratchwork first, beginning with defining $G_B$-reachability as the transitive closure of the belief accessibility relations of a group $G$ of agents.

**Definition 6.10 (G_B-reachability).** Given a model $M_{CG}$ with a set of possible worlds $W$ and a set of agents $A$. The world $w'$ is said to be $G_B$-reachable from a world $w$ for a group of agents $G \subseteq A$ if there is a sequence $(w = w_0, w_1, \ldots, w' = w_n)$ of worlds such that $(w_k, w_{k+1}) \in \bigcup_{j \in G} R_B^j$ for all $0 \leq k < n$. 

146
6.1. Common Ground

Keep in mind that we define $G_B$-reachability here solely in terms of the belief accessibility relations. Let us now define the relation $R_B^G$ as follows:

$$R_B^G = \{(u, v) \mid v \text{ is } G_B\text{-reachable from } u\}$$

That is, $(u, v) \in R_B^G$ if and only if world $v$ is $G_B$-reachable from world $u$ for group $G$. An alternative characterization of mutual belief can then be given as follows (Fagin et al., 1995):

**Lemma 6.1** (Mutual belief). Consider a model $M_{CG}$ as defined above, a particular world $w$, and a context $c \in C$. We have $M_{CG}, w \models MB_c \varphi$ if and only if $M_{CG}, v \models \varphi$ for all worlds $v$ with $(w, v) \in R_B^G$ and $G = X(c)$.

**Proof.** Recall from the semantics that $MB_c \varphi$ iff $EB_c^k \varphi$ for $k = 1, 2, \ldots$, in which $EB_c^1 \varphi \overset{\text{def}}{=} EB_c \varphi$ and $EB_c^{k+1} \overset{\text{def}}{=} EB_c(EB_c^k \varphi)$. From the definition of $R_B^G$, it is clear that $EB_c^k \varphi$ holds if and only if $\varphi$ is reachable in $k$ steps (take the base case of $k = 1$ and perform induction on $k$). For a pair of worlds $u$ and $v$ it holds that $(u, v) \in R_B^G$ if $v$ is reachable from $u$ after any number of steps, which is the same as the relationship between $EB_c \varphi$ and $MB_c \varphi$. From this, and the definition of $EB_c^k \varphi$, the lemma holds.

We now define the following relation that represents collective acceptance:

**Definition 6.11.** We define $R_{CA}^c \subseteq W \times W$ to be a relation such that $(w, u) \in R_{CA}^c$ iff $u \in \bigcup_{i \in X(c)} R_A^c(w) \cup \{u \mid v \in R_B^G(w) \wedge u \in \bigcup_{i \in X(c)} R_A^c(v)\}$. We write $R_{CA}^c(w)$ to denote the set $\{v \in W \mid (w, v) \in R_{CA}^c\}$.

That is, $(w, u) \in R_{CA}^c$ if and only if world $u$ is accessible from world $w$ by any of the acceptance accessibility relations or there is a world $v$ that is $G_B$-reachable from world $w$ and world $u$ is reachable from world $v$ by any of the acceptance accessibility relations.

A collective acceptance of the agents involved in a particular context $c$ is then defined as follows:

- $M_{CG}, w \models CA_c \varphi$ iff $\forall v \in R_{CA}^c(w)$ we have $M_{CG}, v \models \varphi$.

Depending on the type of the context $c$, $CA_c$ denotes context-specific, personal, or communal common ground.

Distributed common ground $DCG_\alpha$ denotes a type of distributed “knowledge” that is defined by the intersection of the sets of possible worlds considered by context-specific common ground and by the communal common ground of the different salient groups: \(^5\)

- $M_{CG}, w \models DCG_\alpha \varphi$ iff $\forall v \in \left[R_{CA}^c(w) \cap \bigcap_{G \in Y(\alpha)} R_B^G(w)\right]$ we have $M_{CG}, v \models \varphi$.

\(^5\)Distributed knowledge in the case of individuals denotes the knowledge an objective third-party would say these individuals have if they combined their knowledge (Fagin et al., 1995).
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

Distributed common ground is thus deduced from the combined information of context-specific common ground and the communal common ground of the different salient groups.

Salient common ground $CG_\alpha$ is based on $DCG_\alpha$ as follows:

• $M_{CG, w} \models CG_\alpha \varphi$ if $M_{CG, w} \models (\neg DCG_\alpha \perp \land DCG_\alpha \varphi) \lor (DCG_\alpha \perp \land CA_\alpha \varphi)$.

Hence, salient common ground arises from combining the information collectively accepted in the context of the joint action and of the different salient groups. This represents the composition of context-specific, personal, and communal common ground (recall that personal common ground is treated as a special case of communal common ground).

However, as identified previously, the different collective acceptances do not contribute to salient common ground with equal priority. We suggested that context-specific common ground in particular would be prioritized over personal and communal common ground. This is represented by $CG_\alpha$. Only if $DCG_\alpha$ is consistent ($\neg DCG_\alpha \perp$), i.e. the set of worlds $R_{CA}^\alpha (w) \cap \bigcap_{G \in Y(\alpha)} R_{CA}^G (w)$ is not empty, salient common ground $CG_\alpha$ consists of distributed common ground. Otherwise ($DCG_\alpha \perp$) salient common ground falls back to context-specific common ground ($CA_\alpha \varphi$). In other words, personal and communal common ground as combined in distributed common ground are only taken into consideration when they are consistent among each other and with context-specific common ground.

We require the following restrictions on accessibility relations:

(S1) $R_B^i$ is serial, transitive, and Euclidean for all $i \in A$.

(S2) $R_A^{i,e}$ is serial, transitive, and Euclidean for all $e \in C$ and $i \in X(e)$.

S1 and S2 ensure that beliefs are consistent and that, when an agent believes/accepts a proposition or does not believe/accept a proposition, the agent also believes/accepts this to be the case. Our belief and acceptance operators are therefore in a standard modal logic of belief $KD45$ (Fagin et al., 1995). Hence, we follow Stalnaker (2002) in assuming that the logic of acceptance is the same as the logic of belief.

In addition, we require the following restrictions on accessibility relations:

(S3) If $(u,v) \in R_B^i$ and $(v,w) \in R_A^{i,e}$ then $(u,w) \in R_A^{i,e}$ for all $i \in A$.

(S4) If $(u,w) \in R_A^{i,e}$ and $(u,v) \in R_B^i$ then $(v,w) \in R_A^{i,e}$ for all $e \in C$ and $i \in X(e)$.

The semantic constraint S3 ensures that when an agent accepts a proposition, the agent also believes that it accepts this proposition. S4 ensures that when an agent does not accept a proposition, it believes that it does not.

We demonstrate in Figure 6.2 with a simple example how Kripke semantics define the truth values for the sentences in $L_{CG}$. The model in this figure consists of the following components (excluding the accessibility relations and the valuation function): $A = \{1, 2\}, G =$
6.1. Common Ground

Figure 6.2: An example of a Common Ground Logic model. On this figure, a label $A_i^c$ on an arrow from world $u$ to world $v$ denotes $(u, v) \in R_i^c$, a label $B_i$ on an arrow from world $u$ to $v$ denotes $(u, v) \in R_i^B$, and a label $CA_c$ on an arrow from world $u$ to $v$ denotes $(u, v) \in R_{CA}^c$. See text for further details.

$\{\{1,2\}\}, Q = \{\alpha\}, C = \{\alpha, \{1,2\}\}, X = \{(\alpha, \{1,2\}\}, Y = \emptyset, W = \{w_1, w_2, w_3, w_4, w_5\}.$

Note in particular that we are only concerned with a subset of the subgroups of $\mathcal{A}$, i.e. $\mathcal{G} \subset 2^\mathcal{A}$. We now determine some of the truth values for sentences in this model. We can immediately see that $M_{CG}, w_1 \vDash \neg p$. Also, $M_{CG}, w_1 \vDash B_1 \neg p$ holds because of $M_{CG}, w_2 \vDash \neg p$ and $R_1^B(w_1) = \{w_2\}$. Because of a similar reasoning, $M_{CG}, w_1 \vDash B_2 \neg p$ holds and therefore $M_{CG}, w_1 \vDash EB_{\alpha} \neg p$. In fact, because $R_1^B(w_2) = R_2^B(w_2) = \{w_2\}$ and $R_1^B(w_3) = R_2^B(w_3) = \{w_3\},$ we have $M_{CG}, w_1 \vDash MB_{\alpha} \neg p$. However, because of $M_{CG}, w_4 \vDash p, M_{CG}, w_5 \vDash p, \text{and } R_{CA}^c(w_1) = \{w_4, w_5\},$ we have $M_{CG}, w_1 \vDash CA_{\alpha} p$. We leave it as an exercise to the reader to demonstrate that $M_{CG}, w_1 \vDash EA_{\alpha} p \land MB_{\alpha} EA_{\alpha} p$.

Axiomatcs

The following axioms and inference rules hold because our belief- and acceptance-operators are normal modal operators (Fagin et al., 1995):

(\textbf{Classical}) All propositional tautologies

(\textbf{MP}) $\frac{\varphi}{\varphi \rightarrow \psi}$

(\textbf{Nec}_B) $\frac{\varphi}{B_i \varphi}$

(\textbf{Nec}_A) $\frac{\varphi}{A_i^c \varphi}$

(\textbf{K}_B) $(B_i \varphi \land B_i (\varphi \rightarrow \psi)) \rightarrow B_i \psi$
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

\[(K_A) \ (A_i^c \varphi \land A_i^c (\varphi \rightarrow \psi)) \rightarrow A_i^c \psi\]

The classical propositional tautologies (Classical), the inference rules modus ponens (MP) and necessitation (Nec_B and Nec_A), as well as the distribution axioms (K_B and K_A) define a normal modal logic. Axioms K_B and K_A and Axioms Nec_B and Nec_A are responsible for the problem of logical omniscience. They cause an agent to believe and accept all the deductive consequences of its beliefs and acceptances and to believe and accept all theorems. This is clearly an assumption that is too strong for resource-bounded reasoners such as humans or artificial agents for that matter. We will return to the problems associated with this assumption later.

The semantic constraints S1 and S2 correspond to the following axioms of consistency, positive introspection, and negative introspection (Fagin et al., 1995):

\[(D_B) \ B_i \varphi \rightarrow \neg B_i \neg \varphi\]
\[(4_B) \ B_i \varphi \rightarrow B_i B_i \varphi\]
\[(5_B) \ \neg B_i \varphi \rightarrow B_i \neg B_i \varphi\]

\[(D_A) \ A_i^c \varphi \rightarrow \neg A_i^c \neg \varphi\]
\[(4_A) \ A_i^c \varphi \rightarrow A_i^c A_i^c \varphi\]
\[(5_A) \ \neg A_i^c \varphi \rightarrow A_i^c \neg A_i^c \varphi\]

The semantic constraints S3 and S4 correspond to the following axioms:

\[(PosIntrA) \ A_i^c \varphi \rightarrow B_i A_i^c \varphi\]
\[(NegIntrA) \ \neg A_i^c \varphi \rightarrow B_i \neg A_i^c \varphi\]

The following additional axioms and inference rule characterize mutual belief (Fagin et al., 1995):

\[(EB) \ \text{EB}_c \varphi \leftrightarrow \bigwedge_{i \in X(c)} B_i \varphi\]
\[(MB) \ \text{MB}_c \varphi \leftrightarrow \text{EB}_c (\varphi \land \text{MB}_c \varphi)\]
\[(Ind) \ \varphi \rightarrow \text{EB}_c (\psi \land \varphi) \quad \varphi \rightarrow \text{MB}_c \psi\]

The following axiom, analogous to EB, follows directly from the semantics:

\[(EA) \ \text{EA}_c \varphi \leftrightarrow \bigwedge_{i \in X(c)} A_i^c \varphi\]
The following axiom, which corresponds to the semantic definition of collective acceptance above, is in line with the definition of collective acceptance by Tuomela (Definition 6.4), and, in addition, accounts for context:

\[(CA) \quad CA_c \varphi \leftrightarrow EA_c \varphi \land MB_c EA_c \varphi\]

Therefore, a proposition \( \varphi \) is collectively accepted in a context \( c \) if and only if every agent involved in context \( c \) accepts \( \varphi \) in \( c \) and this is mutually believed among these agents.

The following axioms are valid with respect to \( DCG \) (inferred from an analysis of distributed knowledge in the case of individuals by Fagin et al., 1995):

\[(DCG) \quad CA_c \varphi \rightarrow DCG_\alpha \varphi \text{ for } c = \alpha \text{ or } c \in Y(\alpha)\]

\[(K_{DCG}) \quad (DCG_\alpha \varphi \land DCG_\alpha (\varphi \rightarrow \psi)) \rightarrow DCG_\alpha \psi\]

That is, a proposition that is collectively accepted in a context \( c \) is also in distributed common ground of a joint action context \( \alpha \) if \( c \) is this relevant joint action context \( \alpha \) or if \( c \) is a group that is otherwise salient. Also, distributed common ground is closed under deduction.

The following axiom characterizes salient common ground and corresponds directly to the semantics:

\[(SCG) \quad CG_\alpha \varphi \leftrightarrow (\neg DCG_\alpha \bot \land DCG_\alpha \varphi) \lor (DCG_\alpha \bot \land CA_\alpha \varphi)\]

The following axiom represents the consistency of salient common ground:

\[(DCG) \quad CG_\alpha \varphi \rightarrow \neg CG_\alpha \neg \varphi\]

**Correspondence**

The correspondence between \( S1 \) and \( D_B, 4_B, \) and \( 5_B \) and the correspondence between \( S2 \) and \( D_A, 4_A, \) and \( 5_A \) are standard results (Fagin et al., 1995). It remains to show the correspondence between \( S3 \) and \( PosIntrA \) and between \( S4 \) and \( NegIntrA \).

**Proof for PosIntrA.** We need to prove for all models \( M_{CG} \) that \( A_i^j \varphi \rightarrow B_i A_i^j \varphi \) holds if and only if \( S3 \) holds.

For the direction from right to left assume an arbitrary model \( M_{GG} \) with \( S3 \). Now assume \( M_{GG}, u \models A_i^j \varphi \) in an arbitrary world \( u \). That is, \( M_{GG}, w \models \varphi \) in all worlds \( w \) with \( (u, w) \in R_i^{jc} \).

Since \( M_{GG}, w \models \varphi \) and there is a world \( v \) with \( (u, v) \in R_i^c \) and \( (v, w) \in R_i^{jc} \) only if \( (u, w) \in R_i^{jc} \) (S3), we have \( M_{GG}, v \models A_i^j \varphi \) for all such \( v \), and hence \( M_{GG}, u \models B_i A_i^j \varphi \).

We prove the direction from left to right by contraposition. Assume that \( S3 \) does not hold in a model \( M_{GG} \). Then there are worlds \( u, v, w \in W \) with \( (u, v) \in R_i^c \) and \( (v, w) \in R_i^{jc} \) but not \( (u, w) \in R_i^{jc} \). We show that there can be a valuation such that \( M_{GG}, u \models A_i^j p \) and
The axiomatics of Theorem 6.2.

Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

Follows from the correspondence proofs above.

To prove the soundness of the axiomatics of

Proof. Assume \( M_{GG}, s \models p \) for all worlds \( s \) with \( (u, s) \in R^e_A \).
So we have \( M_{GG}, u \models A^e_i p \). Hence, we have \( M_{GG}, v \not\models A^e_i p \) and consequently \( M_{GG}, u \not\models B^e_i A^e_i p \). Hence, we do not have \( A^e_i \phi \rightarrow B_i A^e_i \phi \) in such a model. \( \square \)

Proof for NegIntrA. We need to prove for all models \( M_{CG} \) that \( \neg A^e_i \phi \rightarrow B_i \neg A^e_i \phi \) holds if and only if S4 holds.

For the direction from right to left assume an arbitrary model \( M_{GG} \) with S4. Now assume \( M_{GG}, u \models \neg A^e_i \phi \) in an arbitrary world \( u \). That is, there is at least one world \( w \) with \( (u, w) \in R^e_A \) and \( M_{GG}, w \not\models \phi \). Because of S4 we have \( (u, w) \in R^i_A \) and \( (u, v) \in R^i_B \) only if \( (v, w) \in R^i_A \) for every such world \( v \). That is, \( M_{GG}, v \models \neg A^e_i \phi \) for each \( v \) and therefore \( M_{GG}, u \models B_i \neg A^e_i \phi \).

We prove the direction from left to right by contraposition. Assume that S4 does not hold in a model \( M_{GG} \). Then there are worlds \( u, v, w \in W \) with \( (u, w) \in R^i_A \) and \( (u, v) \in R^i_B \) but not \( (v, w) \in R^i_A \). We can now construct a valuation such that \( M_{GG}, w \not\models p \) and such that \( M_{GG}, t \models p \) for a proposition \( p \) for all worlds \( t \) with \( (v, t) \in R^i_A \). On the one hand, it follows that \( M_{GG}, u \models \neg A^e_i p \) because \( (u, w) \in R^i_A \). On the other hand, it follows that \( M_{GG}, v \models A^e_i p \) and thus \( M_{GG}, u \not\models B_i \neg A^e_i p \). Hence, we do not have \( \neg A^e_i \phi \rightarrow B_i \neg A^e_i \phi \) in such a model. \( \square \)

In summary, we have shown that the axioms for positive and negative introspection correspond to particular semantic restrictions on belief and acceptance accessibility relations.

Soundness

Theorem 6.2. The axiomatics of \( LCG \) are sound.

Proof. To prove the soundness of the axiomatics of \( LCG \), we need to show the validity of the axioms and inference rules.

The validity of the standard modal inference rules and axioms Classical, MP, NecB, NecA, Kb, and KA follows from previous work, as does the validity of the axioms DB, 4B, 5B, DA, 4A, and 5A (Fagin et al., 1995). The validity of the axioms PosIntrA and NegIntrA follows from the correspondence proofs above.

The validity of the axioms EB and MB, the inference rule Ind, and the axioms DCG and KD are standard results (Fagin et al., 1995). We sketch the proofs for the latter two axioms briefly. For Axiom DCG, note in the semantics that the set of worlds, \( v \), reachable from \( w \), for which \( \phi \) holds for CA, \( \phi \) to hold, are a superset of those for DCG, \( \phi \). Axiom KD is straightforward distribution axiom that holds by noting that, for any possible world \( v \), if \( \phi \) holds and \( \phi \rightarrow \psi \), then \( \psi \) holds from MP. Axioms EA and SCG hold directly from the semantics. Remains to show the validity of CA and DC.
\( \nu \in R_{CA}^C(w). \) Since \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \cup \{ v \mid u \in R_{B}^{X(c)}(w) \land v \in \bigcup_{i \in X(c)} R_{A}^{iE}(u) \} \) only if \( \nu \in R_{CA}^C(w) \) (see Definition 6.11), we have \( M_{CG}, \nu \vDash \phi \) in all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \cup \{ v \mid u \in R_{B}^{X(c)}(w) \land v \in \bigcup_{i \in X(c)} R_{A}^{iE}(u) \} \). In particular, \( M_{CG}, \nu \vDash \phi \) holds in all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \) for all \( i \in X(c) \). Therefore, we have \( M_{CG}, \nu \vDash A^i_\alpha \) for all \( i \in X(c) \) and hence \( M_{CG}, \nu \vDash EA_\alpha \). \( M_{CG}, \nu \vDash \phi \) also holds in all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \cup \{ v \mid u \in R_{B}^{X(c)}(w) \land v \in \bigcup_{i \in X(c)} R_{A}^{iE}(u) \} \). Therefore, for all worlds \( \nu \) that are \( G_B \)-reachable from world \( w \) for the group \( X(c) \) it holds that \( M_{CG}, \nu \vDash A^i_\alpha \phi \) for all \( i \in X(c) \). Hence, we have \( M_{CG}, \nu \vDash MB_\alpha EA_\alpha \phi \) and therefore \( CA_\alpha \phi \rightarrow EA_\alpha \phi \land MB_\alpha EA_\alpha \phi \).

Second, assume \( M_{CG}, \nu \vDash \phi \) for all worlds \( \nu \) with \( \nu \in R_{A}^i(w) \) for all \( i \in X(c) \) and for all worlds \( \nu \) for which there exists a world \( \mu \) such that \( u \) is \( G_B \)-reachable from world \( w \) for group \( X(c) \) and \( \nu \in R_{A}^{iC}(u) \) for any \( i \in X(c) \). Hence, \( M_{CG}, \nu \vDash \phi \) for all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \cup \{ v \mid u \in R_{B}^{X(c)}(w) \land v \in \bigcup_{i \in X(c)} R_{A}^{iE}(u) \} \). Because \( \nu \in R_{CA}^C(w) \) holds only if \( \nu \in \bigcup_{i \in X(c)} R_{A}^{iC}(w) \cup \{ v \mid u \in R_{B}^{X(c)}(w) \land v \in \bigcup_{i \in X(c)} R_{A}^{iE}(u) \} \) (see Definition 6.11), we have \( M_{CG}, \nu \vDash \phi \) for all worlds \( \nu \) with \( \nu \in R_{CA}^C(w) \). Thus follows \( M_{CG}, \nu \vDash CA_\alpha \) and hence \( CA_\alpha \phi \leftarrow EA_\alpha \phi \land MB_\alpha EA_\alpha \phi \).

The validity of Axiom \( DCG \) is proved by contradiction. Suppose there is a model \( M_{CG} \) in which both \( CG_\alpha \phi \) and \( CG_\alpha \neg \phi \) are true in some world \( w \). Then it follows by \( SCG \) that either:

1. \( M_{CG}, w \vDash (DCG_\alpha \bot \land CA_\alpha \phi) \) and \( M_{CG}, w \vDash (DCG_\alpha \bot \land CA_\alpha \neg \phi) \); or
2. \( M_{CG}, w \vDash (\neg DCG_\alpha \bot \land DCG_\alpha \phi) \) and \( M_{CG}, w \vDash (\neg DCG_\alpha \bot \land DCG_\alpha \neg \phi) \)

The first option implies that \( M_{CG}, \nu \vDash \phi \) for all worlds \( \nu \in R_{CA}^C(w) \) but also \( M_{CG}, \nu \vDash \neg \phi \) for all worlds \( \nu \in R_{CA}^C(w) \) with there being at least one such world \( \nu \), which creates a contradiction immediately. For the second option we argue as follows. From \( M_{CG}, \nu \vDash (\neg DCG_\alpha \bot \land DCG_\alpha \phi) \) it follows that \( M_{CG}, \nu \vDash \phi \) for all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{CA}^i \cap \bigcup_{G \in Y(\alpha)} R_{CA}^{G} \) and there is at least one such world \( \nu \). From \( M_{CG}, \nu \vDash (\neg DCG_\alpha \bot \land DCG_\alpha \neg \phi) \) follows that \( M_{CG}, \nu \vDash \neg \phi \) for all worlds \( \nu \) with \( \nu \in \bigcup_{i \in X(c)} R_{CA}^i \cap \bigcup_{G \in Y(\alpha)} R_{CA}^{G} \) and there is at least one such world \( \nu \). Hence, the contradiction. Thus follows \( \vdash CG_\alpha \phi \rightarrow \neg CG_\alpha \neg \phi \).

**Properties**

In this subsection, we study the properties of Common Ground Logic. We write \( \vdash \phi \) if \( \phi \) is a theorem of \( L_{CG} \) and we write \( \not\vdash \phi \) if \( \phi \) is not a theorem of \( L_{CG} \).

**Mutual Belief** The following theorems and inference rule are valid with respect to our definition of mutual belief (Dunin-Kęplicz and Verbrugge, 2010):

\[(MB\text{ Distribution} ) \quad MB_\phi \land MB_\psi (\phi \rightarrow \psi) \rightarrow MB_\psi\]
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

(MB Generalization) \[ \frac{\varphi}{MB_c \varphi} \]

(MB Conjunction Distribution) \[ MB_c (\varphi \land \psi) \leftrightarrow MB_c \varphi \land MB_c \psi \]

Consistency of CA We have the following theorem pertaining to the consistency of collective acceptance:

Theorem 6.3.

\[ \vdash CA_c \varphi \rightarrow \neg CA_c \neg \varphi \]

Theorem 6.3 implies that collective acceptances are always consistent in the same manner that individual belief and acceptance and salient common ground are. This theorem is important because it shows that collective acceptance behaves in the same way as those standard modal operators when it comes to consistency.

Proof of Theorem 6.3. We prove this by contradiction. Assume there is \( CA_c \varphi \) and \( CA_c \neg \varphi \).

(1) \( CA_c \varphi \) from assumption.

(2) \( CA_c \neg \varphi \) from assumption.

(3) \( EA_c \varphi \) from (1), \( CA \), and \( MP \).

(4) \( EA_c \neg \varphi \) from (2), \( CA \), and \( MP \).

(5) \( A_i^c \varphi \) for all \( i \in X(c) \) from (3), \( EA \), and \( MP \).

(6) \( A_i^c \neg \varphi \) for all \( i \in X(c) \) from (4), \( EA \), and \( MP \).

(7) \( \neg A_i^c \neg \varphi \) for all \( i \in X(c) \) from (5), \( DA \), and \( MP \).

Lines (6) and (7) contradict. Hence follows \( CA_c \varphi \rightarrow \neg CA_c \neg \varphi \).

Positive Introspection and Awareness We also have the following theorems:

Theorem 6.4.

\[ \vdash CA_c \varphi \rightarrow B_i CA_c \varphi \text{ for } i \in X(c) \]

Theorem 6.5.

\[ \vdash CA_\alpha \varphi \rightarrow B_i CG_\alpha \varphi \text{ for } i \in X(\alpha) \]
Thus, agents possess positive introspection to their collective acceptances. However, there is positive introspection into salient common ground only with respect to those propositions that are salient common ground because they are in context-specific common ground. We care about the properties relating to positive introspection because they establish a relationship between actual and perceived common ground. The idea, which we discuss in more detail later, is that the belief of common ground represents a type of perceived common ground.

Proof of Theorem 6.4.

1. \( B_i(EA_c \varphi \land MB_i EA_c \varphi \rightarrow CA_c \varphi) \) from CA and \( \text{Nec}_B \).
2. \( CA_c \varphi \rightarrow MB_i EA_c \varphi \) from CA.
3. \( MB_i EA_c \varphi \rightarrow EB_c(EA_c \varphi \land MB_i EA_c \varphi) \) from MB.
4. \( EB_c(EA_c \varphi \land MB_i EA_c \varphi) \rightarrow B_i(EA_c \varphi \land MB_i EA_c \varphi) \) from EB.
5. \( B_i(EA_c \varphi \land MB_i EA_c \varphi) \rightarrow B_iCA_c \varphi \) from (1) and \( \text{K}_B \).
6. \( CA_c \varphi \rightarrow B_iCA_c \varphi \) from (2), (3), (4), (5), and Classical.

\( \square \)

Proof of Theorem 6.5.

1. \( B_i(CA_\alpha \varphi \rightarrow DCG_\alpha \varphi) \) from DCG and \( \text{Nec}_B \).
2. \( CA_\alpha \varphi \rightarrow B_iCA_\alpha \varphi \) from Theorem 6.4.
3. \( B_iCA_\alpha \varphi \rightarrow B_iDCG_\alpha \varphi \) from (1) and \( \text{K}_B \).
4. \( CA_\alpha \varphi \rightarrow B_i(CA_\alpha \varphi \land DCG_\alpha \varphi) \) from (2), (3), Classical, and the fact that the \( B_i \) are normal modal operators.
5. \( B_i(CA_\alpha \varphi \land DCG_\alpha \varphi \rightarrow CG_\alpha \varphi) \) from SCG, Classical, and \( \text{Nec}_B \).
6. \( B_i(CA_\alpha \varphi \land DCG_\alpha \varphi) \rightarrow B_iCG_\alpha \varphi \) from (5) and \( \text{K}_B \).
7. \( CA_\alpha \varphi \rightarrow B_iCG_\alpha \varphi \) from (4), (6), and Classical.

\( \square \)

Based on Theorems 6.4 and 6.5, we can prove the following theorems:

**Theorem 6.6.**

\[ \vdash CA_c \varphi \rightarrow MB_i CA_c \varphi \]
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

Theorem 6.7.

\[ \vdash \text{CA}_\alpha \varphi \rightarrow \text{MB}_\alpha \text{CG}_\alpha \varphi \]

Theorem 6.6 implies the mutual awareness of collective acceptances. Again, there is mutual awareness of salient common ground only with respect to propositions in context-specific common ground (Theorem 6.7). The relevance of these theorems arises again from their role in establishing the relationship between actual and perceived common ground, considering that the mutual belief of common ground can be considered a type of perceived common ground.

Proof of Theorem 6.6 (adopted from Lorini et al. (2009)).

1. \[ \text{CA}_c \varphi \rightarrow \bigwedge_{i \in X(c)} B_i \text{CA}_c \varphi \] from Theorem 6.4.

2. \[ \bigwedge_{i \in X(c)} B_i \text{CA}_c \varphi \rightarrow \text{EB}_c \text{CA}_c \varphi \] from \text{EB}.

3. \[ \text{CA}_c \varphi \rightarrow \text{EB}_c (\text{CA}_c \varphi \land \text{CA}_c \varphi) \] from (1), (2), and \text{Classical}.

4. \[ \text{CA}_c \varphi \rightarrow \text{MB}_c \text{CA}_c \varphi \] from (3) and \text{Ind}.

The proof of Theorem 6.7 follows analogously.

Negative Introspection and Awareness

Note that we do not have negative introspection:

\[ \not \vdash \neg \text{CA}_c \varphi \rightarrow \bigwedge_{i \in X(c)} B_i \neg \text{CA}_c \varphi \] (6.1)

\[ \not \vdash \neg \text{CG}_\alpha \varphi \rightarrow \bigwedge_{i \in X(c)} B_i \neg \text{CG}_\alpha \varphi \] (6.2)

Assertion 6.1 holds because of the properties of the standard model of mutual belief we adopt here (Fagin et al., 1995). There can be the lack of mutual belief of a proposition without every involved agent individually believing this. Figure 6.3 provides an example of an \text{MC}_G \text{G} \text{M} \text{C}_G \text{G} \text{M} \text{C}_G \text{G} model with \( M_{CG}, w_1 \models \neg \text{CA}_c \varphi \) but with \( M_{CG}, w_1 \not\models B_2 \neg \text{CA}_c \varphi \). This example can also easily be extended to a counter-example for negative introspection for salient common ground (Assertion 6.2). In the simplest case, we can set \( c = \alpha \) and \( Y(\alpha) = \emptyset \) in the model represented by Figure 6.3 and therefore obtain \( M_{CG}, w_1 \models \neg \text{CG}_\alpha \varphi \) but \( M_{CG}, w_1 \not\models B_2 \neg \text{CG}_\alpha \varphi \). As we discuss in more detail later, this represents a way in which actual common ground can deviate from perceived common ground.

156
6.1. Common Ground

Figure 6.3: This example model $M_{CG}$ shows that there is no negative introspection into collective acceptance. On this figure, a label $A^c_i$ on an arrow from world $u$ to world $v$ denotes $(u, v) \in R^c_A$, a label $B_i$ on an arrow from world $u$ to $v$ denotes $(u, v) \in R^c_B$, and a label $CA_c$ on an arrow from world $u$ to $v$ denotes $(u, v) \in R^c_{CA}$. We have $M_{CG}, w_4 \models \neg p$ and $M_{CG}, w_5 \models p$. We have $M_{CG}, w_1 \models \neg CA_c p$ because $(w_1, w_4) \in R^1_A$ and $M_{CG}, w_4 \models \neg p$. However, we have $M_{CG}, w_3 \models CA_c p$ and thereby $M_{CG}, w_1 \models B_2 CA_c p$ and hence $M_{CG}, w_1 \models \neg B_2 \neg CA_c p$ and $M_{CG}, w_1 \not\models B_2 \neg CA_c p$.

Even with positive and negative introspection of collective acceptance and salient common ground, the following unnecessarily strong results would arise:

\[
CA_c \varphi \leftrightarrow MB_c CA_c \varphi \\
\neg CA_c \varphi \leftrightarrow MB_c \neg CA_c \varphi \\
CG_a \varphi \leftrightarrow MB_a CG_a \varphi \\
\neg CG_a \varphi \leftrightarrow MB_a \neg CG_a \varphi
\]

Because we do not have these results, a mutual belief of a collective acceptance (or salient common ground) does not imply this collective acceptance (or salient common ground), and obviously we do not have the results pertaining to negative awareness.

**Distribution and Necessity**  We have the following further theorems:

**Theorem 6.8.**

\[
\vdash (CA_c \varphi \land CA_c (\varphi \rightarrow \psi)) \rightarrow CA_c \psi
\]

**Theorem 6.9.**

\[
\vdash (CG_a \varphi \land CG_a (\varphi \rightarrow \psi)) \rightarrow CG_a \psi
\]
That is, not only distributed common ground is closed under deduction ($K_{DCG}$) but also collective acceptance and salient common ground are. This result matters because it shows that deductive inference is possible on collective acceptance and common ground.

**Proof of Theorem 6.8.**

(1) $CA_c \phi \land CA_c (\phi \rightarrow \psi) \rightarrow EA_c \phi \land MB_c EA_c (\phi \rightarrow \psi) \land MB_c EA_c (\phi \rightarrow \psi)$ from $CA$.

(2) $EA_c \phi \land EA_c (\phi \rightarrow \psi) \rightarrow EA_c \psi$ from $K_A$, $EA$, and Classical.

(3) $MB_c[EA_c \phi \land EA_c (\phi \rightarrow \psi) \rightarrow EA_c \psi]$ from (2) and MB Generalization.

(4) $MB, EA_c \phi \land MB, EA_c (\phi \rightarrow \psi) \rightarrow MB, [EA_c \phi \land EA_c (\phi \rightarrow \psi)]$ from MB Conjunction Distribution.

(5) $MB_c[EA_c \phi \land EA_c (\phi \rightarrow \psi)] \land MB_c[EA_c \phi \land EA_c (\phi \rightarrow \psi) \rightarrow EA_c \psi] \rightarrow MB, EA_c \psi$ from MB Distribution.

(6) $EA_c \psi \land MB, EA_c \psi \rightarrow CA_c \psi$ from CA.

(7) $CA_c \phi \land CA_c (\phi \rightarrow \psi) \rightarrow CA_c \psi$ from (1) to (6) and Classical.

□

**Proof of Theorem 6.9.**

(1) $CG_\alpha \phi \land CG_\alpha (\phi \rightarrow \psi) \rightarrow [\neg DCG_\alpha \bot \land DCG_\alpha \phi] \lor (DCG_\alpha \bot \land CA_\alpha \phi)]$

$\land [\neg DCG_\alpha \bot \land DCG_\alpha (\phi \rightarrow \psi)]$

$\lor (DCG_\alpha \bot \land CA_\alpha (\phi \rightarrow \psi)])$

from SCG.

(2) $[\neg DCG_\alpha \bot \land DCG_\alpha \phi] \lor (DCG_\alpha \bot \land CA_\alpha \phi)]$

$\land [\neg DCG_\alpha \bot \land DCG_\alpha (\phi \rightarrow \psi)] \lor (DCG_\alpha \bot \land CA_\alpha (\phi \rightarrow \psi)])$

$\rightarrow [\neg DCG_\alpha \bot \land DCG_\alpha \phi \land DCG_\alpha (\phi \rightarrow \psi)] \lor [DCG_\alpha \bot \land CA_\alpha \phi \land CA_\alpha (\phi \rightarrow \psi)]$

from Classical.

(3) $\neg DCG_\alpha \bot \land DCG_\alpha \phi \land DCG_\alpha (\phi \rightarrow \psi) \rightarrow \neg DCG_\alpha \bot \land DCG_\alpha \psi$ from $K_{DCG}$.

(4) $\neg DCG_\alpha \bot \land DCG_\alpha \psi \rightarrow CG_\alpha \psi$ from SCG.

(5) $DCG_\alpha \bot \land CA_\alpha \phi \land CA_\alpha (\phi \rightarrow \psi) \rightarrow DCG_\alpha \bot \land CA_\alpha \psi$ from Theorem 6.8.

(6) $DCG_\alpha \bot \land CA_\alpha \psi \rightarrow CG_\alpha \psi$ from SCG.
6.1. Common Ground

(7) \( CG_\alpha \varphi \land CG_\alpha (\varphi \rightarrow \psi) \rightarrow CG_\alpha \psi \) from (1) to (6) and Classical.

Also the inference rules \( \text{Nec}_{CA}, \text{Nec}_{DCG} \) and \( \text{Nec}_{CG} \) analogous to \( \text{Nec}_{B} \) and \( \text{Nec}_{A} \) follow trivially. If \( \varphi \) is a theorem, then this is believed, mutually believed, accepted, etc. These results are important because they show that distributed common ground and salient common ground behave to a large extent in the same way as standard modal operators of belief.

Actual and Perceived Common Ground  Let us consider now in which ways perceived common ground can deviate from actual common ground and which ramifications this has. First, however, we need to discuss in more detail what we consider perceived common ground to be. In Section 6.1.2, we identified perceived common ground as the second component of collective acceptance \( (MB_i, EA_c, \varphi) \), in line with Kashima et al.’s definition of common ground (Definition 6.7). However, in the context of the study on stereotypes discussed in the last chapter, the term has also been used by Lyons and Kashima (2003) to denote an individual’s belief about common ground. Hence, we figure that perceived common ground can more broadly be understood as any (mutual) belief of groups or individuals about (actual) common ground.

From our proof of Theorem 6.4 (line 3 to 5) it follows that \( MB_i, EA_c, \varphi \) is sufficient to imply \( B_i, CA_c, \varphi \) and due to Theorem 6.6 also \( MB_i, CA_c, \varphi \). That is, a (mutual) belief of a collective acceptance and hence personal or communal common ground is implied even if we do not have actual common ground \( (EA_c, \varphi) \). In this way, the common ground perceived by individuals and by a group can deviate from actual common ground.

We also do not have \( B_i, CA_c, \varphi \rightarrow CA_c, \varphi \) or \( MB_i, CA_c, \varphi \rightarrow CA_c, \varphi \), nor \( B_i, CG_\alpha, \varphi \rightarrow CG_\alpha, \varphi \) or \( MB_i, CG_\alpha, \varphi \rightarrow CG_\alpha, \varphi \). Hence, perceived common ground does not necessarily imply corresponding actual common ground.

Furthermore, we do not in general have \( CG_\alpha, \varphi \rightarrow B_i, CG_\alpha, \varphi \). Hence the salient common ground of a proposition does not imply correct beliefs about this.

Example  We demonstrate with a brief example how Common Ground Logic accounts for the bottom-up emergence of the common ground of groups and for the top-down determination of salient common ground.

Assume an interaction between two Keynesian economists \( i \) and \( j \). Let \( \alpha \) denote this interaction, \( X(\alpha) = \{i, j\} \) the set of both economists, and \( Y(\alpha) = \{P, G\} \) with \( P = \{i, j\} \) and \( \{i, j\} \subseteq G \) the salience of two group memberships—of the group \( P \) associated with the economists’ personal common ground and the group \( G \) of all Keynesian economists, which is associated with a respective communal common ground.
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

Now consider the following initial setup, determined by the economists’ previous common experiences and the communal common ground of all Keynesian economists:

\[ CA_{\text{high}} \land CA_{\text{G}(\text{high} \rightarrow \text{insuff})} \]  

(6.3)

where \( \text{high} \) expresses that unemployment is unusually high and \( \text{high} \rightarrow \text{insuff} \) that, if unemployment is unusually high (\( \text{high} \)), monetary and fiscal policies are insufficient (\( \text{insuff} \)). At this point, context-specific common ground is unspecified. From Equation 6.3 follows that distributed common ground is consistent, i.e. \( \neg DCG_{\alpha} \bot \), and the following holds:

\[ DCG_{\alpha \text{high}} \land DCG_{\alpha}(\text{high} \rightarrow \text{insuff}) \]

Because of \( K_{DCG} \) we also have \( DCG_{\alpha \text{insuff}} \). From this follows through \( \text{SCG} \):

\[ CG_{\alpha \text{high}} \land CG_{\alpha}(\text{high} \rightarrow \text{insuff}) \land CG_{\alpha \text{insuff}} \]

It is salient common ground between \( i \) and \( j \) that unemployment is unusually high (\( \text{high} \)), economic policies are insufficient if unemployment is unusually high (\( \text{high} \rightarrow \text{insuff} \)), and economic policies are insufficient (\( \text{insuff} \)). These sentences enter top-down into salient common ground through the interaction of personal and communal common ground.

Consider now that during the course of the interaction economists \( i \) and \( j \) adopt the following attitudes:

\[ A_{\alpha i}^{\neg \text{high}} \land A_{\alpha j}^{\neg \text{high}} \land MB_{\alpha} EA_{\alpha}^{\neg \text{high}} \]

This yields also \( CA_{\text{high}} \), i.e. the context-specific common ground that unemployment is not unusually high. From the clash with \( CA_{\text{high}} \) follows \( DCG_{\alpha} \bot \). Therefore, salient common ground is specified only by context-specific common ground and we have \( CG_{\alpha}^{\neg \text{high}} \).

Assume that the two economists later establish that their context-specific common ground is actually also personal common ground and they accept this also for the communal common ground of \( G \), i.e. \( CA_{\text{high}} \land A_{i}^{G \neg \text{high}} \land A_{j}^{G \neg \text{high}} \). From this follows \( \neg DCG_{\alpha} \bot \) and \( CG_{\alpha}^{\neg \text{high}} \land CG_{\alpha}(\text{high} \rightarrow \text{insuff}) \). Also, the first step towards the bottom-up emergence of \( CA_{\text{high}} \) is done.

6.1.3 Related Work

We compare our Common Ground Logic against the two most similar formalizations: Grounding Logic by Gaudou et al. (2006a) and Acceptance Logic by Lorini et al. (2009). Both are modal logics based on possible world semantics.
6.1. Common Ground

Grounding Logic

The $G$-operator of Gaudou et al. (2006a) expresses what is publicly grounded in a group. A proposition $\phi$ becomes publicly grounded in a group $I$ (i.e. $G_I \phi$) once it is grounded in the group for every group member $i$ that this proposition is grounded for this group member ($\bigwedge_{i \in I} G_i G_i \phi \rightarrow G_I \phi$). $G_I G_i$ is achieved by agent $i$ openly expressing $\phi$ in the presence of group $I$. This represents the bottom-up mechanism by which a proposition enters the common ground of a group or in fact becomes publicly grounded for that group. While this mechanism is suitable for smaller groups, it does not appear appropriate for the emergence of common ground in larger collectives: There is no way for a proposition to become publicly grounded in a collective by being openly expressed by every member of the collective.

In Grounding Logic what is grounded in a group is generally independent of what is grounded in its super- and subgroups. The only exceptions are:

- If a proposition $\phi$ is (not) grounded in a group $I$, then it is grounded in every subgroup $I'$ that $\phi$ is (not) grounded in $I$ ($G_I \phi \rightarrow G_I G_I' \phi$ and $\neg G_I \phi \rightarrow G_I' \neg G_I \phi$ respectively, which are variations of Axioms 4 and 5).

- If a proposition $\phi$ is grounded in a group $I$, then it is grounded in $I$ that $\phi$ is grounded in every subgroup $I'$, for certain “objective” $\phi$ ($G_I \phi \rightarrow G_I G_{I'} \phi$).

In contrast, we do not specify any interactions between the collective acceptances of a group and the collective acceptances of its super- or subgroups. The motivation for this is that groups are not to be limited in the ways in which their collective acceptances can deviate from those of other groups. This means that collective acceptance and common ground in our case miss the variants of Axiom 4 and 5 present in the Grounding Logic.

However, we do specify how the collective acceptances of different groups are composed in specific situations through Axiom SCG to yield salient common ground. This top-down mechanism is stronger than the ones assumed by the Grounding Logic, yet it still does not specify any relationship between different common grounds other than those with salient common ground. In doing so, we also emphasize the distinction and interaction between salient, context-specific, personal, and communal common ground, which is not the case in the Grounding Logic.

Axiom D is specified for Grounding Logic’s $G$-operator. We have Axiom D for individual acceptance and belief but not for CA and CG. However, we show that consistency of collective acceptance and salient common ground are in fact theorems of our logic. Hence collective acceptances and salient common ground are consistent, at least within respective contexts.
Acceptance Logic

The $A$-operator of Lorini et al. (2009) expresses what is (collectively) accepted by a single agent or a group of agents *qua* members of an institution. A (collective) acceptance is always interpreted with respect to a group of agents and an institutional context. The symbol $A_{C,x}\varphi$ denotes that $\varphi$ is collectively accepted by group $C$ in context $x$. In our case, collective acceptance and common ground is interpreted with respect to a context only (a joint action or a group), with the context determining the agents that are involved in what is collectively accepted or common ground.

Similar to Grounding Logic, Acceptance Logic makes a strong assumption about the top-down determination of collective acceptances by including variants of Axioms 4 and 5: If a proposition $\varphi$ is (not) collectively accepted by a group $C$ in context $x$ it is also collectively accepted by every subgroup $B$ in any other context $y$ that this is (not) the case ($A_{C,x}\varphi \rightarrow A_{B,y}A_{C,x}\varphi$ and $\neg A_{C,x}\varphi \rightarrow A_{B,y}\neg A_{C,x}\varphi$).

Furthermore, if a proposition $\varphi$ is collectively accepted by a group $C$ in context $x$, it is also collectively accepted by every subgroup $B$ in that context. A closer analysis of our Common Ground Logic reveals a similar property: Recall that a proposition is collectively accepted in a context if and only if everyone involved in that context accepts this proposition and there is mutual belief about this. When we consider a subgroup of the group involved in this context, it is also clearly true for this subgroup that everyone in that subgroup accepts this proposition and there is mutual belief about this. We did not analyze this property, though.

The bottom-up mechanism of the construction of collective acceptances is represented as follows: A proposition $\varphi$ is collectively accepted by a group $C$ in context $x$ if it is accepted by this group in this context that $\varphi$ is accepted in this context by all group members ($[A_{C,x} \bigwedge_{i \in C} A_{i,x}\varphi] \rightarrow A_{C,x}\varphi$). This is a strict principle of emergence because it represents that collective acceptance comes about by the collective agreement of the group’s members, or by their consensus as Lorini et al. have it. This makes sense for structured groups that have appropriate group decision-making processes in place. It does not seem so appropriate for unstructured groups. Acceptance Logic does not include Axiom D, hence apparently acceptances do not need to be consistent.

Tuomela (2009) remarks that Acceptance Logic in terms of its ideas is close to his account and even compatible in terms of semantics and axiomatics.

6.1.4 Discussion

The formalization of common ground introduced in this section represents the various types of common ground postulated by the grounding model of cultural transmission as well as the bottom-up emergence of communal common ground and the top-down determination of
salient common ground in interactions. The model in particular fulfills the requirements we identified in Section 6.1.1. We revisit these requirements briefly.

Our model of common ground is based on the I-mode collective acceptance account of Tuomela. We consider collective acceptances to be associated with a context, which is either a joint action or a group context. In the first case, we obtain context-specific common ground, in the latter case we obtain personal or communal common ground. We describe a model of salient common ground that is a composition of context-specific, personal, and communal common ground. Common ground in our case is non-summative and hence does not imply individual beliefs corresponding to what is common ground. However, we see common ground as reducible to individual mental attitudes, namely beliefs and acceptances. We do not make a technical distinction between personal and communal common ground and we find in an analysis of our logic (Section 6.1.2) that salient common ground shares a majority of properties with our model of collective acceptance. Both share the features of consistency, distribution, and necessity and the lack of negative introspection. The only feature salient common ground lacks compared to collective acceptance is full positive introspection.

In the following, we comment on a few further details. As Acceptance and Grounding Logic, so do we accept the problem of logical omniscience. An agent believes (accepts) all the deductive consequences of its beliefs (acceptances) as well as all theorems (the same for groups and collective acceptances). This is a strong assumption, yet a simplifying one. We feel that this idealization does not impair the value of the model as a framework for more concrete instantiations. In fact, various possible solutions for this problem have been proposed within the context of epistemic modal logics (see, for example, the discussion by Halpern and Moses, 1992).

Our model distinguishes carefully between the different parts of common ground and hence allows for powerful extensions. For example, the model keeps track of the source of any proposition in common ground, i.e. whether it is part of context-specific, personal, or communal common ground. This allows an agent to treat propositions in salient common ground according to where they come from. In our simple model of salient common ground, context-specific common ground has priority over personal and communal common ground, and personal and communal common ground are given equal priority. One could imagine a more expressive model in which there is a preference relation between different salient group memberships. In this case, different communal common grounds have priority over others.

Note that consistency of collective acceptance and salient common ground is restricted to a context, i.e. the content of the common ground in one context (group) does not impose any restrictions on the common ground in other contexts (groups), even if they share some of their group members.
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

On a brief note, the satisfiability problem, i.e. the problem of determining whether there is a model for a given sentence, for the modal logic KD45 with common knowledge (a variant of common belief) and multiple agents is EXPTIME-complete (Halpern and Moses, 1992). Common Ground Logic is even more expressive and possibly computationally even more complex. For small models and sentences, this is not necessarily a practical problem but it certainly is for larger models.

6.2 The Joint Activity

We turn now towards formalizing the other two major components of the grounding model of cultural transmission: the joint activity and the grounding process. From here on, our treatise is going to be rather semi-formal because of the complexity of properly integrating the three components of the grounding model of cultural transmission. However, we seek to provide a comprehensive computational-conceptual framework that can give rise to more precise formalizations.

The role of the joint activity is to provide a frame in whose context grounding and hence cultural transmission occurs. The joint activity poses epistemic and relational goals and thus determines to a large extent which information is communicated in which way. We have discussed in Chapter 2 how relational goals interact with common ground to influence grounding: If relational goals are strong, it is easier to ground information (i.e. add to common ground) that is compatible with common ground. The joint activity also determines whether the grounding of a particular proposition is adequate to advance this activity.

Different philosophical and computational approaches have been proposed to describe the mental attitudes of a group of agents engaged in a joint activity (Bratman, 1992; Cohen and Levesque, 1991; Grosz and Kraus, 1996; Tuomela, 2006). In one way or the other each of these models presumes that action comes about through practical reasoning over higher-level mental attitudes such as beliefs, goals, and (joint) intentions. The widely accepted view on practical reasoning advanced by Bratman (1987) in terms of the BDI model, which we briefly mentioned in Section 3.2.2, is as follows: An actor represents the current state of the world in terms of beliefs. Desired world states represent goals. Once the actor commits to a particular goal, an intention is formed that specifies the goal and the recipe or plan with which the actor intends to achieve the goal. Hence an intention refers both to an end and a means to achieve this end. Goals and intentions can be organized in complex hierarchies, in which intentions at one level are means to goals at the higher level but also include subgoals as means themselves. Because of its commitment, the actor will not drop intentions arbitrarily but will seek to achieve them. It is generally accepted that some form of joint intentions are required in the case of joint activities. We expand on this discussion in the next chapter.
By and large the models of joint activity mentioned above do not specify in detail how the dynamics of the joint activity play out, i.e. how agents collaboratively plan and execute their plans. Grosz’s and Kraus’ computational SharedPlan formalism (1996) and Dunin-Keplicz’s and Verbrugge’s TEAMLOG (2010) formalism are notable exceptions. We adopt the SharedPlan formalism here to describe the joint activities that coactors might be engaged in. The main benefit gained from the SharedPlans formalism is that it can specify how grounding processes are evoked by a joint activity, as we shall see later. Recall that the role of grounding as a sub-activity of everyday joint activities is a key assumption of the grounding model of cultural transmission. We note, though, that it is not quite clear how appropriate the SharedPlans model is for representing human joint activities. However, the model is informed by information about human joint activities (Grosz and Kraus, 1996) and the model of practical reasoning described above has found some acceptance in philosophy and psychology (see next chapter).

In the following, we describe the SharedPlan formalism to a degree sufficient for this thesis. The reader is referred to Grosz and Kraus (1996) for a comprehensive discussion. A SharedPlan describes a joint activity of a group of agents in terms of their beliefs and intentions (Grosz and Kraus, 1996). Together with a set of axioms and collaborative planning and decision processes (Grosz and Hunsberger, 2006), a SharedPlan ensures basic features of joint activities (Bratman, 1992): (i) the agents are mutually responsive to each other’s intentions and actions, (ii) they are committed to the joint activity, and (iii) they are committed to mutual support.

The formalization of SharedPlans relies on a first-order modal logic with the usual first-order connectives and modal operators for belief, mutual belief, intentions, and further operators relating to action. Intentions are either intentions-that or intentions-to, and either potential or actual intentions. Intentions-that ensure that agents are committed to the success of the joint activity and the other agents’ actions. Intentions-to ensure that agents are committed to performing their part of the activity, and they entail means-end reasoning. The target of an intention-that is a proposition, while the target of an intention-to is an action. Potential intentions are those intentions available as options before the agent has committed to any particular ones. Modal operators are interpreted over accessibility relations between possible worlds with each world corresponding to a temporal structure. We ignore the temporal aspect and other details for the sake of comprehensibility.

We denote by $L_{SP}$ the language of this logic and use a syntax compatible with the one in the previous section. We use $p, q$ for propositions and $\varphi, \psi$ for sentences. We denote actions by $\alpha, \beta,$ and $\gamma$. An action $\alpha$ with $\alpha(p_1, \ldots, p_n)$ has type $\alpha$ and parameters $p_1$ to $p_n$ and is defined by preconditions and effects, accessible by $\text{Prec}(\alpha)$ and $\text{Eff}(\alpha)$ respectively. We denote agents by $i, j$ and agent $i$’s belief of $\varphi$ by $B_i \varphi$ and mutual belief between a group of
agents $G$ by $MB_G \phi$. An intention of agent $i$ that a certain proposition $\phi$ holds is denoted by $I_i \phi$ (or $P_i \phi$ for a potential intention) and an intention to perform an action $\alpha$ by $I_i \alpha$ (or $P_i \alpha$ for a potential intention). An agent cannot have any conflicting intentions. The operator $Do_i \alpha$ denotes that agent $i$ does action $\alpha$, or alternatively $Do_G \alpha$ denotes that group $G$ does action $\alpha$. The predicate $CBA(i, \alpha, R_\alpha)$ (“can bring about”) denotes that agent $i$ is able to perform action $\alpha$ with recipe $R_\alpha$.

A group of agents $G$ essentially has a SharedPlan $SP(P_\alpha, G, \alpha)$ with name $P_\alpha$ for an action $\alpha$ if a set of mutual beliefs is established, e.g. that there is a recipe for this joint activity and that all agents intend to identify necessary parameters. A recipe for an action consists of a set of sub-actions to be executed and a set of constraints on their execution. A sub-action might be a complex action itself, in which case the agents need to establish a subordinate SharedPlan for the execution of this sub-action.

The reason for a SharedPlan of a group of agents is the mutually believed intention that they perform the action: \[ MB_G(\bigwedge_{i \in G} I_i Do_G \alpha). \]

For each basic subordinate action $\beta$ for a recipe to do action $\alpha$, the agents need to establish the mutual belief that a selected group member $j$ can carry out this action using recipe $R_\beta$:

\[ MB_G[(\exists R_\beta)CBA(j, \beta, R_\beta)] \]

and that they all intend this to happen:

\[ MB_G[\bigwedge_{i \in G} I_i[(\exists R_\beta)CBA(j, \beta, R_\beta)]]. \] (6.4)

Recall the example dialogues we introduced in Chapter 2 about the conversations that Alice has with others about the incident concerning Gary. Assume in the following this alternative sample dialogue between Alice and her husband Bob:

(1) Alice: Gary, a player from our club, got caught drunk-driving yesterday.
He spent the night in jail.
(2) Bob: I thought he was playing today?
(3) Alice: Yeah, but you know. It’s the typical story: He got drunk, drove his Porsche at 150, and

---

\[ ^6 \text{Note that if an agent believes that it has an intention-that, it actually does have this intention (Grosz and Kraus, 1996, p. 16).} \]
abused the police. Everyone knows how these guys tick.

(4) Bob: Yeah.

We now formalize how the joint activity of two agents $i$ and $j$ described as a Shared-Plan $P_\alpha$ entails grounding and suggest how the agents’ common ground affects individual interactions. The actual grounding process is the topic of the next section.

In addition to the modal operators specified by the SharedPlan formalism, we require the operators for acceptance and the interactions between acceptance and belief as specified in Section 6.1. Note that it is in principle possible to adjust our propositional modal logic of belief and acceptance in Section 6.1 to fit with the first-order modal logic of the SharedPlans formalism. However, we feel that this is not important for describing our ideas.

A SharedPlan’s recipe can pose epistemic goals that are to be fulfilled by communication (PREM8). We assume that the epistemic goal of agents $i$ and $j$ to ground proposition $\phi$ is described by the complex action $\text{Ground}(i, j, \phi)$, which is simply one of the subactions referred to by the SharedPlan’s recipe. We come back to these epistemic goals later. While the establishing of the SharedPlan’s mutual beliefs requires communication itself, we focus here on epistemic goals posed by the recipe.

In line with our discussion in the beginning of this section, we assume that the joint activity determines how an epistemic goal can be met by grounding one proposition from a set of alternative propositions that are adequate to advance the joint activity. For example, Alice has multiple different ways to tell the story about Gary’s arrest. She opts for the one leaving out the information that Gary got drunk on a bottle of wine. She could as well include this information as this would still be adequate to advance their joint activity. Within the context of a different joint activity, her choice might not have been precise enough to advance the activity. Consider, for example, that Alice and her husband are planning to invite Gary for dinner and they are trying to figure out what kind of drinks to serve. In that case, Alice might tell the story as follows: “Someone told me recently that Gary actually drank wine”, omitting here the irrelevant information about Gary’s drunk-driving.

Given an epistemic goal to ground a sentence $\phi$ described by the complex action $\text{Ground}(i, j, \phi)$, we denote by $\text{Substitutable}(\phi, \psi, P_\alpha)$ the normal modal operator that is true if and only if $\psi$ can be added to common ground instead of $\phi$, within the context of the SharedPlan $P_\alpha$. That is, the grounding of $\phi$ and $\psi$ is equally adequate to advance the joint activity. Obviously, $\text{Substitutable}$ can be given many different (and interesting) semantics. Here we simply consider $\text{Substitutable}(\phi, \psi, P_\alpha)$ to be true in a possible world if and only if $\text{CG}_\alpha(\psi \rightarrow \phi)$ is true in that world. That is, $\psi$ is adequate to be substituted for $\phi$ if and only if it is common ground that $\psi$ implies $\phi$. Because of Theorem 6.9 (Section 6.1.2), the common ground of $\psi$ is sufficient to infer the common ground of $\phi$. Consequently,
Chapter 6. Towards a Formal Model of the Grounding Model of Cultural Transmission

Substitutable($\phi, _\psi, P_\alpha$), where _ stands for any $\psi$, determines the precision required by the joint activity with regard to the epistemic goal $\text{Ground}(i, j, \phi)$.

We assume that the relational goals of the joint activity are expressed in terms of the intentions of the participating agents (PREM8). For example, the agents might be committed to improving their social relationship. Any subsequently adopted intentions need to be compatible with these prior intentions. Thus, the agents will not attempt any communication that would violate relational goals. That is, relational goals and epistemic goals might also be incompatible (PREM8). Most importantly, the evaluation of relational goals also depends on common ground as we discussed earlier with regard to Alice’s omission of Gary’s wine-drinking.

6.3 Grounding

As discussed in Chapter 2, we consider the grounding process to be some kind of negotiation of the exchanged information. Accordingly, we conceptualize this process as an argumentation dialogue. In the following, we illustrate with our example how such a grounding process can be described and how this process is driven by the agents’ involvement in the SharedPlan (PREM5). In particular, we provide a tentative formalization of an argumentation dialogue capable of representing the sample dialogue. We then describe how the sample dialogue relates to the provided formalism, paying particular attention to how the grounding process affects and is affected by common ground. We take a normative view here but discuss later how human agents might deviate from these expectations. We do by no means claim that our formalization can cover all types of dialogue that could play a role in the transmission of cultural information. However, we do provide a starting point for further elaboration.

6.3.1 The Argumentation Dialogue

As indicated, we describe the grounding process between two agents $x$ and $y$ as an argumentation dialogue, i.e. as an exchange of arguments. Argumentation dialogues can take different forms depending on the epistemic goal to achieve and the roles of the agents (Walton and Krabbe, 1995). For example, an agent can seek to persuade a disagreeing partner to ground a certain proposition, or they can deliberate or negotiate about which proposition to ground. We assume that the recipe of a casual conversation that is supposed to fulfill a higher-level relational goal requires from Alice to tell a story to Bob and she opts for the one about Gary.

An argument here is a pair $e = (H, h)$ with $H \subseteq \mathcal{L}_{SP}$ and $h \in \mathcal{L}_{SP}$ such that $H$ logically implies $h$ and $H$ is consistent and minimal (Parsons et al., 2002). We call $H$ the support of $e$ and $h$ its conclusion, denoted by $H = \text{Supp}(e)$ and $h = \text{Conc}(e)$. Let $e$ and $f$ be two
6.3. Grounding

arguments. We say that $e$ defeats $f$ if and only if $(\exists h \in \text{Supp}(f))[\text{Conc}(e) \equiv \neg h]$ (undercut) or if and only if $\text{Conc}(e) \equiv \neg \text{Conc}(f)$ (rebuttal).

An argumentation framework is a pair $AF = (\mathcal{H}, \mathcal{R})$ where $\mathcal{H}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{H} \times \mathcal{H}$ is a defeat relation between arguments (Dung, 1995). Which arguments an agent is able to assert or to accept\(^7\) in a dialogue depends on the argumentation framework the agent can construct from its beliefs, from salient common ground, and from the current dialectical commitments (refer to Parsons et al. (2002) for a discussion on assertion and acceptance attitudes). Obviously, in particular the agent’s beliefs about common ground should play a role in determining assertable and acceptable arguments and thereby affect cultural transmission (IMPL1). Let $C_i \subset L_{SP}$ be the set of agent $i$’s commitments. We denote by the predicate $\text{Assertable}(i, e)$ that agent $i$ is able to assert argument $e$ and by the predicate $\text{Acceptable}(i, e)$ that argument $e$ is acceptable for agent $i$. We do not commit to any particular assertion or acceptance attitudes here. However, we note that an argument is more likely acceptable the more information is common ground and the more compatible this argument is with this common ground (PREM6). Hence, more effort is needed during grounding of an argument the less acceptable this argument is (PREM6). Note that communal common ground plays a role in salient common ground and hence culture affects the grounding process and therefore cultural transmission (IMPL5).

Arguments are exchanged as the content of communicative actions. Here, we distinguish three communicative actions: (1) A present allows a speaker to present a proposition to a listener in form of an argument. (2) An approve allows the listener to approve (accept) a previously presented argument. (3) A challenge allows a speaker to challenge a previously uttered argument with another argument. A present opens a dialogue and an approve closes the dialogue. A challenge can either attack a previous present or a previous challenge. Obviously, this is but a subset of the communicative actions one would want to consider in a full-blown account of the grounding process. In the following, we describe the semantics of these communicative actions semi-formally in terms of their preconditions and effects over the agents’ mental attitudes, their common ground, and their dialectical commitments.

**Definition 6.12.** A present $(i, j, e)$ of argument $e$ by agent $i$ to agent $j$ requires that $e$ is assertable for $i$ and causes the conclusion and support of $e$ to be added to $i$’s commitments:

\[
\begin{align*}
\text{Prec}(\text{present}(i, j, e)) & \overset{\text{def}}{=} \text{Assertable}(i, e) \\
\text{Eff}(\text{present}(i, j, e)) & \overset{\text{def}}{=} C_i := C_i \cup \{\text{Conc}(e)\} \cup \text{Supp}(e)
\end{align*}
\]

\(^7\)Note that this notion of acceptance is different from the epistemic state of acceptance we were concerned with in Section 6.1.
**Definition 6.13.** An approve\((i, j, e)\) by agent \(i\) of a previous argument \(e\) of agent \(j\) requires that \(e\) was previously presented by \(j\) and that it is acceptable for \(i\). This action causes the conclusion and the support of \(e\) to be added to \(i\)’s commitment store. It also causes the conclusion and the elements of the support to be added to context-specific common ground:

\[
\text{Prec}(\text{approve}(i, j, e)) \overset{\text{def}}{=} \text{present}(j, i, e) \text{ has been done} \land \text{Acceptable}(i, e)
\]

\[
\text{Eff}(\text{approve}(i, j, e)) \overset{\text{def}}{=} C_i := C_i \cup \{\text{Conc}(e)\} \cup \text{Supp}(e) \land \text{CA}_a \text{Conc}(e) \land \text{CA}_a \phi \text{ for all } \phi \in \text{Supp}(e)
\]

**Definition 6.14.** A challenge\((i, j, e, f)\) by agent \(i\) of a previous argument \(f\) of agent \(j\) using argument \(e\) requires that \(f\) was previously presented by \(j\) or that \(f\) was used by \(j\) in a challenge against argument \(g\), that \(e\) is assertable for \(i\), and that \(e\) defeats \(f\). This action causes the conclusion and the support of \(e\) to be added to \(i\)’s commitment store:

\[
\text{Prec}(\text{challenge}(i, j, e, f)) \overset{\text{def}}{=} [\text{present}(j, i, f) \text{ has been done} \lor \text{challenge}(j, i, f, g) \text{ has been done}] \land \text{Assertable}(i, e) \land e \text{ defeats } f
\]

\[
\text{Eff}(\text{challenge}(i, j, e, f)) \overset{\text{def}}{=} C_i := C_i \cup \{\text{Conc}(e)\} \cup \text{Supp}(e)
\]

The effects of communicative actions also include “side-effects” that determine whether an action contributes to or violates relational goals, represented by changes of the agents’ mental attitudes. For example, a failed or tedious grounding process would be represented as socially-disruptive (PREM7). Therefore, arguments that are acceptable and hence compatible with common ground are more likely exchanged (IMPL2). Side-effects are not detailed here but referred to later.

### 6.3.2 The Grounding Convention

In the following, let \(a\) denote Alice and \(b\) denote Bob. As defined above, common ground of \(\phi\) between Alice and Bob, represented by the epistemic goal \(\text{Ground}(a, b, \phi)\), can be established by a subsequent execution of the communicative actions \(\text{present}(a, b, e)\) and
6.3. Grounding

approve\((b, a, e)\) where the first one is performed by Alice and the second one by Bob, and \(e\) is an argument \(e = \langle \{\phi\}, \phi \rangle\). We further assume that this can be achieved by a SharedPlan \(P_G\):

\[SP(P_G, \{a, b\}, \text{Ground}(a, b, \phi)).\]

whose recipe specifies that first the present-action has to be executed by Alice and then the approve-action by Bob.

To initiate such a SharedPlan, Alice needs to adopt the following intention-that:

\[I_aDo_{(a, b)}\text{Ground}(a, b, \phi).\]

In order to make this a complete SharedPlan, Bob also needs to adopt this intention and further (mutual) beliefs need to be created (see Section 6.2). We assume that \(P_G\) is a convention and fully specified once it is initiated. We represent this by a communication convention which ensures that a full SharedPlan \(P_G\) is instantiated between agents \(i\) and \(j\) immediately when an agent \(i\) produces a communicative action present\((i, j, e)\):\(^8\)

\[\forall (i, j, e)[Do, present(i, j, e) \rightarrow SP(P_G, \{i, j\}, \text{Ground}(i, j, \text{Conc}(e))))]\]

6.3.3 Prerequisites

Let \(\rho_F\) stand for the stereotype of careless and beer-drinking football players:

\[\rho_F \overset{\text{def}}{=} \forall i[FPlayer(i) \rightarrow (\text{Careless}(i) \land \text{Prefers}(i, \text{Beer}))]\]

and \(\rho_B\) for the stereotype of beer-only drunk-drivers:

\[\rho_B \overset{\text{def}}{=} \forall i[\text{Prefers}(i, \text{Beer}) \rightarrow \neg(\text{Drank}(i, \text{Wine}) \land \text{DrankDrove}(i))].\]

\(^8\)Note that we lack the appropriate precision here because of our decision to ignore the temporal aspect of SharedPlans.
Let $E$ stand for the group including everyone. Assume the following initial common ground and initial beliefs of Alice and Bob:

$$
CA_E \rho_F \land CA_E \rho_B
$$

$$
B_a [\text{FPlayer}(\text{Gary}) \land \text{GotCaught}(\text{Gary})
\land \text{Drank}(\text{Gary}, \text{Wine}) \land \text{DrankDrove}(\text{Gary})]
$$

$$
B_a \forall i [\text{Careless}(i) \rightarrow \neg \text{Dedicated}(i)]
$$

$$
B_a \text{HadMatch}(\text{Gary})
$$

$$
B_b \text{HadMatch}(\text{Gary})
$$

$$
B_b \forall i [\text{FPlayer}(i) \land \text{HadMatch}(i) \rightarrow \text{Dedicated}(i)]
$$

$$
B_b \forall i [\text{FPlayer}(i) \land \text{Dedicated}(i) \rightarrow \{\text{HadMatch}(i) \rightarrow \neg \text{GotCaught}(i)\}]
$$

The two stereotypes are part of communal common ground shared by everyone. Let us assume that the membership of Alice and Bob in the group that includes everyone is salient in their interaction. Alice believes the information about Gary’s incident and she believes that someone who is careless is not dedicated. She also believes that Gary had a match. Bob also believes that Gary had a match and he moreover believes that a football player that had a match would have been dedicated. He also believes that a dedicated football player would not get caught if he had a match.

From $CA_E \rho_F$, $CA_E \rho_B$, and SCG (see Section 6.1.2) follow $CG_a \rho_F$ and $CG_a \rho_B$ respectively. That is, the two stereotypes are also part of salient common ground between Alice and Bob in the context of the joint action $\alpha$.

### 6.3.4 The Dialogue

We assume Alice adopts the epistemic goal $\text{Ground}(a, b, \varphi)$ to ground the essence of the story about Gary, which is dictated by a higher-level recipe, with:

$$
\varphi \triangleq \text{FPlayer}(\text{Gary}) \land \text{DrankDrove}(\text{Gary}) \land \text{GotCaught}(\text{Gary}).
$$

For all $\psi$ such that $B_a \text{Substitutable}(\varphi, \psi, P_a)$, Alice adopts the following potential intentions to ground proposition $\psi$: $P_a \text{Do}_{(a,b)}(\text{Ground}(a, b, \psi))$. From among those potential intentions, Alice commits to $P_a \text{Do}_{(a,b)}(\text{Ground}(a, b, \varphi))$ whose achievement she deems compatible with her relational goals in the sense that it does not disagree with hers and Bob’s common ground. Amongst others, she discards this alternative option:

$$
\psi \triangleq \text{FPlayer}(\text{Gary}) \land \text{Drank}(\text{Gary}, \text{Wine})
\land \text{DrankDrove}(\text{Gary}) \land \text{GotCaught}(\text{Gary}),
$$
which contradicts with the stereotypes $\rho_F$ and $\rho_B$. $\text{Ground}(a, b, \phi)$ is a complex action, which Alice intends her and Bob to execute together:

$$ I_a \text{Do}_{(a,b)} \text{Ground}(a, b, \phi). $$

As described above, this can be achieved by a SharedPlan $P_G$. Because $\text{Ground}(a, b, \phi)$ is part of the recipe of a higher-level SharedPlan that describes the joint activity, $P_G$ is effectively a subordinate activity of the joint activity.

Alice issues the present-action with the argument $e = \langle \{\phi\}, \phi \rangle$ in utterance (1). This augments her commitments as follows:

$$ C_a := C_a \cup \{F\text{Player}(\text{Gary}) \land \text{DrankDrove}(\text{Gary}) \land \text{GotCaught}(\text{Gary})\} $$

Alice’s performance of the present-action also evokes the communication convention, which creates the SharedPlan $P_G$. Due to Equation 6.4 both Alice and Bob are now committed to $P_G$ and in particular to Bob being able to perform the approve-action. The commitment to Bob being able to perform the approve-action represents the cooperative nature of this dialogue: Alice and Bob intend to arrive at a common understanding of matters. While an immediate approval by Bob would complete $P_G$ and add $\phi$ (Alice’s presentation of the story about Gary) to common ground, it turns out that $\phi$ is not acceptable by Bob. In utterance (2), he uses a challenge to present an assertable argument that rebuts Alice’s presentation:

$$ \langle\{\text{FPlayer}(\text{Gary}), \text{HadMatch}(\text{Gary}), \text{FPlayer}(x) \land \text{HadMatch}(x) \rightarrow \text{Dedicated}(x) \}, \forall x[\text{FPlayer}(x) \land \text{Dedicated}(x) \rightarrow \{\text{HadMatch}(x) \rightarrow \neg \text{GotCaught}(x)\}]\rangle $$

In creating this argument, Bob makes use of Alice’s commitments ($F\text{Player}(\text{Gary})$) as well as his beliefs about football players being dedicated when they have a match. This adds the content of Bob’s argument to his commitments. Bob’s utterance advances $P_G$ by informing Alice that and why Bob cannot perform the approve-action:

$$ B_a B_b \neg \text{Acceptable}(b, e). $$

Because acceptability is a precondition of the approve-action, Alice reasons that:

$$ B_a B_b [\neg \exists R_{\text{acc}} CBA(b, \text{approve}(b, a, \phi), R_{\text{acc}})]. $$
That is, Alice comes to believe that Bob believes that he is unable to execute the approve-action. However, both Alice and Bob are committed to Bob’s action. To support Bob in approving her original presentation, in utterance (3) Alice challenges Bob’s argument by undercutting it and reinstating her initial presentation with another argument:

$$\langle \{FPlayer(Gary), \rho_F, \forall i [Careless(i) \rightarrow \neg Dedicated(i)]\}, \neg Dedicated(Gary) \rangle.$$  

Here, Alice exploits her own beliefs about being careless and dedicated as well as the salient common ground between Bob and her ($\rho_F^G$). Again, her commitments are augmented.

Bob then performs the approve-action in utterance (4), indicating that Alice’s original presentation is now an acceptable argument. This adds Alice’s original presentation of the story about Gary to Bob’s commitments and in fact to their context-specific common ground. We can also assume that Alice and Bob are both committed to Alice’s last statement, which remains unchallenged. They should not be committed to $Dedicated(Gary)$ and $\neg GotCaught(Gary)$ in Bob’s challenge in utterance (2) because these propositions have been defeated without reinstatement. So this sub-dialogue is terminated. However, it is not obvious which of the remaining commitments should become common ground as well.

A change of context-specific common ground does not by itself reflect a change of culture. Cultural change is driven by changes in communal common ground. Let us consider how this would be represented in our model (PREM9, IMPL3). Assume that an agent $i$ from group $G$ repeatedly experiences others from that group assert that $CA_G \phi$. This is in fact the case in utterance (3) of the sample dialogue when Alice asserts that everybody knows the stereotype about football players. Given this repeated experience, agent $i$ can develop the belief that indeed everyone in that group accepts that $\phi$ is common ground of that group and that this is mutually believed (the premises of Axiom CA, see Section 6.1.2). Because of this belief, agent $i$ might at some point itself decide to accept that $\phi$ is common ground in the group (i.e. $A_G^i \phi$). It is not obvious under which conditions this transition would be made. However, this observation points to the role of the perceived common ground in a group, i.e. the beliefs of group members about the common ground of their group.

6.4 Discussion and Conclusions

In this chapter, we have described a formalization of common ground and a computational-conceptual account of its interrelation with joint activities and the grounding processes that these activities evoke. The model introduced in this chapter is more faithful to the grounding model of cultural transmission than the models introduced in previous chapters. Thereby, this model contributes to RQ1. The model is distinguished from previous formalizations of
6.4. Discussion and Conclusions

dialogue, common ground, and grounding by its close adherence to the grounding model of cultural transmission. On the nano-to-macro axis, this model covers the nano and micro level. The computational complexity of this model surpasses the one of the models presented previously, due to the richer cognitive mechanisms it requires from the agents.

A few remarks are expedient: Despite its detail, our model is still an idealization of the grounding model of cultural transmission. However, the models of cultural transmission in the last two chapters were even more so; and idealization is not a flaw of models but a feature as discussed in Chapter 3. We did not provide a complete formal dialogue protocol of the grounding process. Further work is necessary to explicate the formalization and to sort out the characteristics of different types of dialogues that might occur during a joint activity. We did not analyze in detail the requirements on the agents’ internal reasoning. An interesting point to consider is that agents might still communicate information that violates relational goals, or they might opt for committing to unacceptable statements in order not to violate a relational goal. This is certainly something that humans might consider doing. A decision-theoretic approach appears suitable for describing the agents’ weighing of these alternatives. The question remains to what extent the SharedPlan formalism is an appropriate description of human joint activities. For now, the SharedPlans formalism offers the benefit that it can explain how grounding processes appear naturally as the result of joint activities.

Despite these limitations, we believe that our contribution is valuable in two regards. First, our formalization can be used as a starting point for other and potentially more complete formal studies of cultural dynamics. In that, our model can serve as an auxiliary model for other attempts at formalizing cultural transmission and dynamics. Then the model adopts the role of a framework that is evaluated rigorously only through its concrete instantiations. Second, by working towards a comprehensive formalization of the grounding model of cultural transmission we have revealed aspects of that model that require further discussion. In fact, we have provided first answers to these new questions. This is the contribution of this chapter towards RQ2. We name a few of these aspects in the following.

Our analysis has shown that there are many subtleties involved in defining common ground. We have opted for a definition in terms of collective acceptance. This deviates from the original description of the grounding model of cultural transmission. However, the insights we have gained through our analysis might be carried over to the original model. We have also described axioms that we believe describe best the rules that govern the emergence of common ground and its composition in joint activities. Discussion is particularly necessary with respect to the axioms that describe how common ground emerges bottom-up from individual mental attitudes and how it affects joint activities top-down.

Our formal analysis has also revealed that the term “perceived common ground” and its relation to actual common ground has been used in ambiguous ways within the context of
the grounding model of cultural transmission. Making this notion more precise is important because it is perceived common ground that is accessible for individuals, not actual common ground. Our work has also raised the question when commitments incurred in a dialogue become actually part of common ground. In our sample dialogue, not only primary information is exchanged but also secondary information that arises during the grounding process. We have made particular assumptions but there are certainly opportunities for further empirical studies that can shed light on these questions. We have also discussed that it is not clear how precisely communal common ground is modified during joint activities. Yet as discussed previously it is the change of communal common ground that reflects the change of culture.

In addition, our analysis has led us to introduce the term “salient common ground” to distinguish more carefully between the common ground that actually holds in a situation and context-specific common ground, which is manipulated during a social interaction.

As other similar models (e.g. Gaudou et al., 2006b) our model has not been put to the test against actual dialogue data. An interesting prediction of this model that could be tested empirically is that if group membership salience is manipulated, salient common ground is manipulated as well. Such an empirical study could also shed more light on the composition operator that yields salient common ground. For example, one could hypothesize that normally different group memberships are salient to a different extent. If that was the case, some communal common grounds would take precedence over others in the designation of salient common ground.
Chapter 7

A Minimal Computational Architecture of Joint Action

As outlined in Section 3.4, this chapter presents a computational-conceptual model (terminology introduced in Section 3.2.2) of the interaction between higher- and lower-level coordination mechanisms in joint action. As we shall see, this interaction plays an important role in cultural transmission. To some extent we expand on the previous chapter, which has introduced a computational-conceptual model of the interplay between a joint action and the coactors’ task-incidental dialogue (grounding). From the perspective of this chapter, grounding is a higher-level coordination mechanism that enables coactors to align their information about the joint action. Grounding is a by-product of the coactors’ joint intentions and of their collaborative, goal-oriented planning. This analysis stands in the tradition of philosophical approaches to the study of practical reasoning and joint intentions based on higher-level mental attitudes such as beliefs, goals, and intentions. We have mentioned this research throughout this thesis a few times, most notably in Sections 2.2 and 6.2. This approach is useful in explaining how coactors coordinate by collaboratively planning their actions, including the use of symbolic communication.

Recent research from social and ecological psychology, on the other hand, has emphasized that lower-level mechanisms, which are not mediated by any higher-level cognition, enable basic forms of coordination in joint action (refer to Knoblich et al., 2011, for a comprehensive review). In fact, higher-level coordination mechanisms can be supported by lower-level ones. For example, internal predictive motor models allow actors to infer the intentions of coactors, which serves collaborative planning (Knoblich and Sebanz, 2008). Coactors also align their communication at multiple linguistic and non-linguistic levels, e.g. by using the same words or similar grammatical structures, which supports that alignment
Chapter 7. A Minimal Computational Architecture of Joint Action

of information which we have discussed in the previous chapter as the grounding process (Garrod and Pickering, 2009).

So far there is no satisfactory account showing how theories operating at either of these two levels could account for the phenomena explained by the theories operating at the other level. Therefore, a full understanding of joint action depends on an integration of these two perspectives. A few such attempts have been made but they leave open a lot of details. In the spirit of this thesis, we believe that a computational approach can contribute to the discussion by (a) providing a more precise model of joint action in a language suitable for describing the mechanisms involved, (b) directing discussion and empirical research, and (c) enabling the development of computational simulations of joint action to complement empirical research.

The goal of this chapter is to propose a computational-conceptual architecture of joint action that accounts for the interplay between higher- and lower-level coordination mechanisms. While the former mechanisms rely on practical reasoning based on higher-level mental attitudes, the latter mechanisms are independent of any higher-level cognition. More specifically, we provide a computational-conceptual model of joint action after extracting a minimal set of requirements from related literature and an analysis of a particular experimental task—the social Simon task (Sebanz et al., 2003). The next chapter provides a concrete (and simplified) implementation of this architecture to reproduce the human performance in the social Simon task.

In the social Simon task, two participants share a simple task: Each participant is instructed to act on the presentation of a stimulus with a particular attribute (e.g. color or location) by pushing a button in front of them. Surprisingly, an increase in reaction time is observed if the participant has been instructed about the other participant’s task and if the presented stimulus at the same time requires a response by the participant and also relates to the other’s task. This interference effect suggests that people represent others’ tasks and actually have a natural tendency to do so even if this is not necessary for a successful performance of the joint action. Moreover, the interference effect is partly attributed to lower-level coordination mechanisms as we shall explain in detail later. Therefore, higher-level representations of the joint action appear to interact with lower-level coordination mechanisms in this task. Being able to model tasks such as the social Simon task is consequently an essential step towards a computational architecture of joint action that spans higher and lower levels of coordination, which is the goal of this chapter.

Within the context of this thesis, the architecture described in this chapter is to some degree an extension of the one introduced in the previous chapter. As we shall see, we basically retain the idea of the higher-level planning component from the previous chapter and add lower-level coordination mechanisms and their interactions with the higher-level
7.1 Perspectives on Joint Action

This section provides an overview of the different relevant perspectives on the study of coordination in joint action. The first perspective to be discussed is the philosophical one of practical reasoning and joint intentions. This follows on from the context assumed in the previous chapter. Joint intentions are considered to be those mental constructs that link the coactors’ intentions and practical reasoning to each other’s actions and to the overall joint action. Practical reasoning and joint intentions provide higher-level mechanisms that enable more complex collaborative planning based on symbolic information. Joint intentions are assumed to ensure that coactors are committed to their joint action and hence drive collaborative planning towards the goal of the joint action.

The second perspective is the one taken by ecological and social psychology, which focuses on more basic, lower-level forms of coordination. The social Simon task is among the tools used by psychology to study this type of coordination. From this psychological

component. In that, the architecture presented here is also located at the nano and micro levels of the nano-to-macro axis.

The relevance to the topic of this thesis—computational representations of the grounding model of cultural transmission—is two-fold: First, lower-level coordination mechanisms, which were not considered in the previous chapter, play a key role in the alignment of mental models (grounding) and hence in the transmission of cultural information. Second, lower-level coordination mechanisms enable the transmission of cultural information without symbolic communication or explicit reasoning. We expand on these points later in this chapter. This chapter contributes to RQ1 by providing a more comprehensive model of joint actions, which are a core component of the grounding model of cultural transmission. The chapter addresses RQ2 by suggesting factors of joint action that affect cultural transmission but have previously not been considered within the context of the grounding model of cultural transmission.

We begin this chapter in Section 7.1 with a more extensive discussion of joint action and the different levels of coordination in joint action. We describe the social Simon task and associated experimental results in Section 7.2. In Section 7.3, we follow on by extracting the requirements of a computational architecture of joint action that is able to represent the social Simon task. Section 7.4 describes an architecture that fulfills these requirements and Section 7.5 discusses how it is able to reproduce the human performance in the social Simon task. We close this chapter with a discussion and summary in Section 7.6. We delay a detailed discussion of related work to the end of the next chapter when we have described the architecture and its implementation.
perspective, Knoblich et al. (2011) distinguish between phenomena of planned and emergent coordination and they describe some of the mechanisms responsible for these phenomena. Planned coordination is enabled by representations of the coactors’ tasks in the joint action. This allows, for example, to predict how a coactor will react to a particular stimulus. Emergent coordination in contrast rests on low-level mechanisms that do not involve any higher-level cognition. These mechanisms generally cause coactors to perform similar actions. Planned and emergent coordination are generally enabled by procedural knowledge.

In the following subsections, we discuss practical reasoning and joint intentions, planned coordination, and emergent coordination. The reader is advised to bear in mind that the first perspective describes a mechanism while planned and emergent coordination are seen as phenomena that are but enabled by certain mechanisms. Of course, we emphasize in our discussion the mechanisms of planned and emergent coordination. In the last subsection, we discuss briefly how the mechanisms postulated by the philosophical perspective could be reconciled with the mechanisms postulated by the psychological perspective. In doing so, we rely on the obvious overlap between the mechanism of practical reasoning and joint intentions and the mechanisms of planned coordination: Joint intentions naturally presume some form of corepresentation of the coactors’ tasks, which is also assumed to play a key role in planned coordination (Knoblich et al., 2011).

7.1.1 Practical Reasoning and Joint Intentions

Much philosophical research has been conducted on practical reasoning in individual action. We have briefly described the widely accepted perspective by Bratman (1987), which rests on the mental attitudes goals, intentions, and beliefs, in Section 6.2.

The study of joint planning and execution in joint action stands in this tradition. The focus has been on the mental attitudes and in particular intentions that are required to drive planning and execution towards a joint goal. It is generally accepted that some form of a joint intention is required but it is debated whether joint intentions are reducible to individual mental attitudes or not (Bratman, 1992; Gilbert, 1992; Searle, 1990; Tuomela, 2000b). Refer to Pettit and Schweikard (2006) for a more in-depth discussion of these issues. Regardless of this ontological debate, the mental attitudes of actors in joint actions need to be appropriately interlocked and make reference to each other (Castelfranchi, 2006). For a joint action to be robust against the failure of individual contributions, for example, coactors need to have appropriate intentions towards the success of their peers’ tasks to ensure mutual support.

For our purpose, the ontological status of joint intentions is not important. More relevant are the consequences we assume joint intentions to have on the behavior of the coactors that hold them. Bratman (1992) identifies the following necessary characteristics of joint action:
7.1. Perspectives on Joint Action

**Mutual responsiveness** Coactors are trying to be responsive to each other’s intentions and actions while knowing that the other party is doing so as well.

**Commitment to the joint action** Coactors are committed to the joint action, which causes their mutual responsiveness. The reasons why the agents are committed to the joint action, however, do not need to be the same. In particular, the commitment to a joint action is supported by an individual intention to bring about a certain state of affairs.

**Commitment to mutual support** The agents are committed to helping each other in order to complete the joint action successfully.

From the philosophical perspective, language is simply understood as an extension of the action repertoire of the coactors. Hence, communication is a product of reasoning as are any other actions, and in fact communication is a form of joint action itself (Clark, 1996c; Garrod and Pickering, 2009). This is the perspective taken in the previous chapter.

Philosophical research of joint action is also closely (though not necessarily exclusively) tied to the *theory theory* version of the *theory of mind*—the idea that people use an explicit mental model of human reasoning to infer the mental states of their peers (see Flavell, 2004, for a review). The mental model is assumed to play a similar role in reasoning as mental models of the physical world. This mental model of reasoning is conceptualized as a set of causal links between the mental states attributed to others. Obviously, these mental states can be the same ones we have talked about in this section. There is disagreement about the origin of this mental model, whether it is an innate capability or whether it is acquired during childhood (Gallagher, 2001). This debate is only of marginal relevance to this thesis.

Research on practical reasoning and joint intentions attempts to answer the question by which means coactors are committed to their joint action and how they jointly plan their actions towards a common goal. However, the intricate details of the coordination of motor actions, for example the ones necessary to hand an object from one person’s to the other person’s hand, are not considered by this perspective. Also, philosophers are typically not concerned with the question under which situational conditions coactors create a mental construct representing a joint intention.

### 7.1.2 Planned Coordination

Planned coordination is the first of two types of phenomena of coordination in joint action considered by social and ecological psychology (Knoblich et al., 2011). Planned coordination arises from coactors planning their actions with the actions of each other in mind. Vesper et al. (2010) propose that an actor in a joint action represents at least the goal of the joint action, its own part of the task, and the fact that some part of the task is the responsibility
of another actor or (physical) force. The details of the other’s part of the task are only represented when this is deemed necessary. Empirical research, however, indicates that coactors are prone to developing shared task representations, i.e. they do tend to represent their partners’ parts of the joint action (Sebanz et al., 2003, 2005, 2006b).

Along with the representation of the task, an actor might represent attributes of its partner (Vesper et al., 2010) or their social relationship. This information can affect whether and how a partner’s task is represented. It has been observed, for example, that a shared task representation is not exercised when the coactor is perceived as unfriendly in a condition that usually evokes the creation of a shared task representation (Hommel et al., 2009). Similarly, one can expect that the trust in the coactor’s intention and competence to achieve their part of the task affects whether their task is corepresented or not. We will return to this observation.

Another mechanism by which representations can be shared is joint attention or joint perception (Eilan et al., 2004). Sharing the focus of attention allows coactors to share their representations of objects and events, which creates a kind of perceptual common ground (Sebanz et al., 2006a). Knowing what others attend to allows individuals to some extent to infer which information others do have or do not have and what their current intentions are. This insight can be used to support the coactor, e.g. by guiding their attention to a particular object or by communicating information assumed to be unknown. Central to joint attention is the ability to distinguish between one’s own and the other’s perceptions (Knoblich and Sebanz, 2008).

Planned coordination is based on shared mental representations, which enable phenomena of coordination that are more basic than the ones supported by practical reasoning and collaborative planning. Research on planned coordination has provided insights into the conditions under which coactors share mental representations. However, social and ecological psychologists by and large remain silent about how these shared representations might be used to enable collaborative planning. Also, planned coordination does not consider phenomena such as the synchronization of movements discussed in the following in the context of emergent coordination.

7.1.3 Emergent Coordination

Precise coordination of actions in time and space in real-time is the domain of emergent coordination (Knoblich et al., 2011). Mechanisms enabling emergent coordination rest on direct links between perception and action, which complement slow and computationally expensive cognitive mechanisms.

Ecological psychology has put forward an embodied cognition perspective to joint action (Marsh et al., 2009). According to the embodied cognition view, perception and action enable and constrain each other mutually without the involvement of any intermediary cognitive
mechanisms (Clark, 1996a). The actor and its environment is conceptualized as a dynamical system. The same principles are assumed to determine the interactions between an actor and its environment and between coactors. Coactors can become coupled in two different ways (Knoblich and Sebanz, 2008): First, the behavior of people can become entrained when their actions and the perceptions of each other’s actions affect each other. For example, people tend to synchronize rhythmic limb movements when they perceive each other even if they have not been told to do so (Richardson et al., 2005). The synchronization of rhythmic movements in this and other instances has been shown to follow general mathematical principles of dynamical systems (Schmidt and Richardson, 2008). Second, coactors can become coupled if objects in the environment afford similar action opportunities to them. Action opportunities or so called affordances (Gibson, 1979) are assumed to be encoded in the environment itself. Likewise, intentions are considered to be part of the environment and not part of the minds of the actors (Shaw, 2001).

Research on entrainment and common affordances has focused on phenomena rather than the mechanisms that enable them. More specific explanations are offered by the hypothesis that perception and action share the same representations because actions are planned and controlled in terms of their effects. As a consequence of this, the same representations are activated both during action planning and during the observation of others’ performing this action. This hypothesis is known as the common coding theory developed by Prinz (1990, 1997) as an extension to James’ ideomotor theory (1890). Evidence for this hypothesis comes both from neuroscientific and behavioral studies (Knoblich and Sebanz, 2008; Sebanz and Knoblich, 2009).

Neuroscientific work has shown that some of the neurons in the motor area of the macaque monkey cortex are active when the monkey performs a particular object-directed action (e.g. grasping) and when the monkey observes a conspecific or an experimenter performing that action. This observation was first made by Di Pellegrino et al. (1992) and the respective part of the cortex was termed the mirror neuron system. Mirror neurons have been shown to be capable of generalization: Their activation during action observation is independent of factors such as the distance to the participant that performs the observed action or the object that this action is directed at (Rizzolatti and Craighero, 2004). It was later proposed that a mirror neuron system is also present in the human brain (Gallese et al., 1996) and further investigations have provided support for this hypothesis (see reviews in Glenberg, 2011; Rizzolatti and Craighero, 2004; Rizzolatti and Sinigaglia, 2010). Interestingly, there is evidence that mirror neurons in the human brain also react to actions that are not object-oriented (see discussion in Rizzolatti and Craighero, 2004; Rizzolatti and Sinigaglia, 2010). Various functional roles have been attributed (at least partly) to the mirror neuron system, from action understanding to empathy to language comprehension to imitation (Glenberg,
However, some question the validity of the evidence that has been put forward in support of these functions (e.g. Dinstein et al., 2008; Hickok and Hauser, 2010). Despite the debate surrounding the functional role of the mirror neuron system, its discovery provides convincing evidence for a common representational system for action and perception.

Behavioral studies have shown that observing an action interferes with action execution (for a brief overview see Knoblich and Sebanz, 2008). Perceiving an action causes an involuntary tendency in the observer to perform the observed action and essentially facilitates the execution of that action. Performing an action opposite to the observed one is more difficult and suppressing an involuntary action tendency requires modulation from a higher level. These observations suggest that a common representational system is used both during action planning and perception. From a functional perspective, perceiving an action allows an individual to simulate the performance of that action, a process termed action simulation (Knoblich and Sebanz, 2008). Action simulation provides an interface to higher-level intentional mechanisms: Simulating an observed action can be used to predict the outcome of that action and infer the goals and intentions of the actor (Fogassi et al., 2005; Sebanz and Knoblich, 2009; Wolpert et al., 2003). Hence, action simulation is a mechanism that facilitates imitation and synchronization.

The summary of these results is that actors can use their own action repertoire to make sense of others’ behavior in joint actions. In effect, this helps to establish a procedural common ground in joint action (Knoblich and Sebanz, 2008). This is compatible with the so-called simulation theory version of the theory of mind, which postulates that people can use their own cognitive mechanisms to infer mental states of their peers (see Shanton and Goldman, 2010, for a review). Simulation theory competes with theory theory, which we discussed in Section 7.1.1, because it rejects that humans have an explicit mental model of others’ minds. According to simulation theory, individuals construct pretend mental attitudes to describe another person’s mental state and then use these attitudes as an input to their own cognitive mechanisms. The output of this simulation is attributed to the other person. In effect, the person takes the perspective of the other person and thereby simulates what they themselves would do or feel or think if they were in the situation of that other person. This simulation does not necessarily need to be a conscious process but might execute sub-consciously as well. Some hybrid theories of mind reading that draw on both simulation theory and theory theory have been proposed (Shanton and Goldman, 2010).

The results outlined in this subsection show that some forms of coordination can arise even without the involvement of higher-level mental representations. While these approaches provide explanations for low-level coordination, it is unclear how they can explain more complex coordination that requires coactors to take into account each other’s parts of the joint action.
7.1.4 Synthesis

In this subsection, we discuss how the philosophical and the psychological perspectives of joint action relate to each other. The picture we advance is one in which practical reasoning drives the creation of joint intentions, whose parts in turn specify the contributions of individual actors and thereby constitute shared task representations. This gives rise to planned coordination. In turn, emergent and planned coordination support higher-level coordination driven by practical reasoning and joint intentions.

Let us first summarize which coordination phenomena can be accounted for by practical reasoning and joint intentions and which ones can be accounted for by the mechanisms of planned and emergent coordination. Mechanisms of emergent coordination can explain how coactors come to execute similar actions but they cannot easily explain how coactors take each other’s parts into consideration during action planning. Mechanisms of planned coordination explain how information relevant to coactors is represented and considered in planning. However, psychology has generally not taken planned coordination beyond phenomena of basic coordination. For example, planned coordination has not been evoked to explain how coactors become committed to the joint action and how they collaboratively plan towards a joint goal. Similarly, those mechanisms of planned coordination by which an actor relates different parts of the joint action to the overall task are not explicated.

This is the point where philosophical research on practical reasoning and joint intentions comes in. Philosophical approaches explain how practical reasoning causes the instantiation of joint intentions. And joint intentions do specify how joint planning is driven by the commitment of the actors to their joint action and how different parts of the joint action relate to each other. We believe that practical reasoning on joint intentions can be seen as a mechanism of planned coordination that enables more complex phenomena of planned coordination. In doing so, joint intentions rely on or rather constitute the shared task representations of planned coordination. As discussed in Section 6.2, philosophers have proposed a hierarchical organization of goals and intentions in which means and ends alternate. This suggestion has been picked up by some psychologists as well (e.g. Tomasello et al., 2005). If we assume that joint intentions follow the same structure as individual intentions (hierarchies of goals and plans), they neatly integrate into the practical reasoning of individual actors. This also yields a structure that allows actors to keep the contributions of different coactors apart mentally (Knoblich and Sebanz, 2008; Tomasello et al., 2005) and to relate individual contributions to the overall joint action (Knoblich and Sebanz, 2008) and to their own intentions. These capabilities and the ability to infer others’ intentions from their actions are essential for the engagement in joint intentions (Knoblich and Sebanz, 2008).
Chapter 7. A Minimal Computational Architecture of Joint Action

The role of emergent coordination remains to explain the precise coordination of actions in time and space that neither practical reasoning on joint intentions nor other mechanisms of planned coordination can account for. With planned and emergent coordination in place, their mechanisms can be employed to support higher-level planning as discussed previously. Even by itself, emergent coordination enables some of the mutual responsiveness that is required by joint actions according to Bratman (1992).

This perspective suggests that bringing practical reasoning and joint intentions, planned coordination, and emergent coordination together would facilitate a more holistic understanding of joint action. The idea of integrating practical reasoning and joint intentions, planned coordination, and emergent coordination is not novel. Yet by specifying a computational architecture of these interactions, we take one step more than previous approaches to this problem. We also put more emphasis on the role of practical reasoning and joint intentions in joint action than was done previously.

Knoblich and Sebanz (2008) discuss how entrainment, common affordances, action simulation, and joint perception support joint intentions and are controlled by joint intentions:

- Entrainment allows coactors who have the intention to synchronize their actions (e.g. musicians) to execute the same actions at the same time.

- Common affordances turn into joint affordances when a joint goal requires coactors to perform different actions on an object at the same time.

- Because joint intentions enable the representation of intentions and actions separately by actor, action simulation can be used to predict and plan one’s own as well as others’ actions.

- Joint perception enables coactors to infer which information is available to their coactors and what their current intentions are within the context of the joint action.

Vesper et al. (2010) propose an architecture of joint action that bridges the gap between emergent and planned coordination. They assume that coactors represent the goal of the joint action, their own part of the task, and possibly the part of the task that is the responsibility of another actor or force. Joint action is supported by prediction and monitoring processes operating on some or all of these representations. Monitoring processes keep track of the progress towards the goal or individual tasks or the execution of individual actions. Predictions can be made with regard to how the coactors’ combined actions achieve the joint goal and with regard to the effects of individual actions. Prediction is based on action simulation using internal predictive models. In addition, so called coordination smoothers facilitate the joint action, e.g. by rendering one’s own actions more predictable for coactors.
7.2 The Social Simon Task

In this section, we describe the social Simon task and associated empirical findings. Together with the review in the previous section, the following discussion paves the way for the identification of the minimum requirements for a computational architecture conducted in the next section.

The social Simon task is an experimental paradigm which has revealed that a participant’s performance in a joint task is impaired by a coactor’s part of the task. This suggests that people construct and employ shared task representations in joint tasks, which lead to an interference between higher- and lower-level coordination mechanisms. First, we describe briefly the individual Simon task, on which the social Simon task is based.

In the individual Simon task, participants attend to a screen and push one of two buttons (left or right) depending on a non-spatial attribute of an object appearing on the screen. The object on the screen is called the stimulus and the non-spatial attribute is typically the color of this object. We also call this attribute the task-relevant attribute. For example, a participant could be instructed to push the left button when the object is green and the right button when the object is red. It turns out that reaction time increases if spatial, task-irrelevant attributes of the object are incompatible with spatial attributes of the expected response (push left or right button). The spatial, task-irrelevant attribute is typically the location of the object on the screen. For example, the object can appear on the left or right side of the screen and be either on the same side as the button that is to be pushed according to the task-relevant attribute of the stimulus or on the other side. In the first case, we talk about stimulus-response compatibility and about stimulus-response incompatibility in the second case. Figure 7.1 displays these two conditions. The increase in reaction time in the stimulus-response incompatibility condition is called the Simon effect because Simon and Rudell (1967) were the first ones to observe this effect. The Simon effect is not observed...
Chapter 7. A Minimal Computational Architecture of Joint Action

Figure 7.1: Stimulus-response compatibility and incompatibility in the Simon task. In (a), the stimulus appears on the side of the button that is to be pushed. In (b), the stimulus appears on the other side. Note that buttons are not colored in the actual task.

when the participant performs only one part of the task, e.g. when the participant is instructed to push the left button when the stimulus object is green without receiving any instruction with respect to the right button.

Various explanations for the Simon effect have been proposed (e.g. Hommel, 1993; Hommel et al., 2001; Nicoletti and Umiltà, 1994; Sheliga et al., 1997). It is beyond the scope of this thesis to provide a comprehensive discussion. We side with the following explanation offered by Hommel et al. (2001) because it is compatible with common coding theory, which we accept as a description of a key mechanism of emergent coordination.

Based on the common coding theory of Prinz (1990), Hommel et al. (2001) assume that stimuli and action effects share the same representational domain. In particular, stimuli and action effects are assumed to be represented as compositions of so called feature codes or simply features (see Treisman, 1988, for an early discussion around the concept of features and Hommel, 2010, for a more recent one). Feature codes are considered not to represent proximal attributes of stimuli and action effects but more distal ones. Essentially, distal attributes are abstractions of proximal attributes (see Hommel, 2009, for a discussion of the difference between proximal and distal). Perceiving a stimulus leads to the activation of a set of features while planning an action amounts to activating the features that constitute that action’s effects. This shared use of features allows for stimuli to interfere with action planning because features activated due to a stimulus can be shared with an action. In the Simon task, the task-irrelevant spatial attribute of the stimulus (left or right) can lead to the activation of a feature that is shared with the action that is not the appropriate response (push left or push right) given the task-relevant stimulus attribute. In this case, this inappropriate action receives activation, which needs to be explicitly suppressed so that the correct action is executed. This action conflict as Sebanz et al. (2005) call it causes the Simon effect, i.e. an increase in reaction time. Table 7.1 summarizes the setup for the two conditions in the
7.2. The Social Simon Task

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Setup</th>
<th>No conflict</th>
<th>Action Conflict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task rules</td>
<td>green → left</td>
<td>green → left</td>
<td>green → left</td>
</tr>
<tr>
<td>Stimulus</td>
<td>red/right</td>
<td>red → right</td>
<td>red/lef</td>
</tr>
</tbody>
</table>

Table 7.1: The setup of the two experimental conditions in the individual Simon task consisting of the participant’s task and the stimulus. By green → left we mean that the participant has been instructed to react to a green stimulus by pushing the left button. The parts of each task rule that are activated by the stimulus either because of the relevance to the rule’s precondition or an overlap with the features of its action are presented in bold face.

individual Simon task, showing the combination of task rules and stimulus attributes that constitute these conditions.

In the social Simon task, two participants carry out the original Simon task together, i.e. each participant is responsible for one of the two stimulus-response mappings (task rules). For example, the participant sitting on the left side would push their button only if the stimulus object was green. Surprisingly, a task-irrelevant spatial attribute referring to the other participant’s action leads to an increase in reaction time similar to the one in the individual Simon task (Sebanz et al., 2003). For example, the participant mentioned above would react more slowly if the object appears on the right side of the screen, thus being compatible with the other participant’s action of pushing the right button. Such an increase in reaction time does not occur if the participants carry out their part of the task individually. The interpretation is that participants corepresent their partner’s action in the joint task (action corepresentation). The partner’s action can then be activated by a compatible stimulus feature and therefore cause an action conflict and consequently a Simon effect as in the individual Simon task.

Sebanz et al. (2005) distinguish an action conflict from a task conflict and action corepresentation from task corepresentation respectively. For an increase in reaction time is also observed when a stimulus calls for both participants to carry out an action at the same time (task conflict). The interpretation is that a participant does not only corepresent the action of their partner but also the entire task rule of that partner (task corepresentation). If the participant corepresents their partner’s task rule, the associated action is activated when the stimulus triggers that rule’s precondition. As with the action conflict, a task conflict is not observed when the participant performs its part of the task individually. Sebanz et al. find that action and task corepresentation can affect task performance in isolation as well as simultaneously. Results show that the reaction time increase due to task conflicts is larger
Chapter 7. A Minimal Computational Architecture of Joint Action

<table>
<thead>
<tr>
<th>Setup</th>
<th>No conflict</th>
<th>Action Conflict</th>
<th>Task Conflict</th>
<th>Both conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left participant</td>
<td>green → left</td>
<td>green → left</td>
<td>right → left</td>
<td>left → left</td>
</tr>
<tr>
<td>Right participant</td>
<td>red → right</td>
<td>red → right</td>
<td>red → right</td>
<td>red → right</td>
</tr>
<tr>
<td>Stimulus</td>
<td>red/right</td>
<td>red/right</td>
<td>red/right</td>
<td>red/left</td>
</tr>
</tbody>
</table>

Table 7.2: The setup of the four experimental conditions in the social Simon task consisting of the left participant’s task, the right participant’s task, and the stimulus. The conditions refer to the right participant, e.g. in the action-conflict condition, there is an action conflict for the right participant. The parts of each task rule that are activated by the stimulus either because of the relevance to the rule’s precondition or an overlap with the features of its action are presented in bold face.

than the one due to action conflicts. If action and task conflict occur simultaneously, the increase in reaction time is more than the sum of the reaction time increase when action and task conflict occur in isolation. In summary, participants create shared task representations that affect task performance beyond what happens in the individual Simon task. Table 7.2 lists the different experimental conditions and the conflicts they evoke from the perspective of the participant sitting on the right-hand side. Note that the no-conflict and action-conflict conditions mirror the same conditions in the individual Simon task. The difference is that in the social Simon task responsibilities are shared by two participants.

The analysis of Sebanz et al. (2005) suggests that in the action-conflict and the task-conflict condition a representation of the incorrect action is activated, which causes an action selection problem. Electrophysiological studies provide evidence that in the task-conflict condition of the joint task the coactor’s action is explicitly suppressed (Sebanz et al., 2006b; Tsai et al., 2006). This indicates that the Simon effect is due to a conflict arising after actions have been activated, i.e. an action selection problem.

Hommel et al. (2009) find a Simon effect in the social Simon task only if the partner is perceived to be a likable person. No Simon effect is observed when the partner behaves in an intimidating manner. This indicates that the social relationship between participants affects whether shared task representations are constructed. There is also evidence that mood plays a factor in whether shared task representations are created (Kuhbandner et al., 2010). These findings suggest that lower-level mechanisms responsible for the Simon effect are modulated by higher-level ones.
7.3 Requirements

In this section, we extract the requirements for a minimal architecture of joint action from the description of the social Simon task in the previous section and the different perspectives on coordination in Section 7.1. We label these requirements so that we can refer back to them in the next section, in which we describe the architecture.

We start with the requirements relating predominantly to emergent coordination processes, which are based on direct perception-action links:

REQ1 The architecture needs to represent an actor’s perception and action mechanisms, which are required by any actor that is to engage with its environment in a meaningful way.

REQ2 Perceptions are to be encoded by the activation of distal, possibly subsymbolic features of varying level of abstraction (Prinz, 1997).

REQ3 Action effects are also encoded by features (Prinz, 1997).

REQ4 Perceptions and action effects are encoded by the same set of features. Such shared representations for perception and action are postulated by common coding theory (Prinz, 1997) and claimed to be responsible for the Simon effect in the action-conflict condition (Hommel et al., 2001). In order to allow for the encoding both of the actor’s own actions and the ones of another actor as postulated by common coding theory, at least some of these features need to provide for actor-independent coding.

REQ5 Actions are planned in terms of their effects as assumed by common coding theory (Prinz, 1997). Therefore, a mechanism is required to translate action effects into appropriate motor commands. This mechanism is called an inverse model (see Kawato, 1999, for an early review of internal models of motor control).

REQ6 To enable action simulation and thereby the prediction of action outcomes and another actor’s intentions (Fogassi et al., 2005; Sebanz and Knoblich, 2009; Wolpert et al., 2003), a mechanism complementary to the inverse model is required that translates motor commands into action effects (forward model, Kawato, 1999). As we will discuss later, forward models are assumed to be crucial to any sort of motor control.

REQ7 Higher-level mechanisms need to be available to suppress the execution of an action that is not due to planning but due to action simulation or another involuntary activation of its features (Sebanz et al., 2006a). A requirement for this is also that self and other
can be distinguished. Self-other distinction is the responsibility of a higher level mechanism (Böckler et al., 2010).

The next set of requirements predominantly relates to planned coordination and practical reasoning processes:

**REQ8** An actor employs a mechanism of practical reasoning based on goals, intentions, and beliefs in order to perform higher-level reasoning and planning postulated by the philosophical view on (joint) action (Bratman, 1987, 1992).

**REQ9** An actor needs to be able to represent the goal of the joint action, its own task, and possibly the tasks of coactors (Vesper et al., 2010).

**REQ10** The task representation needs to be embedded appropriately in the actor’s intentions and its practical reasoning so that it can be accounted for in the actor’s action planning. Part of this is the ability to attribute mental attitudes to coactors (Flavell, 2004) and to distinguish its own intentions from those of coactors and relate its own and the coactors’ intentions to the overall joint action (Knoblich and Sebanz, 2008). This represents the theory theory perspective on the theory of mind. In the social Simon task, the participant clearly attributes some form of intentions to its partner.

**REQ11** The representation of the joint action (in terms of intentions, etc.) needs to impose appropriate attributes of joint action onto the actors, as postulated by Bratman (1992).

**REQ12** Whether a coactor’s task is represented needs to be modulated by information such as the current mood of the actor or its representation of the social relationship with the coactor. This follows from the results of Hommel et al. (2009) and Kuhbandner et al. (2010). Obviously, this depends on **REQ8** as it is not easy to see how these kind of effects could be represented without higher-level mental representations of the social relationship with coactors, for example.

**REQ13** The actor needs to be able to infer the intentions of others (Knoblich and Sebanz, 2008). Intention inference can rely on action simulation by observing a coactor’s behavior or it can rely on practical reasoning based on the mental attitudes attributed to the coactor. The former instance is compatible with the simulation theory of mind reading and the latter with theory theory.

**REQ14** The actor’s practical reasoning mechanisms need to be able to employ forward and inverse models as described above for the planning, prediction, and monitoring of its own and others’ actions. Using forward and inverse models to account for a coactor’s mental states is in line with the simulation theory of mind reading.
REQ15 REQ14 implies that the representation of motor commands and action effects by features needs to be compatible with the representation of actions and goals used by practical reasoning mechanisms.

REQ16 The actor’s practical reasoning mechanisms need to be able to consider the actor’s own and the coactors’ mental attitudes for prediction of future states and the monitoring of the actor’s own and the coactors’ actions. Prediction and monitoring have been identified by Vesper et al. (2010) as central processes in joint action.

In effect, the social Simon task reveals an interaction between practical reasoning and mechanisms of emergent and planned coordination. In the social Simon task, practical reasoning and joint intentions entail planned coordination in the form of shared task representations that account for the responsibilities of a partner. Shared task representations, on the other hand, allow action conflicts to arise because perceptions and corepresented actions share the same representations, which is a mechanism of emergent coordination. Task conflicts arise because all planned actions share the same representational domain, which requires resolution through the suppression of actions that are not to be executed. Later in this chapter we will demonstrate that an architecture fulfilling the requirements above is able to reproduce the human performance in the social Simon task. In fact, some of the features listed above are not fully required for a representation of the social Simon task, but they are essential mechanisms of joint action: (1) In the social Simon task, intention inference is likely to rely only on higher-level reasoning mechanisms because the social Simon task does not really endorse or require inferring the intentions of a coactor from observing their actions (REQ13); (2) it is also not clear which role prediction and monitoring processes play in the social Simon task (REQ14, REQ16).

### 7.4 Architecture

We describe an architecture of joint action that fulfills the requirements identified in the previous section. Thereby, we contribute to bridging the gap between mechanisms of emergent and planned coordination on the one hand and practical reasoning and joint intentions on the other. From that perspective, our work expands on the one of Knoblich and Sebanz (2008) and Vesper et al. (2010). Yet we intend to be more specific about the interplay between the different mechanisms and forms of coordination. This entails a departure from the holistic perspective of those authors and a zoom in onto the representations and the processes that glue them together. We also put a larger emphasis on the role of practical reasoning and joint intentions in joint action.
In line with the topic of this thesis, we adopt a computational perspective and provide a computational-conceptual model of joint action. On the one hand, our architecture provides a framework for computational models of joint action that consider the interplay between mechanisms of emergent and planned coordination as well as practical reasoning and joint intentions. On the other hand, we contribute to the theoretical study of joint action by providing an account in terms of computational primitives, which map naturally onto mental representations and processes. In that, we contribute to extending the discussion about what a joint action is to how mechanisms at all levels of coordination interact to affect individual behavior. As we will argue later in this chapter, these mechanisms play a crucial role in the transmission of culture, which is the topic of this thesis.

We start this section with an overview of the architecture and then proceed by describing its mechanisms, i.e. representations, processes, and their interactions. Whenever possible, we justify our choices with pointers to the research discussed in Section 7.1 and refer back to the requirements presented in the previous section. In the next section, we demonstrate how the actual human performance in the social Simon task can be reproduced by the mechanisms of this architecture.

### 7.4.1 Overview

We distinguish between the inter-personal and intra-personal side of joint action. At the inter-personal side, the representation of a joint action is based on joint intentions between the coactors. This reflects the philosophical view on joint action, which emphasizes joint planning and the commitment of coactors to the joint action. The inter-personal dynamics of a joint action play out on two different planes: The one at which the coactors’ actions contribute to the completion of the joint action, i.e. the fulfilling of the joint intentions, and the one at which the coactors coordinate their contributions to the joint action. This coordination manifests itself, for example, in communication or the synchronization of movements.

The inter-personal dynamics are generated from the interaction of mechanisms of emergent and planned coordination and intentional planning at the intra-personal side. We note that the objective, inter-personal state of the joint action is generally unavailable to an actor. Therefore, intentional planning operates on a subjective view of the joint action. At the intra-personal side of joint action we distinguish between the following functional levels:

**Perception-action level**  Sensorial input is received and mental action representations are translated into muscular movements (REQ1).

**Non-intentional level**  Shared representations for perception and action support emergent coordination and the engagement in the joint action, partly by facilitating synchronization and partly by providing supplementary information to the intentional level.
7.4. Architecture

Intentional level  Intentional planning is performed based on subjective representations of the joint action. Monitoring and prediction processes, which partly rely on the non-intentional level, are initialized in service of planning. This is the level at which planned coordination, practical reasoning, and joint intentions are given meaning.

As we shall see, the distinction between these levels is rather blurred and we adopt it mainly to guide our discussion and not to make strict ontological distinctions.

7.4.2 Memory

We assume that there are two different memory systems with their own modes of processing. Dual-process and dual-system models of memory and cognition have been studied extensively across various disciplines within the last 30 years or so (for recent reviews see Frankish, 2010; Gawronski and Creighton, 2012). There is ample empirical evidence that humans employ two different types of cognitive processing in various tasks (see citations in Darlow and Sloman, 2010; Evans, 2003). The first mode of processing is commonly accepted as fast, automatic, and subconscious. The second mode of processing is commonly considered to be slow, controlled, and conscious (Frankish, 2010). Following Smith and DeCoster (2000), we assume that the first mode of processing is an associative one that operates on a slow-learning associative memory. The second mode of processing is a logic-based one that operates on a fast-learning symbolic memory. The important bit here is that the first memory and its processing mode enable the mechanisms of the non-intentional level, and the second memory and its processing mode enable the mechanisms of the intentional level. We will later discuss this point in detail. Via interactions between these two memories, intentional and non-intentional level and thereby mechanisms of emergent and planned coordination as well as intentional planning interact. Therefore, those memory systems provide the backbone of the architecture.

Associative Memory

The first memory system is an associative memory that underlies the non-intentional level. The basic elements of this memory are features in the sense introduced above. Each feature is represented by a processing unit. At any point in time, a feature has a certain activation level. Via inhibitory and excitatory connections between features, activation spreads from feature to feature. If there is a connection between two features, the activation of one feature affects the activation of the other. Hence, this memory encodes associations between features and processing is concerned with exercising these associations by propagating feature activations. This processing mode is assumed to be effortless and automatic, not requiring any cognitive resources. As an example consider a feature “house”. This feature is
likely linked to other features such as “door”, “window”, “brick”, etc. Hence activating the features “door”, “window”, and “brick”, the “house”-feature is going to be activated, and vice versa. However, the features “windows” and “doors” are likely to have a strong connection with the feature “car” as well. Hence, there is an overlap between the features “house” and “car”. Activating the “house”-feature affects the activation level of the “car”-feature. We shall assume throughout this chapter that features represent information at the level of abstraction exemplified here and that perceptual input is resolved to this level as well, by some process operating outside of this architecture.

The features currently activated constitute a kind of working memory. Learning adjusts the connections between features based on how often they are activated at the same time. When two features are often activated at the same time, their connection strength is increased and the activation of one of them becomes sufficient to activate the other. However, this learning process is assumed to proceed slowly. Connectionist models offer a framework for the computational implementation of the associative memory system (Rumelhart et al., 1986). We have discussed connectionist models with more detail in Chapter 5.

**Symbolic Memory**

The second memory, which we call *symbolic memory*, consists of representations in a language that allows for symbolic reasoning, e.g. propositional or first-order logic. This is the memory system employed by the intentional level. Processing on this memory amounts to logical inference based on the content of this memory. This type of processing is a slow and cognitively expensive process, which is applied only within limits of cognitive capability and motivation. However, new content for this memory can be learnt quickly, which enables the rapid acquisition of episodic knowledge, whose truth value is determined by a single experience. Episodic knowledge includes knowledge extracted from communication. Symbols exchanged during communication can be interpreted by the symbolic information in this memory and stored in the same format.

We assume that sentences in symbolic memory are represented by sets of features in associative memory. Activating the set of features representing a sentence causes that sentence to be made true in symbolic memory (if it is not yet true). However, only those sentences whose features are sufficiently activated at the moment are available for reasoning. This constitutes a second kind of working memory and allows the current context, which is encoded by activated features, to influence which information in memory is accessible at the moment. Inferring a sentence by symbolic reasoning, on the other hand, causes the features representing this sentence in associative memory to be activated, which leads to the activation of further features. Assume, for example, that symbolic memory is based on a propositional logic. A propositional sentence could be *house-with-window*. This sentence
Figure 7.2: An example of the interaction between symbolic and associative memory. Rectangles denote sentences in symbolic memory and circles denote features in associative memory. Blue shapes denote activated entities, while red ones denote inactive entities. Solid black lines without arrows show which features are associated with which sentences and gray lines with arrows show the strength of associations between different features. In (a) only the proposition \textit{house-with-window} and its features are active. Because the “window” and “house” features are associated with the “door” feature, the “door” feature is activated in (b). Because the “house” and the “door” feature are activated, also the proposition \textit{house-with-door} becomes activated in (c).

could be represented in associative memory by the features “house” and “window”. Via the feature “house”, further features such as “door” can be activated, which might lead to further activations of sentences in symbolic memory, e.g. \textit{house-with-door}. Figure 7.2 depicts this process. Both kinds of processing hence are assumed to operate in parallel and mutually affect each other. Obviously, the expressiveness of associative memory limits the complexity of sentences that can be represented by configurations of features. An important remark is that despite the interaction between both memories, the content of symbolic memory is typically more permanent and can outlive the transient feature activations of associative memory.

The extraction of sentences from features is a learning process that can be based, for example, on the frequency with which a number of features is activated simultaneously. The benefit of this extraction is that an extracted sentence can be used in other contexts as well, independent of the current configuration of associative working memory. In the other direction, if a combination of sentences is used repeatedly at the same time in symbolic working memory, their features in associative memory are repeatedly activated simultaneously and
become associated. Learning is outside the scope of this architecture, albeit an important component to consider in some instantiations of this architecture.

7.4.3 Actions

At the core of any architecture of joint action have to be actions, of course. Here, we understand as an action any muscular movement by an actor that causes a change in the environment of this actor. An action can be grasping an object but it can also be the production of sound for communication. This amounts to the external side of action. Internally, any muscular movement is matched by a representation of the motor command that produces this movement. We assume that part of the representation of the motor command are the preconditions under which this action is executable. As indicated earlier, actions are likely planned in terms of their effects. Hence, action effects are the second representation associated with an action. Essentially, a motor command is the means to the end represented by particular effects. As we shall see shortly, an actor is effortlessly able to retrieve the motor command associated with particular effects and vice versa. In the following, when we talk about an action representation, we refer to the effects and the motor command of that action.

We assume that both motor commands and effects are represented by sets of features in associative memory (REQ3). Hence, the representations of different actions can overlap via their features. Likewise, the representations of actions and perceptual input overlap because perception and action share features as a common representational system (REQ4). Considering that the features of the motor command and the effects of an action are activated at the same time frequently, there is a strong association between motor command and action effects in associative memory. Consequently, an activation of features associated with the effects of an action also activates the associated motor command and vice versa. This implements the above mentioned effortless translation between effects and motor commands. We suggest that this mechanism is also related to affordances. An action is activated by the mere observation of an object whose features overlap with the features of the action’s effects, e.g. because the effect of this action is some manipulation of that object. We assume that features representing motor commands or action effects are represented by symbols in symbolic memory.

Another consequence of encoding action representations in terms of features is that the same action representation can be activated multiple times, resulting in an increased activation level, and multiple action representations can be activated at the same time. The interaction between symbolic and associative memory enables action representations to be shared between the intentional and non-intentional levels (REQ15).
7.4.4 The Non-intentional Level

Following recent models of motor control (e.g. Grush, 2004; Hurley, 2008; Kawato and Wolpert, 1998), we adopt a control system perspective to perception and action for the non-intentional level. The commonality of these models is the assumption of internal models of motor control, which have been hypothesized to be used by the central nervous system (e.g. Wolpert et al., 1998). As indicated before, two types of models are distinguished: Inverse models determine the motor command required to lead to particular effects. Forward models predict the effects of motor commands. This allows an actor to translate between both components of actions as indicated above. Obviously, in our architecture inverse and forward models are implemented by associative memory, resolving associations between the motor commands and effects of actions in both directions (REQ5, REQ6). In that, associative memory underlies the non-intentional level and inverse and forward models share the same representational system. As we will discuss in Section 8.3, others before us have considered inverse and forward models to be implemented by an associative memory (e.g. Wolpert and Kawato, 1998).

An appropriate composition of inverse and forward models enables the actor to deal with basic motor control without the involvement of any higher-level cognition, which is in line with the embodied cognition perspective to joint action. We consider the setup depicted in Figure 7.3. The figure shows processes and representations at the non-intentional and perception-action levels. Also the control and information flow between these processes and representations as well as the information exchange with the intentional level at the top is shown. An important point to make upfront is that the perceptual input received from the environment is assumed to be encoded at the abstract level exemplified in the previous section. In fact, input basically amounts to an activation of features (REQ2).

Based on Figure 7.3, we discuss now the different functions implemented by the non-intentional level. The inverse model translates effects into a motor command given the current input, which specifies the preconditions that the motor command has to satisfy. In basic execution mode (Figure 7.4a), the effects correspond to a goal state (point 2 in Figure 7.3) that the control system is to achieve given the current input. This goal state is represented by an activation of corresponding action effect features. The execution of the subsequently activated motor command acts on the environment in terms of muscular movement, which thereby affects the input that the inverse model is receiving. The system also relies on an internal forward model that produces a rapid estimation of the effects of the current motor command (Figure 7.4b). These estimations are more rapidly available than the feedback from the environment and therefore can support the inverse model in its control task. When the inverse model fails to control for the error between the input and goal state, an error
Figure 7.3: Control and information flow at the non-intentional and perception-action levels. Rectangles depict representations, rectangles with rounded corners depict processes. Solid lines represent control flow and dashed lines show the points at which information exchange with the intentional level happens. Dotted lines mark the separation between the three levels.

A signal (point 3) is produced that indicates that higher-level control is required to correct for the error. Hence the inverse model implements some form of action monitoring that is available to the intentional level (REQ14).

As indicated in the figure, this system is used beyond basic control. When the system is taken offline by inhibiting its output (point 6), the inverse model can be used to generate motor commands (point 4) for different goal states (point 2) (Figure 7.4c). The forward model can then be used to make predictions for the effects (point 1) of these motor commands (point 5), a process we shall call action prediction (Figure 7.4d, REQ14). In fact, this kind of prediction can also be applied to another actor’s actions. Considering that the intentional level is not necessarily ensuring output inhibition, there might be a tendency to execute an activated motor command even if it was not voluntarily planned for.

We assume that observing the effects of others’ movements activates corresponding action effect features, thereby causing the respective motor command to be activated via the inverse model (Figure 7.4e). This enables basic action understanding postulated by common coding theory, which supports imitation. When the extracted motor command (point 4) is used as an input to the forward model (point 5), action simulation is obtained, which allows...
individuals to predict another actor’s movement (point 1) and infer their intentions at a basic level (Figure 7.4f, REQ13, REQ14). If output is not inhibited, copying of the other actor’s movements results, which facilitates mimicry and synchronization. These mechanisms require that the system can distinguish between possible and actual actions, i.e. those that should be executed and those that should be inhibited, and between one’s own and others’ actions. We assume that the intentional level is in control of these aspects (REQ7).

Essentially, the same system is used for perception and action. In particular, the inverse model is active during action observation and action planning. Also, inverse and forward model share the same representational system as we discussed earlier. This marks the connection with neuroscientific results about the mirror neuron system. However, because of the complexity of this topic (Oztop et al., 2006), we refrain from making any further claims regarding the biological plausibility of the implementation of motor control models in this architecture.

 Strictly, an additional component is necessary to translate between the subjective perception of another actor’s movements and the effects projected onto the observer’s own body. Considering the abstract representation of input described earlier, the low-level distinction
between the actor’s own and other’s actions is assumed to be resolved outside of this architecture. Observing another actor’s movements leads here directly to the activation of features that represent the effects of this action from an actor-independent perspective.

The non-intentional level implements the emergent coordination phenomena described in Section 7.1.3, thereby enabling basic forms of coordination in joint action. For example, the actor and its environment are conceptualized as a dynamical system, which is a central component of the embodied cognition perspective. Shared representations for action and perception as postulated by common coding theory enable action simulation, which allows individuals to infer the intentions of others and also facilitates imitation and synchronization. Intentional-level coordination processes, which we discuss in the next section, build partly on what we have described here.

### 7.4.5 The Intentional Level

In this subsection, we describe the processes that enable intentional planning and coordination. We assume that these processes operate on symbolic representations of beliefs, goals, and intentions as in the agent model presented in the previous chapter (REQ8), which stands in the tradition of formal models of intelligent agent reasoning (Wooldridge, 2009). Because these entities are symbolic constructs, symbolic memory is naturally the system for their storage and manipulation. Figure 7.5 provides an overview of the interrelations between the different processes and representations at the intentional level and those at the other two levels. We begin by outlining the content of symbolic memory. Then we describe goals and intentions that represent the actor’s current state of planning towards their individual goals and towards the joint action. Finally, we discuss the planning processes operating on these representations.

#### Symbolic Memory

Symbolic memory represents the mental attitudes of the agent, including a mental representation of the state of the world in general and of the joint action in particular. First and foremost we assume that goals and intentions are part of symbolic memory. Thereby, these entities take part in the interaction between symbolic and associative memory. For example, an action referenced by an intention in symbolic memory activates corresponding features in associative memory.

The beliefs of the agent contain further information about the joint action, e.g. the state of relevant objects or coactors, as suggested by Vesper et al. (2010). Also the social relationship with coactors is likely to be encoded in some way, including their perceived common ground—the information they assume to share—and their trust relationship.
Figure 7.5: Information flow and dependencies between the intentional level and the two other levels. Rectangles denote representations, rectangles with rounded corners denote processes. The solid lines at the bottom depict the feedback loop at the non-intentional and the perception-action level. Dashed lines represent some kind of information flow (apart from the one between planning and the intention structure).

If an actor trusts in the intentions and competence of a coactor, there might be no reason to monitor that coactor’s performance and hence no reason to represent the part of the task of that coactor. We hypothesize that this is related to the findings of Hommel et al. (2009) that a shared task representation is not created when the coactor is perceived as unfriendly. In that case, the participant might doubt that the coactor has the appropriate intentions to perform their part of the task. Because then there is no way the actor can make up for the coactor’s “misbehavior”, a shared task representation is not constructed. Hence, higher-level information can modulate the construction and employment of a shared task representation (REQ12).

Recall that feature activations in associative memory partly determine the content of symbolic memory and also which part of memory is accessible at any point in time. Hence, symbolic memory retrieves some of its information from perceptual input indirectly via associative memory.

**Goals and Intentions**

Goals and intentions represent the current state of planning towards the achievement of individual goals and the joint action from the perspective of the actor. A goal is a state of
the world that the actor desires to hold. We distinguish between goals that the agent has not committed to pursue actively (i.e. desires) and goals that the agent has selected for pursuit. An intention is a commitment to execute a plan or an action in order to achieve one or more goals. An intention is the adopted means to the end represented by a goal, and an intention can include further subgoals if it refers to a plan and not to a primitive action. Note the relation to the action representations consisting of effects (end) and motor commands (means) described previously. Goals and intentions are arranged in a hierarchy of alternating levels of goals and intentions, which we call intention structure in the following. We assume that a goal referring to action effects and an associated intention referring to a motor program can form the lowest level of this intention structure. This enables the non-intentional level to be integrated with the intentional level. Note that the intention structure is not necessarily a tree because an intention at one level might be responsible for achieving two or more different goals at the next higher level.

Goals and intentions are annotated with the actors that are believed to hold the respective attitude, which contrasts with action representations at the non-intentional level that are based on actor-independent features (REQ7). Actor annotations mark the responsibility of coactors for particular goals and intentions. If a goal has more than one actor, this goal as well as the part of the hierarchy below it form a joint intention. However, parts of this joint intention can be associated with only a subset of the actors. In that case, this part of the task is those actors’ responsibility. Hence this structure is a representation of the goal of the joint action and the individual tasks of the coactors (REQ9). The intention structure distinguishes intentions by actor and relates the intentions of different actors to the overall joint intention (REQ10).

The kind of intention structure described here has been hypothesized, for example, by Bratman (1987) and Tomasello et al. (2005). Our intention structure represents both the agent’s own intentions as well as its beliefs about its coactors’ intentions as top-level elements. This contrasts with other formalisms of intentions and joint intentions (e.g. Cohen and Levesque, 1991; Grosz and Kraus, 1996), which typically separate the actor’s own intentions from its beliefs about coactors’ intentions. Of course, an actor in our model does not act on the intentions of another participant. However, we will see later how those parts of the intention structure relating to coactors are expanded during planning. These other-related intentions are crucial in enabling prediction and monitoring of coactors’ actions. In contrast to existing formalisms, our conception of intentions and joint intentions as outlined here is minimalistic. We leave open the possibility for a more sophisticated formal elaboration on the structures of intentions within our framework. We do assume, though, that joint intentions impose requirements of joint action on the coactors as identified by Bratman (1992), namely mutual responsiveness, commitment to the joint action, and commitment
7.4. Architecture

to mutual support (REQ11). We discuss how Bratman’s features of joint action affect the evolution of the intention structure when we go into the detail of planning.

We emphasize again that the intention structure is only the participative view of the joint action by an actor and different coactors are likely to perceive the state of their joint action differently. In particular, an actor needs to be able to infer the intentions of a coactor or plan for the coactor’s actions to fill in the blanks of the intention structure. Likewise, there needs to be some sort of error handling that resolves the discovery of mismatches in the representation of the joint action. This would likely involve coordination smoothers in the sense of Vesper et al. (2010), which we assume to operate on a second plane on the inter-personal level of the joint action. For example, coactors could engage in communication to resolve a misunderstanding or they could slow down or exaggerate their movements to emphasize their intentions.

We have not talked about the resolution of intentions or actions yet. It is well known that many actions do not require an intention because they have been internalized and can be executed subconsciously. On the other hand, often actions are executed consciously. We are going to leave it open to the planning process at which level of intentionality particular intentions are executed. Therefore, we discuss later how exactly intentions can be expanded to and be executed at different levels of granularity. However, we assume here that consciousness and intentionality are equal in that intentional action is conscious action and non-intentional or automatic action is subconscious action. This distinction is well maintained among some psychologists (Knoblich and Sebanz, 2008) even though there is no universal agreement (Dijksterhuis and Aarts, 2010).

Planning

Planning includes a host of different processes. Its main responsibility though is the incremental expansion of the intention structure. Planning relies on two strategies to fill in the gaps of the intention structure: means-end reasoning and intention inference.

Given a particular goal, means-end reasoning determines an intention that can achieve it. In case the intention requires joint action, communication might be necessary to establish a proper joint intention. Means-end reasoning can rely on symbolic planning as well as processes at the non-intentional level. Symbolic planning yields intentions or action representations to achieve goals. Inverse models can be used to determine a motor command from a particular goal or in fact expected action effects (REQ14). The choice between symbolic planning and the use of inverse models determines at which level of intentionality an intention is encoded. Symbolic planning leads to a representation in terms of further intentions while the use of inverse models yields a low-level action representation.
Chapter 7. A Minimal Computational Architecture of Joint Action

Based on observations of a coactor’s actions, intention inference produces those intentions of the coactor that are deemed to have led to these actions, which can then be used to augment the intention structure. Intention inference has been suggested, for example, by Sebanz and Knoblich (2009). A coactor’s intentions can be inferred both by employing action simulation as described previously and by symbolic reasoning (REQ13).

Every action representation added to the intention structure is shared with the non-intentional level via the association between symbolic and associative memory. Recall that action representations at the non-intentional level are not annotated with their actors and that their activation entails an indiscriminate tendency to execute the respective actions. It is intentional planning that retains the control over which actions are in the end executed by the actor and which actions are suppressed.

As described before, parts of the intention structure are annotated with the responsibility of other parties. Whether planning actually accounts for these parts of the intention structure depends on a number of factors, e.g. cognitive limitations, the importance of avoiding mistakes in this joint action, or the social relationship between the actor and the coactor(s) responsible for those intentions (REQ12). Planning in general relies on the content of symbolic memory, of which the social relationship between the coactors is only one component.

Planning also relies on processes that monitor the progress towards adopted intentions. In contrast to the previously discussed action monitoring, we call this higher-level process intention monitoring (REQ16). This process is closely tied to planning and will cause replanning when conflicts between intentions arise or when an intention becomes unachievable. Obviously, this type of monitoring is only possible because of the higher-level representation of the joint action. Complementary to action prediction, intention prediction uses symbolic memory in general and the intention structure in particular to predict future states and actions (REQ16). This is a process of reasoning.

Planning is central to the achievement of Bratman’s (1992) criteria for joint action. We assume that coactors are committed to the success of the joint action because of their joint intentions. Due to this commitment, actors have a vested interest in the success of the joint action. To ensure success, they take their coactors into account during planning and deploy monitoring to be aware of any arising problems. Hence, coactors are mutually responsive to each other’s intentions and actions, which enables them to engage in supportive action if required. Supportive action includes taking over responsibilities from coactors in need or any form of coordination smoothing. However, we saw earlier that actors do not always take their coactors into account. Depending on the situation, a coactor’s part of the joint action might not be represented at all and therefore cannot be monitored in detail either.

The intentional level implements shared task representations postulated by planned coordination as well as practical reasoning and joint intentions.
7.5 Representing the Individual and Social Simon Task

We now describe how to instantiate the architecture in order to represent the individual and social Simon task. We start this section with a description of the assumptions we make and the basic agent setup. In the remaining subsections, we describe how each of the experimental conditions of the individual and social Simon task would play out in this architecture given the assumptions. We close this section with a brief summary.

7.5.1 Assumptions

We label the following assumptions to be able to refer back to them later.

Assume that the knowledge representation in symbolic memory is based on a formal logic with constant symbols, predicate symbols, function symbols, action symbols, and actions (ASS1). We require as part of this logic the constant symbols $Me, You, Color, Location, Right, Left, Red, Green$. In the following, let $x$ and $y$ stand for arbitrary constant symbols. We require the predicate symbols $nice(x)$, $pushed-right$, $pushed-left$ and the function symbol $attr(x)$. The symbols $push-right$, $push-left$, $Simon-task(u,u',w,w')$ denote actions. The $Simon-task$ action has parameters $u$ (the left-hand side participant), $u'$ (the right-hand side participant), $w$ (the characteristic of the stimulus the left participant responds to), $w'$ (the characteristic of the stimulus the right participant responds to). We denote by $Goal_{\{x,y\}}(\varphi)$ the goal by agents $x$ and $y$ that proposition $\varphi$ holds and by $Int_{\{x,y\}}(\alpha)$ their (joint) intention to execute action $\alpha$. We denote by $Do_{\{x,y\}}(\alpha)$ that agents $x$ and $y$ do action $\alpha$. We do not define the semantics of this logic but rather rely on the intuitive meaning of these symbols to make our point.

We assume here that only a limited set of sentences is mapped onto feature sets in associative memory (ASS2). The reason for this simplification is that mapping complex sentences onto features would require a more expressive representation in associative memory and quickly become intractable. Because of this simplification, we also neglect the mechanism of feature activations affecting the activation level of symbols (ASS3). This part of the architecture is not required in this example. We only assume that feature activations in associative memory due to perceptions of stimulus attributes are translated into corresponding sentences in symbolic memory (ASS4).

Table 7.3 lists those mappings between symbolic and associative memory that are relevant in the following. We assume that these mappings have been established a priori (ASS5). In addition, there is a prior bidirectional association between the features representing the effects $pushed-left$ ($pushed-right$) and the features representing the associated motor command $push-left$ ($push-right$) (ASS6).
Table 7.3: Mapping between representations in symbolic and associative memory.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>pushed-left</td>
<td>left, index-finger-left, tactile-left</td>
</tr>
<tr>
<td>pushed-right</td>
<td>right, index-finger-right, tactile-right</td>
</tr>
<tr>
<td>attr(Color) = Red</td>
<td>red</td>
</tr>
<tr>
<td>attr(Color) = Green</td>
<td>green</td>
</tr>
<tr>
<td>attr(Location) = Left</td>
<td>left</td>
</tr>
<tr>
<td>attr(Location) = Right</td>
<td>right</td>
</tr>
<tr>
<td>push-left</td>
<td>$x_1, x_2, x_3, \ldots$</td>
</tr>
<tr>
<td>push-right</td>
<td>$y_1, y_2, y_3, \ldots$</td>
</tr>
</tbody>
</table>

Figure 7.6 visualizes the mappings between the elements of symbolic and associative memory, and the associations between different groups of features. There are four different ways in which features in associative memory can receive activation:

**Attitude Representation** When a belief, goal, or intention is created, associated features are activated (ASS7). For example, when a goal makes reference to pushed-left, the following features receive activation: left, index-finger-left, and tactile-left.

**Stimulus Perception** When a stimulus is perceived, features associated with that stimulus are activated (ASS8). For example, if the perceived stimulus is green and appears on the left side, the features green and left receive activation.

**Action Planning** When the intentional level employs the inverse model of the non-intentional level, it effectively activates features associated with the state to be achieved in order to retrieve an appropriate motor command (ASS9). For example, if the state to be achieved is pushed-left, the features left, index-finger-left, and tactile-left receive activation, which leads to the features $x_i$ to be activated.

**Action Execution** When the intentional level invokes the execution of an intention that refers to a primitive action, the features associated with that action receive activation (ASS10). For example, if the action push-right is to be executed, the features $y_1, y_2$, and $y_3$ are activated.

For activation propagation from the features representing pushed-left (pushed-right) to the features representing push-left (push-right) we make the following assumption (ASS11): Only together, attitude representation and stimulus perception generate enough activation on the features representing pushed-left (pushed-right) so that this activation is propagated to the features representing push-left (push-right). In Figure 7.6, this is
7.5. Representing the Individual and Social Simon Task

pushed-left    pushed-right

left

index-finger-left

tactile-left

push-left     push-right

right

index-finger-right

tactile-right

attr(Location) = Left
attr(Location) = Right
attr(Color) = Red
attr(Color) = Green

x_1     x_2     x_3

\[ \cdots \]

y_1     y_2     y_3

\[ \cdots \]

Figure 7.6: Elements of symbolic and associative memory and their relationships. Rectangles denote sentences in symbolic memory and circles denote features in associative memory. Solid black lines without arrows show which features are associated with which sentences and gray lines with arrows show the strength of associations between different features. The threshold symbol on the two gray arrows indicates that activation along this link is only propagated if it exceeds a certain threshold.

represented by the threshold symbol on the links between the different sets of features. For example, representing a goal which refers to pushed-left is not sufficient to have activation be propagated to the features of push-left. Likewise, activating the left feature, for example, is not sufficient to have any activation be propagated to the features of push-left. We do assume that activation due to action planning and execution is sufficient to overcome this threshold (ASS12). We also assume that only activation for the execution of a motor command is sufficient to actually lead to a motor command being executed after activation has settled to a stable pattern (ASS13). In that sense, the intentional level is to some extent in control of which actions are actually executed, and in this instance, unintended output is always inhibited.

7.5.2 No-Conflict Condition in the Social Simon Task

We describe how a no-conflict condition trial in the social Simon task could unfold given the representation described above, from task preparation to the perception of the stimulus to
action execution. Naturally, two agents would need to be instantiated but apart from different task instructions, their setup would be the same. Therefore, we focus on one agent here, our participant sitting on the right-hand side.

1. **Task Preparation I** In this condition, the agent is instructed that the one on the left (You) reacts to the green stimulus and the one on the right (Me) to the red stimulus. This is represented by the following goal at the top of the intention structure of the right-hand side agent:

   \[ \text{Goal}_{\{\text{you,me}\}}(\text{Do}_{\{\text{you,me}\}}(\text{Simon-task}(\text{You,Me,Green,Red}))) \]  

   (7.1)

   Obviously, establishing this goal as a joint intention would require communication with the other participant or be based on the implicit assumption that the other participant would cooperate. The agent is going to adopt this goal for pursuit.

2. **Task Preparation II** This goal is reasoned to be achievable by the following intention:

   \[ \text{Int}_{\{\text{you,me}\}}(\text{Simon-task}(\text{You,Me,Green,Red})) \]  

   (7.2)

3. **Task Expansion** Let the Simon-task action specify subgoals, which lead to the adoption of the following two goals, depending on the provided parameters:

   \[ \text{Goal}_{\text{you}}(\text{attr(Color)} = \text{Green} \leftrightarrow \text{pushed-left}) \]  

   (7.3)

   \[ \text{Goal}_{\text{me}}(\text{attr(Color)} = \text{Red} \leftrightarrow \text{pushed-right}) \]  

   (7.4)

   This represents action and task corepresentation in the sense of Sebanz et al. (2005): The other participant’s task rule as well as its action (via the action’s effects) are corepresented. Corepresentation depends on nice(You) because only when this proposition holds, the agent might represent the part of the task belonging to its partner. If nice(You) is false, Equation 7.3 would not be represented. Because of the two goals in this example, the features representing pushed-left and pushed-right are activated in associative memory (ASS7). This activation is below threshold level and does not lead to any activation flowing to the features representing the motor commands push-left and push-right that achieve these two goals (ASS11). Also the features representing attr(Color) = Green and attr(Color) = Red are activated.

4. **Stimulus Presentation** A red stimulus is presented on the right side.

5. **Feature Activation** This stimulus is perceived and translated (by a process outside of this architecture) into an activation of the features representing red, and right in associative
memory (ASS8). The feature right is part of the feature set of pushed-right, which thereby receives further activation. Now the activation of these features is sufficient to overcome the threshold and propagate activation over to the features representing the push-right motor command (ASS11). Still, this activation is below the threshold for the motor command to be executed immediately (ASS13).

6. Proposition Activation Feature activations cause the sentences attr(Color) = Red and attr(Location) = Right to be made true in symbolic memory (ASS4).

7. Action Planning I Now the goal Goal(me)\{attr(Color) = Red \leftrightarrow pushed-right\} becomes relevant because pushed-right needs to be made true to achieve this intention. Let us assume that symbolic memory is unable to provide for the action to achieve this goal. Therefore, planning falls back to the inverse model at the non-intentional level to resolve an appropriate motor command. The goal pushed-right is translated into an activation of its features in associative memory above threshold level, which causes an increase in the activation of the features representing the motor command push-right (ASS9, ASS12). However, the activation of the features constituting this motor command is still not sufficient to evoke its execution (ASS13).

8. Action Planning II From the activation of the motor command features, planning obtains a handle to push-right and creates the intention Int(me)(push-right), which causes the features of push-right to receive further activation (ASS7). Figure 7.7 visualizes the intention structure after this last expansion.

9. Action Execution Then this intention is exercised in that its features in associative memory are activated further without output being inhibited (ASS10). This causes the agent to execute that motor command and push the right button (ASS13). As part of this process, other potentially activated motor commands need to be suppressed. This takes longer, the more these motor commands are activated. In this case, however, there are no other activated motor commands.

10. Feedback The new perceptual input activates the feature tactile-right as well as the features index-finger-right and right (ASS8), which makes the pushed-right proposition true in symbolic memory (ASS4). This in turn renders goals and intentions fulfilled.

This is but one example of the different possible experimental conditions of the individual and social Simon task. As we have seen in Section 7.2, these conditions differ along the following two dimensions:
Chapter 7. A Minimal Computational Architecture of Joint Action

Figure 7.7: The fully expanded intention structure in the no-conflict condition of the social Simon task.

1. Task distribution: Different experimental conditions require different instructions as to who is to act on which stimulus. This requires adjustments in the task representation, i.e. Equation 7.1, which implies changes to Equations 7.2, 7.3, and 7.4.

2. Stimulus: This requires changes in steps (4) and (5) of the procedure described above.

We apply these changes in the following to reconstruct the other experimental conditions of the individual and social Simon task.

7.5.3 Action-Conflict Condition in the Social Simon Task

An action conflict can occur when one of the (task-irrelevant) stimulus features overlaps with a corepresented action. In the example above, assume that in step (4) the red stimulus is presented on the left side instead of the right one. Recall that this is precisely the condition under which an action conflict is observed. This stimulus configuration would activate the left feature, which is among the features representing pushed-left. Activating the left feature therefore lifts the activation of the features representing pushed-left above the threshold level, which then leads to an activation gain of the features representing the push-left motor command in step (5) of the above procedure. Therefore, in step (9) the activation of the push-left motor command competes with the activation of the push-right motor command, which is actually to be executed. This yields an increase in response time—the Simon effect.

7.5.4 Task-Conflict Condition in the Social Simon Task

To reconstruct the condition under which a task conflict obtains requires more changes to the example above. Recall that a task conflict is observed when (task-relevant) features of the
observed stimulus match the preconditions of two different task rules. The following joint intention is the one represented in the condition under which a task conflict is observed:

\[
\text{Goal}_{\text{you,me}}(\text{Do}_{\text{you,me}}(\text{Simon-task}(\text{You, Me, Right, Red})))
\]  

(7.5)

Let us assume \(\text{nice(You)}\) is true and the intentions in Equations 7.3 and 7.4 are represented appropriately for this condition, specifying the agent’s own and the other agent’s part of the task:

\[
\text{Goal}_{\text{you}}(\text{attr(Location)} = \text{Right} \leftrightarrow \text{pushed-left})
\]  

(7.6)

\[
\text{Goal}_{\text{me}}(\text{attr(Color)} = \text{Red} \leftrightarrow \text{pushed-right})
\]  

(7.7)

If now a red stimulus is presented on the right side, both goals need to be fulfilled by intentions as in step (8) above. Using the inverse model to create both these intentions yields an increase in the activation level of both the features representing the \(\text{push-left}\) and \(\text{push-right}\) motor commands. Therefore, an increase in reaction time emerges from the competition in activation of these two motor commands in step (9). Recall that feature activation due to stimulus perception is not sufficient to overcome the threshold while feature activation due to action planning is. Hence, the \(\text{push-left}\) motor command here receives overall more activation than in the action-conflict condition because of its activation by the inverse model during action planning. Hence, the competition effect and therefore the Simon effect is larger in the task conflict than in the action-conflict condition, which is in line with the results of Sebanz et al. (2005).

### 7.5.5 Action- and Task-Conflict Condition in the Social Simon Task

To represent the condition under which both types of conflict add up, the following joint intention is required:

\[
\text{Goal}_{\text{you,me}}(\text{Do}_{\text{you,me}}(\text{Simon-task}(\text{You, Me, Left, Red})))
\]  

(7.8)

Let us assume \(\text{nice(You)}\) is true and the intentions in Equations 7.3 and 7.4 are represented appropriately for this condition, specifying the agent’s own and the other agent’s part of the task:

\[
\text{Goal}_{\text{you}}(\text{attr(Location)} = \text{Left} \leftrightarrow \text{pushed-left})
\]  

(7.9)

\[
\text{Goal}_{\text{me}}(\text{attr(Color)} = \text{Red} \leftrightarrow \text{pushed-right})
\]  

(7.10)
A red stimulus is presented on the left side. Hence both task rules become relevant and there is an additional activation of the push-left motor command because the stimulus appears on the left side and therefore contributes to the activation of the features representing pushed-left. In effect, the push-left motor command is activated both according to the action-conflict condition and the task-conflict condition. Overall, this causes the activation level of this motor command to be increased by the sum of the increase in the two conditions.

7.5.6 No Corepresentation

If nice(You) is false, Equation 7.3 is not represented and hence in none of the conditions above any conflict would occur because push-left would never receive any activation. This reflects the findings of Hommel et al. (2009) that a Simon effect is observed only when the partner is perceived as a likable person. Similarly, if the participant is instructed to carry out one part of the task alone, no corepresentation is obtained and therefore a Simon effect does not occur.

7.5.7 Individual Simon Task

The four conditions of the individual Simon task (no-conflict, action-conflict, both with and without corepresentation) mirror the same conditions in the social Simon task. The difference is that intentions relate to the participant itself only and not to any partner. Other than that, the procedure is the same as the one above.

7.5.8 Summary

Table 7.4 summarizes in which steps of the procedure above the two motor commands push-left and push-right and their corresponding effects pushed-left and pushed-right are activated in the different conditions of the individual and social Simon tasks.

The first thing to note is that the features of the push-left motor command are never activated in the no-conflict conditions of the social and the individual Simon task, and they are also not activated when there is no corepresentation. Therefore, there is no Simon effect created in these conditions. In the action-conflict condition with corepresentation of the individual Simon task and the social Simon task, the features of the push-left motor command receive activation from corepresentation (step 3) and from the activation of the feature left (step 5). Therefore, the push-right motor command competes with the push-left motor command during step 9 and hence there is a Simon effect. In the task-conflict condition with corepresentation, both the features of the push-left and the features of the push-right motor command receive activation from corepresentation (step 3) and action planning (steps 7 and 8). The activation of the push-left motor command in this case is larger than the one.
7.5. Representing the Individual and Social Simon Task

in the action-conflict condition. Therefore, there is more activation to be suppressed and the Simon effect in the task-conflict condition can be seen to be larger than the one in the action-conflict condition. In the task-conflict condition with corepresentation, both types of Simon effects add up because the individual activation contributions of the action and the task conflict to push-left add up.

At least qualitatively, these results are compatible with the empirical results we described in Section 7.2. The only minor exception is that the increase in reaction time in the task and action-conflict condition was found by Sebanz et al. (2005) to be larger than the sum of the reaction time increases observed in the action and in the task-conflict condition. We do not account for this overadditivity at the architectural level in order to leave open the exact implementation of activation suppression, which appears to be the most likely component responsible for this result.

Having described how the representation chosen here plays out in those various conditions of the Simon task, we briefly reiterate the role of the individual assumptions we made in the beginning of this section:

**ASS1** We have attempted to keep our requirements of the logic that describes the content of symbolic memory minimal. We believe that this choice is appropriate for the demonstration here but we do acknowledge that a complete formal treatment of this instantiation of our architecture would require a more careful selection and handling of this logic. This is the type of simplification that we believe is characteristic of computational-conceptual models.

**ASS2, ASS3, ASS4** The interaction between symbolic and associative memory is crucial to our architecture, and it is a powerful feature. However, with this power comes complexity: The more elaborate the interaction between both memories is supposed to be, the more complex the structure of associative memory and the mapping between elements of both memories have to be. To represent the individual and social Simon task, we were able to limit this complexity and build on a subset of the possibilities that the architecture offers.

**ASS5, ASS6** In a full-blown instantiation of the architecture, there has to be a learning mechanism that establishes the mapping between elements of symbolic memory and structures of features in associative memory. Also, there needs to be a mechanism for learning the associations between different features in associative memory. However, learning effects are not in the focus of research on the individual and social Simon task and are therefore neglected in our model as well.
### Table 7.4

<table>
<thead>
<tr>
<th>Step</th>
<th>Feature Set</th>
<th>Social Task</th>
<th>Individual Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>With Corep.</td>
<td>Without Corep.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With Corep.</td>
<td>Without Corep.</td>
</tr>
<tr>
<td>3 (Task Expansion)</td>
<td>pushed-left</td>
<td>+ + + +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>push-left</td>
<td>+ + + +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>pushed-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td>5 (Feature Activation)</td>
<td>pushed-left</td>
<td>+ + + +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>push-left</td>
<td>+ + + +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>pushed-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>push-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td>7 (Action Planning I)</td>
<td>pushed-left</td>
<td>+ +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>push-left</td>
<td>+ +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>pushed-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>push-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td>8 (Action Planning II)</td>
<td>pushed-left</td>
<td>+ +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>push-left</td>
<td>+ +</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>pushed-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>push-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td>9 (Action Execution)</td>
<td>pushed-left</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>push-left</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>pushed-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
<tr>
<td></td>
<td>push-right</td>
<td>+ + + +</td>
<td>+ + + +</td>
</tr>
</tbody>
</table>

Table 7.4: This table shows which sets of features are being activated in each of the relevant steps of the procedure above in the different conditions of the individual and social Simon task. The condition names are abbreviated here: NC stands for “No Conflict”, AC for “Action Conflict”, and TC for “Task Conflict”. A plus in a cell means that the respective feature set gains in activation during that step of the procedure.
7.5. Representing the Individual and Social Simon Task

ASS7 This assumption is central because it implements the mechanism by which shared task representations at the intentional level engage in mechanisms at the non-intentional level. Thereby, action and task conflicts become possible.

ASS8 This assumption implements a basic perception mechanism, which also plays a critical role in the invoking of action conflicts.

ASS9, ASS10 These assumptions allow the intentional level to use mechanisms at the non-intentional level for action planning and execution.

ASS11 This assumption ensures two things:

- An overlap between stimulus and action effect features does not by itself cause the activation of a motor command and an interference effect.
- Corepresentation alone does not cause any activation of motor commands and therefore does not cause any interference effect by itself.

These two points are important in ensuring the appropriate activation levels of motor commands as shown in Table 7.4, and they follow from the interpretation of the results of Sebanz et al. (2005).

ASS12 If action planning was not able to overcome the activation threshold between the features of pushed-left (pushed-right) and push-left (push-right), the intentional level would not be able to employ the inverse model for action planning.

ASS13 This assumption ensures that no activated motor command is executed that is not called for by the intentional level. An alternative realization of this would be to have the intentional level ensure for appropriate output inhibition. For the representation of the Simon task, this is not necessary but it might be for other instantiations of this architecture.

This section has highlighted the role of a few components of this architecture. In particular, symbolic memory and the intentional level enable an intuitive and expressive description of the individual and social Simon task and the social relationship between participants. The intentional level also allows for complex goal-oriented planning required by joint actions. Such mechanisms could certainly not have been represented at the non-intentional level without excessive complexity. Yet, it is associative memory at the non-intentional level that provides an appropriate representation of the feature-overlap between different representations, which plays a crucial role in action and task conflicts and hence the Simon effect. The interplay between symbolic and associative memory enables corepresentation determined at the intentional level to condition associative memory for these effects to occur.
This closes our demonstration that our architecture is able to reproduce the human performance in the individual and social Simon task—an experiment that has revealed phenomena spanning different levels of coordination in joint action. This is first evidence that this architecture can account for joint actions whose analysis would require an integration of the philosophical and the social and ecological psychological perspectives.

7.6 Discussion and Conclusions

This chapter has introduced a computational architecture of joint action in terms of a computational-conceptual model. The key feature of this architecture is that it accounts for high-level coordination mechanisms that enable joint planning as well as low-level mechanisms that enable the precise coordination of actions in time and space. The architecture also accounts for interactions between these high- and low-level coordination mechanisms. The requirements of this architecture were identified from a discussion of relevant philosophical and psychological research and an analysis of a particular experimental task. The previous section has demonstrated how the architecture is able to reproduce the human performance in this experimental task.

The key component of this architecture is a dual-process memory model, which enables the interaction of two levels of coordination: At the intentional level, mechanisms of practical reasoning and joint intentions employ symbolic reasoning and memory to coordinate actions with a coactor. At the non-intentional level, low-level coordination mechanisms are enabled by an associative memory and its motor control models. Elements of symbolic and associative memory interact, which leads to the emergence of coordination phenomena that are not possible with either of the two levels in isolation. We have left some of the details open and demonstrated in Section 7.5 how detail can be added to create a concrete instantiation of this architecture. For example, we have not completely specified exact learning, memory retrieval, and practical reasoning mechanisms.

Nevertheless, this architecture is distinguished by providing a computational model of joint action that accounts for high- and low-level coordination mechanisms and their interactions. We contribute to research on joint action by providing a more precise model of the interplay between different levels of coordination in a language suitable for describing the involved mechanisms. This architecture is loosely inspired by a conceptual architecture of joint action by Vesper et al. (2010), which is an effort towards an integration of mechanisms of emergent and planned coordination. We have fleshed out the mechanisms their architecture includes. In particular, the intention structure of our architecture enables actors to relate their part of the task to the overall joint action and the tasks of coactors. Like Vesper et al., we account for processes that monitor the progress of the joint action towards the overall goal.
7.6. Discussion and Conclusions

or individual tasks (intention monitoring), and towards the execution of individual actions (action monitoring). We also account for processes that enable predictions with regard to how the coactors’ combined actions achieve the joint goal (intention prediction) and with regard to the effects of individual actions (action prediction). We have also commented on possible applications of coordination smoothers.

With regard to RQ1, this architecture is a step towards a more complete computational representation of the grounding model of cultural transmission because it expands on the role of joint action in cultural transmission. We shall see in the remainder of this section which modes of cultural transmission can be represented by the architecture introduced in this chapter but not by the one discussed in the previous chapter. In fact, the following discussion goes beyond the current theorizing concerning the grounding model of cultural transmission. Thereby, this architecture can contribute to the refinement of the grounding model of cultural transmission and inspire further empirical research, which is a contribution to RQ2. However, as a computational-conceptual model the architecture itself is underspecified. The architecture can be evaluated rigorously only through concrete models that are constructed in accordance with its assumptions. In that, our architecture is in fact a framework that can be used to inform the construction of other computational models of cultural dynamics, potentially with the purpose of simulation. As argued previously, simulation can play a complementary role to experiments in the study of a system—cultural dynamics in this case. On the nano-to-macro axis, the presented architecture is located at the nano and micro level.

We close this chapter by discussing briefly how this architecture can accommodate for further phenomena of joint action, in particular those that are relevant to cultural transmission. We comment first on the architecture’s ability to accommodate for mimicry and imitation effects. Subsequently, we discuss the influence of the low-level coordination mechanisms considered here on the symbolic transmission of culture that the previous chapter was concerned with. We also describe the ability of the architecture to accommodate for the embodied transmission of culture. This discussion is by no means an exhaustive one and only supposed to indicate what is possible with this architecture. We close the chapter with an outlook to future work.

7.6.1 Mimicry and Imitation

Both mimicry and imitation involve copying behavior from another individual by observation. While mimicry is a subconscious process, imitation is always intentional.

There is ample evidence that people tend to unconsciously mimic each other during joint actions (see citations in Knoblich et al., 2011). Mimicry ranges from the copying of bodily movements and facial expressions to the alignment of accent, tone of voice, and the use of words and grammatical structure (see citations in Böckler et al., 2010). The apparent function
of mimicry is the establishing of social relationships between coactors by the increase of liking and of their willingness to cooperate (Knoblich et al., 2011). At least the mimicking of bodily movements could be attributed to common representations for perception and action, which create a tendency to execute perceived actions (Böckler et al., 2010). These direct links between perception and action are a crucial part of our architecture, which is hence able to account for at least some phenomena of mimicry.

Imitation is a crucial process of social learning and hence cultural transmission as it enables the intentional acquisition of procedural knowledge. Children employ imitation in their development from an early age (Tomasello et al., 2005). Most importantly, from a certain age on children tend to imitate action goals, not necessarily the means to achieve these goals. For example, children can observe an actor achieving a certain goal but then apply a different action to achieve the same goal—thereby essentially imitating the goal of that observed action but not the action itself (Bekkering et al., 2000). This indicates that imitation in humans is inherently goal-directed. Böckler et al. (2010) identify three requirements for goal-directed imitation: 1) a system to map observed actions onto one’s own action repertoire, 2) the ability to distinguish intentions of self and other, and 3) the ability to divide action sequences into meaningful units. The first requirement can be fulfilled by common representations for action and perception, which is implemented by the non-intentional level of our architecture. The second requirement is implemented by the intention structure at the intentional level. The third requirement implies a mechanism at the intentional level that extracts sequences of observed motor commands from the non-intentional level. Hence, in principle our architecture is able to account for goal-directed imitation.

7.6.2 Symbolic Transmission of Culture

The architecture in this chapter extends the one introduced in the previous chapter by adding low-level coordination mechanisms that are not mediated by any higher-level cognition. The previous subsection has indicated how these mechanisms can enable the cultural transmission of procedural knowledge by imitation. However, these mechanisms also enable phenomena not considered previously in this thesis that play a role in the symbolic transmission of culture.

The grounding model of cultural transmission makes the implicit assumption that the alignment of mental models during the grounding process is an intentional process driven by a joint goal and involving symbolic communication. This is the view taken by Clark (1996c) and the one adopted in the previous chapter. The mechanisms at the intentional level of our architecture represent this idea.
7.6. Discussion and Conclusions

Pickering and Garrod (2004) suggest that mimicry\textsuperscript{1} leads to automatic, subconscious alignment at all linguistic levels of the conversation: Interlocutors tend to use the same words and grammatical structures and also align their accent and speech rate. Pickering and Garrod attribute these alignment effects to common representations for language production and comprehension. If a word is perceived, for example, its representation is activated and therefore the word is more likely to be used during subsequent production. This obviously mirrors the more general system of common representations for action and perception we have discussed previously. In our architecture, associative memory provides for this common representation of language production and comprehension: If a word is perceived, for example, relevant features are activated in associative memory, which causes related symbols to be activated in symbolic memory. Thereby, language production at the intentional level is modulated by comprehension.

Based on empirical evidence from previous studies, Pickering and Garrod (2004) suggest that alignment at one level facilitates alignment at other levels and ultimately the alignment of mental models. For example, when interlocutors use the same words, they also tend to look at the world in a similar way. Again, the mechanism enabling inter-level alignment in our architecture is associative memory, which can encode associations between various different representations of a concept, for example between the semantic and grammatical aspects of a particular word. Hence, the non-intentional level plays a crucial role in the alignment of mental models—the grounding process.

Garrod and Pickering (2009) extend the scope of alignment to non-linguistic levels: Interlocutors also mimic each other’s posture and they align their gaze during conversations. These alignment effects can also facilitate the alignment of mental models. In effect, the mechanisms yielding alignment at different linguistic and non-linguistic levels of the conversation serve as coordination smoothers in the sense of Vesper et al. (2010).

Considering the discussion here, our architecture can contribute to understanding how the grounding process during interactions is affected by low-level coordination mechanisms.

7.6.3 Implicit Transmission of Embodied Culture

Cultural information is not only encoded in symbolic information but is often also encoded and embodied in nonverbal behaviors such as bodily postures and movements (Leung and Cohen, 2007). This is what we call an embodied culture. The questions one would want to ask are, for example, when and how embodied cultural information is transmitted.

To illustrate this point, we rely on embodied attitudes as an example of embodied culture (Semin and Smith, 2008). Attitudes are often encoded by approach or retreat

\textsuperscript{1} Pickering and Garrod (2004) use the word “imitation” but we feel that their use of that word is more in line with our use of the word “mimicry”.

221
bodily movements such as leaning forward or backwards (Neumann et al., 2003). It has been observed that the learning of attitudes towards objects is influenced by the currently performed behavior if that behavior is associated with a positive or negative evaluation (see references in Niedenthal et al., 2005). For example, when a participant is nodding their head while perceiving a message, the participant develops a more positive attitude towards the content of the message than when they are shaking their head (Wells and Petty, 1980).

The question arises whether the imitation or subconscious mimicking of another person’s behavior can be a carrier for the transmission of embodied attitudes. For example, consider a coactor constantly leaning forward when presented with a certain stimulus and an observer mimicking/imitating this behavior without having a prior attitude towards this stimulus. Does this observer then develop a positive attitude towards this stimulus (because of the prior positive association with the action of leaning forward)?

In our architecture, a prior association between an evaluation and a particular behavior (e.g. leaning forward) would be represented in associative memory. If the agent performs such a behavior because of mimicry or imitation while being presented with a particular stimulus, features representing that stimulus and features representing the evaluation of the executed behavior are activated simultaneously. Because of the coactivation of the features representing the evaluation and the features representing the current stimulus, their association will be strengthened. This causes the implicit adoption of an attitude towards the current stimulus. Hence, the mechanisms present in our architecture do predict such a transmission of embodied culture. In fact, our architecture even predicts that observation of the behavior in question without actual mimicry is sufficient for this effect because observation alone already causes the activation of action representations and their evaluations. In fact, first empirical investigations have confirmed this hypothesis (Kashima, 2012).

7.6.4 Future Work

The architecture described here is a computational-conceptual model that allows a discussion based on computational terms. This is helpful for an inter-disciplinary dialogue. A complete implementation of this model, however, would require a more rigorous formalization. This is one of the goals for future research. In particular, we are interested in fleshing out the details necessary for representing the embodied transmission of culture. Of course, one can argue that a lot of other details are missing from an architecture of joint action. For example, emotions are known to play a crucial role in human behavior. However, we have here focused on a few key mechanisms, which we believe to be enablers of other optional mechanisms. The effect of emotions, for example, could be integrated with the practical reasoning mechanism of the intentional level (see Dignum et al., 2008, for such an approach).
An Implementation of the Joint Action Architecture

In accordance with our outline in Section 3.4, this chapter provides a simplified version and implementation of the joint action architecture presented in the previous chapter. The core components of this architecture are a higher-level practical reasoning module and a lower-level motor control module. A symbolic memory underlies the higher level and an associative memory the lower level. Via synergies between these two memories, both levels interact to give rise to mechanisms and phenomena of joint action that would otherwise not be possible. Our analysis of the architecture has relied on empirical results obtained from a particular experimental task—the Simon task. The Simon task has revealed interaction effects between higher- and lower-level mechanisms of joint action.

In the last part of the previous chapter we have sketched how our architecture could be instantiated to represent the Simon task. In this chapter, we provide the description of a concrete computational model of the architecture that reproduces the Simon effect qualitatively. We hence provide evidence that the mechanisms postulated in the last chapter can indeed produce the phenomena observed in experiments involving the Simon task and its derivatives. This exercise also demonstrates the benefit of auxiliary models such as the architecture described in the previous chapter in the development of computational models. In particular, by relying on a theoretically and empirically grounded architecture, the roots of the model presented here are clear. The use of auxiliary models was claimed to be beneficial in the development of agent-based models in Section 3.2.2. This exercise also suggests that the architecture presented in the previous chapter can be instantiated in different concrete models, each of which can contribute to the study of cultural dynamics. This is the contribution of this chapter to RQ1.
We begin this chapter with a description of our model in Section 8.1. This is followed by an evaluation of the model in Section 8.2 and a discussion of related work in Section 8.3. We conclude this two-chapter treatise of the joint action architecture by a more detailed discussion of limitations and avenues for future work in Section 8.4.

8.1 Model

In the following, we describe a concrete implementation of the architecture presented in the previous chapter that represents the various conditions of the Simon task. The general idea of the model was already outlined in Section 7.5. In this section, we provide a detailed description of an implementation of that idea. For this presentation we make a few minor simplifications, which however maintain the form of the model.

8.1.1 Associative Memory

The constituents of associative memory are processing units representing features and their weighted connections. Recall that there are mainly two different pools of units: A pool of units representing stimulus and action effect features, and a pool of units representing motor commands. Let $\mathcal{U}$ be the set of units corresponding to stimulus and action effect features, and $\mathcal{V}$ the pool of units corresponding to motor command features. Let $\mathcal{T}$ be the set of threshold units that mediate activation propagation between features in $\mathcal{U}$ and features in $\mathcal{V}$. Let $\mathcal{N} = \mathcal{T} \cup \mathcal{U} \cup \mathcal{V}$ be the set of all units in the network with $\mathcal{T} \cap \mathcal{U} = \emptyset$, $\mathcal{T} \cap \mathcal{V} = \emptyset$, and $\mathcal{U} \cap \mathcal{V} = \emptyset$. Recall from Chapter 5 that $w_{ij} (i, j \in \mathcal{N})$ denotes the weight of the link from unit $j$ to unit $i$, $x_i(t)$ the external input to unit $i$ at time (tick) $t$, $s_i(t)$ the net input to unit $i$ at time $t$, and $a_i(t)$ the activation level of unit $i$ at time $t$. Here, activations are restricted to the interval $[-1, 1]$. In the following, we describe for each of the type of units how their net input and activation level is calculated. We consider the state of the network for each time $t$ in an interval $[t_0, t_1]$.

Stimulus and action effect features do not receive any input from within the network because we do not consider any forward model here. Therefore we have the following net input and activation (see also Chapter 5) for all $i \in \mathcal{U}, t \in [t_0, t_1]$:

\begin{align}
s_i(t) &= x_i(t) \\
a_i(t) &= f(s_i(t)) = \frac{2}{1 + e^{-s_i(t)}} - 1
\end{align}

Here and in the following, $f(\cdot)$ is the sigmoid function projected onto the range $[-1, 1]$. In effect, stimulus and action effect features simply forward input onto other network units.
8.1. Model

For all threshold units \(i \in T\) and all \(t \in [t_0, t_1]\), we have:

\[
s_i(t) = \sum_{j \in \mathcal{N}} w_{ij} a_j(t) \tag{8.3}
\]

\[
a_i(t) = \begin{cases} 
  f(s_i(t)) & \text{if } s_i(t) \geq \Omega \\
  0 & \text{otherwise}.
\end{cases} \tag{8.4}
\]

Hence threshold units accept input only from other units in the network and they are only activated if their net input exceeds a threshold \(\Omega\).

Motor command units in the set \(V\) are arranged in a recurrent, auto-associative network, which settles into that previously learnt pattern that is closest to the current input. This settling behavior reflects that multiple motor command patterns might compete for activation. We build on a recurrent network structure proposed by Rumelhart et al. (1986, chapter 17). As described in Chapter 5, this type of network has successfully been used as a model for human memory and cognition. To accommodate for the time-varying activation behavior, we assume that every time tick \(t\) is split into a number of cycles from the perspective of the recurrent network component. Algorithm 8.1 shows the procedure of activation propagation in the subnetwork spanned by the units in \(V\). While the change in activation since the last cycle is above a certain threshold (line 3), the net input and activation for each unit \(i \in V\) is reassessed (lines 5 to 11). Units in the recurrent network can receive input from external and internal sources. This leads to the input calculation in line 5, where \(s_i(t)\) represents the net input to unit \(i\) at time \(t\) and \(\hat{s}_i(t)\) the total input consisting of the sum of net input and external input. This input is scaled by a factor \(E\), which reflects its excitation level. The equations in line 8 and 10 cause activation values to approach the minimum activation value \(-1\) or the maximum activation value \(1\) slowly and not to exceed them.

Provided with an input that is part of a previously learnt pattern, the recurrent network will settle into an activation pattern that reflects the whole of this previously learnt pattern. In fact, the learning of a pattern consists of (1) providing this pattern as an input, (2) propagating activation through the network by the \textsc{UpdateRecurrent} operation until activation levels have settled, and (3) adjusting the weights in the network. The weight adjustment procedure is shown in Algorithm 8.2. The idea is that a unit \(i\) needs to receive more input from other units if its net input falls short of the externally provided input/target. In that case, the weights of links from other units to this one are increased. Otherwise, they are decreased. Note that external input also plays the role of the target because this is an auto-associative network, unlike the one presented in Chapter 5. In an auto-associative network, the input pattern corresponds to the pattern that is to be learnt. Typically, the learning procedure needs to be repeated multiple times for each pattern to be learnt.
Chapter 8. An Implementation of the Joint Action Architecture

Algorithm 8.1: \texttt{UPDATE\textsc{RECURRENT}}(\mathcal{N}, \mathcal{V}, w, t, x(t), a(t), D, E, \varepsilon)

\begin{enumerate}
\item \textbf{Data:}
- Set of network units \(\mathcal{N}\).
- Set of units that are part of the recurrent network \(\mathcal{V} \subset \mathcal{N}\).
- Weight matrix \(w\) of size \(|\mathcal{N}| \times |\mathcal{N}|\).
- Current time \(t\).
- External input \(x_i(t)\) to all units \(i \in \mathcal{V}\) at time \(t\).
- Activation \(a_i(t)\) of all units \(i \in \mathcal{N}\) at time \(t\).
- Activation decay parameter \(D \in [0, 1]\).
- Excitation parameter \(E \in [0, 1]\).
- Change threshold parameter \(\varepsilon\).

\item \textbf{Result:}
- Updated activation \(a_i(t)\) of all units \(i \in \mathcal{V}\) at time \(t\).
- Updated net input \(s_i(t)\) of all units \(i \in \mathcal{V}\) at time \(t\).
- Number of \texttt{cycles} required for settling.

begin
\item \texttt{cycles} ← 0
\item \textbf{while} \(|\Delta a(t)| \geq \varepsilon\) \textbf{do}
\item \texttt{s}(t) ← \(\sum_{j \in \mathcal{N}} w_{ij} a_j(t)\) for all \(i \in \mathcal{V}\)
\item \texttt{\hat{s}}(t) ← \(E \times (s_i(t) + s_i(t))\) for all \(i \in \mathcal{V}\)
\item \textbf{for} \(i \in \mathcal{V}\) \textbf{do}
\item \textbf{if} \(\hat{s}_i(t) > 0\) \textbf{then}
\item \(\Delta a_i(t) \leftarrow \hat{s}_i(t)(1 - a_i(t)) - Da_i(t)\)
\item \textbf{else}
\item \(\Delta a_i(t) \leftarrow \hat{s}_i(t)(a_i(t) + 1) - Da_i(t)\)
\item \(a_i(t) \leftarrow a_i(t) + \Delta a_i(t)\)
\item \texttt{cycles} ← \texttt{cycles} + 1
\item \textbf{return} \(a(t), s(t)\)
end
\end{enumerate}

We denote the operation that propagates activation through the whole network by \texttt{UPDATE\textsc{NETWORK}}(\mathcal{N}, w, x(t), a(t), D, E, \Omega), where \(\mathcal{N}\) is the set of network units, \(w\) the weight matrix for links between network units, \(x(t)\) the external input to all units at current time \(t\), \(a(t)\) the existing activation of all units at time \(t\), \(D\) and \(E\) the parameters of the recurrent network as required by the \texttt{UPDATE\textsc{RECURRENT}} operation, and \(\Omega\) the threshold used by all threshold units.

Having described the constituents of associative memory, let us briefly indicate how they are orchestrated in our model. We will return with more detail later when we describe how we set up the architecture to represent the Simon task. Basically, sets of different units in \(\mathcal{V}\) correspond to different motor commands. Some of the units of each motor command receive input via threshold units (\(T\)) from stimulus and action effect features in \(U\). Because action effect and stimulus features belong to the same set of features \(U\), there can be an overlap between the representation of stimuli and action effects. Hence, stimulus attributes can
Algorithm 8.2: LEARNRECURRENT(\(\mathcal{N}, \mathcal{V}, w, \eta, a(t), x(t), s(t)\))

Data:
- Set of network units \(\mathcal{N}\).
- Set of units that are part of the recurrent network \(\mathcal{V} \subset \mathcal{N}\).
- Weight matrix \(w\) of size \(|\mathcal{N}| \times |\mathcal{N}|\).
- Learning rate \(\eta\).
- Activation \(a_i(t)\) of all units \(i \in \mathcal{N}\) at time \(t\).
- External input \(x_i(t)\) of all units \(i \in \mathcal{N}\) at time \(t\).
- Net input \(s_i(t)\) of all units \(i \in \mathcal{N}\) at time \(t\).

1 begin
2   for \(i, j \in \mathcal{V}, i \neq j\) do
3     \(w_{ij} \leftarrow \eta(x_i(t) - s_i(t))a_j(t)\)

contribute to an activation of motor commands. This architecture has also been illustrated in Figure 7.6.

8.1.2 Symbolic Memory

Symbolic memory contains beliefs, goals, and intentions, whereby the latter two make reference to actions. The beliefs of the agent are retained in a knowledge (belief) base \(KB\) of sentences in a propositional logic \(\mathcal{L}\) with the set of propositional variables \(\Sigma\) denoted by \(p, q, \ldots\) and the normal logical connectives \(\neg, \land, \lor, \rightarrow, \leftrightarrow\). Later we use more descriptive names for the propositional variables. We denote semantic entailment of a sentence \(\phi \in \mathcal{L}\) by a set of sentences \(S\) by \(S \models \phi\). Inference is implemented by simple model checking through truth-table enumeration, which is sound and complete. We denote syntactic entailment of \(\phi \in \mathcal{L}\) from a set of sentences \(S\) using this inference algorithm by \(S \vdash \phi\). By \textsc{Tell}(\(KB, \phi\)) we denote the operator that adds sentence \(\phi\) to \(KB\), i.e. \(KB \leftarrow KB \cup \{\phi\}\). By \textsc{Retract}(\(KB, \phi\)) we denote the operator that removes sentence \(\phi\) from \(KB\), i.e. \(KB \leftarrow KB \setminus \{\phi\}\). For the implementation of the knowledge base we rely on a module provided by Russell and Norvig (2003).\(^1\) Maintaining the consistency of this knowledge base is the responsibility of its user.

The other primitive constituents of symbolic memory are symbols representing actions. We distinguish primitive from complex actions. Primitive actions relate to a motor command directly, complex actions consist of a set of subgoals. A primitive action is a tuple \(\langle \phi, F \rangle\) where \(\phi \in \mathcal{L}\) is the effect of executing that action and \(F \subseteq \mathcal{N}\) the set of features in associative memory this action corresponds to. Hence we have \(A^P \subset \mathcal{L} \times \mathcal{P}(\mathcal{N})\) for the set of primitive actions. Let \(G\) be the set of all possible goals as we will describe in the next paragraph. A complex action is a tuple \(\langle \phi, G \rangle\) where \(\phi \in \mathcal{L}\) is the effect of executing that action and \(G \subseteq \mathcal{G}\) the set of subgoals of this action. Hence we have \(A^C \subset \mathcal{L} \times \mathcal{P}(\mathcal{G})\) for the set of complex

---

\(^1\)http://aima-python.googlecode.com/svn/trunk/logic.py
actions. We denote by $A = A^p \cup A^c$ the set of all actions and by $A \subseteq A$ the set of actions available to the agent.

In the following, we denote by $\text{Effect} : A \rightarrow L$ the operator that determines the effect of executing an action. We denote by $\text{ActionFeatures} : A^p \rightarrow \mathcal{P}(N)$ the operator that yields the features associated with an action. By $\text{Subgoals} : A^c \rightarrow \mathcal{P}(\mathcal{G})$ we denote the operator that delivers the subgoals that constitute a particular complex action.

Based on these primitives, we can define goals and intentions, which make reference to the agents responsible for them. By $\mathcal{H}$ we denote the set of agents considered to be involved in the joint action by our focal agent. We distinguish three types of goals: achievement, performance, and maintenance goals. An achievement goal or a maintenance goal is a tuple $\langle \phi, H, s \rangle$ where $\phi \in L$ is the goal state to be achieved, $H \subseteq \mathcal{H}$ the set of agents that are associated with this goal, and $s \in \{\text{True}, \text{False}\}$ denotes whether the goal has been selected for active pursuit by this agent. Hence we have for the set of achievement goals $G^A \subset L \times \mathcal{P}(H) \times \{\text{True}, \text{False}\}$, and for the set of maintenance goals $G^M \subset L \times \mathcal{P}(H) \times \{\text{True}, \text{False}\}$. A performance goal is a tuple $\langle \alpha, H, s \rangle$ where $\alpha \in A$ is an action to perform, $H \subseteq \mathcal{H}$ the set of agents that are associated with this goal, and $s \in \{\text{True}, \text{False}\}$ denotes whether the goal has been selected for active pursuit by this agent. Hence we have for the set of performance goals $G^P \subset A \times \mathcal{P}(H) \times \{\text{True}, \text{False}\}$. We denote by $G = G^A \cup G^P \cup G^M$ the set of all possible goals. The current set of goals of the agent is denoted by $G \subset G$.

The objective of an achievement or maintenance goal can be retrieved by the operator $\text{State} : G^A \cup G^M \rightarrow L$. The objective of a performance goal can be retrieved by the operator $\text{Action} : G^P \rightarrow A$. The agents assumed to be responsible for a goal are denoted by the operator $\text{Agents} : G \rightarrow \mathcal{P}(\mathcal{H})$. Whether a goal has been selected for active pursuit or not is denoted by the operator $\text{Selected} : G \rightarrow \{\text{True}, \text{False}\}$.

An intention is a tuple $\langle \alpha, H \rangle$ where $\alpha \in A$ is an action to perform and $H \subseteq \mathcal{H}$ the set of agents that are associated with this intention. Hence we have for the set of possible intentions $I \subset A \times \mathcal{P}(H)$. We denote by $I \subset I$ the current set of intentions referenced by the agent. As in the case of goals, we denote by $\text{Agents} : I \rightarrow \mathcal{P}(\mathcal{H})$ the set of agents considered responsible for an intention. By $\text{Action} : I \rightarrow A$ we denote the operator that produces the action referenced by an intention.

### 8.1.3 Interaction between Symbolic and Associative Memory

Recall that there are four ways in which features in associative memory can receive input (see Section 7.5.1): (1) when a belief, goal, or intention is created (attitude representation); (2) when a stimulus is perceived (stimulus perception); (3) when the intentional level employs the inverse model at the non-intentional level by providing a goal state to retrieve an appropriate motor command (action planning); (4) when the intentional level invokes the execution of a
8.1. Model

primitive action (action execution). Recall that we only require that feature activations due to stimulus perception are translated into corresponding symbols (symbol activation). We do not make any further assumptions about the effect of feature activations in associative memory on symbolic memory. In this section, we describe stimulus perception, symbol activation, and attitude representation as well as the operations these mechanisms are based on. Action planning and execution are described in Section 8.1.4.

We denote by $\text{LITERALS TO FEATURES} : P \rightarrow \mathcal{P}(\mathcal{N})$ the operation that maps propositional literals in symbolic memory onto sets of features in associative memory. By $\text{FEATURES TO LITERALS} : \mathcal{P}(\mathcal{N}) \rightarrow P$ we denote the corresponding reverse operation that maps feature sets onto literals. We also require an operator $\text{FEATURES TO ACTIONS} : \mathcal{P}(\mathcal{N}) \rightarrow \mathcal{A}$ to retrieve the action associated with a particular set of features. The operator $\text{ACTION FEATURES}$ to retrieve the features associated with an action has already been given above. As discussed before, the mappings between literals and features as well as actions and features are due to learning, which we do not consider here.

Algorithm 8.3 shows the procedure of stimulus perception and consequential symbol activation. First, the input provided to units representing attributes of a perceived stimulus are incremented by a fixed amount $\gamma_{\text{stimulus}}$ (line 2). This additional input causes activation to be propagated through the network as discussed above (line 3). This, in turn, causes a change in the activation of some of the units in $\mathcal{N}$ (line 4). In the most obvious case, those units that have just received new external input will change their activation. We consider those units whose activation then exceeds a certain threshold $\Gamma$ (line 4). For subsets of these units (line 5), corresponding literals are searched (line 6). If any such literal is found, that literal’s complement is retracted from the knowledge base and the literal itself is added (lines 8 and 9).

Attitude representation proceeds similarly to lines 2 and 3 in Algorithm 8.3. If the created attitude is an intention, the action referenced by this intention is translated by $\text{ACTION FEATURES}$ into corresponding features. These features receive additional external input $\gamma_{\text{representation}}$. $\text{REPRESENT INTENTION}(e, \mathcal{N}, w, x(t), a(t), D, E; \gamma_{\text{representation}})$ denotes this operation, where $e \in \mathcal{I}$ is the intention, $\gamma_{\text{representation}}$ the additional external input to provide to the network units representing the action of this intention, and all other parameters have the meaning already introduced before.

If the created attitude is a goal, the literals contained in this goal are extracted. Each of these literals is translated into corresponding features by $\text{LITERALS TO FEATURES}$, which then receive additional external input $\gamma_{\text{representation}}$. This operation is denoted by $\text{REPRESENT GOAL}(g, \mathcal{N}, w, x(t), a(t), D, E; \gamma_{\text{representation}})$, where $g \in \mathcal{G}$ is the goal, $\gamma_{\text{representation}}$ the additional external input to provide to the network units representing the goal, and all other parameters have the meaning already introduced before.
Chapter 8. An Implementation of the Joint Action Architecture

Algorithm 8.3: \textsc{PerceiveStimulus}(KB,\mathcal{N},w,t,x(\cdot),a(t),F,\gamma_{stimulus},\Gamma,D,E,\Omega)

Data:
- The agent’s propositional knowledge base $KB$.
- Set of network units $\mathcal{N}$.
- Weight matrix $w$ of size $|\mathcal{N}| \times |\mathcal{N}|$.
- Current time $t$.
- External input $x_i(t')$ to all units $i \in \mathcal{N}$ for all $t' \in [t_0,t_1]$.
- Current activation $a_i(t)$ of all units $i \in \mathcal{N}$ at time $t$.
- Set $F \subseteq \mathcal{N}$ of features to activate.
- External input $\gamma_{stimulus}$ to add to the input of these features.
- Threshold of feature activations $\Gamma$ for features to be considered for the activation of symbols.
- Parameters $D$ and $E$ for activation propagation in the recurrent network.
- Threshold $\Omega$ applied by threshold units.

Result:
- Updated activation $a_i'(t)$ of all units $i \in \mathcal{N}$ at time $t$.
- Updated external input $x(\cdot)$.

\begin{algorithmic}
  \State $x_i(t') \leftarrow x_i(t') + \gamma_{stimulus}$ for all $t' \geq t, i \in F$
  \State $a'(t) \leftarrow \text{UPDATENETWORK}(\mathcal{N},w,x(\cdot),a(t),D,E,\Omega)$
  \State $F' \leftarrow \{i \in \mathcal{N}|a'_i(t) > a_i(t) \land a'_i(t) > \Gamma\}$
  \For{$F'' \in \mathcal{P}(F')$}
  \State $p \leftarrow \text{FEATURESTOLITERALS}(F'')$
  \If{such a $p$ is found}
  \State $\text{RETRACT}(KB,\neg p)$
  \State $\text{TELL}(KB,p)$
  \EndIf
  \EndFor
  \State return $a'(t),x(\cdot)$
\end{algorithmic}

In both cases of attitude representation, a change in the external input of a unit triggers the execution of the \textsc{UpdateNetwork} operator to propagate activation through the network. The procedure for the representation of beliefs in associative memory would proceed accordingly but we do not require this for the representation of the Simon task.

8.1.4 The Non-intentional Level

The non-intentional level provides mechanisms to plan and execute actions. Let us look at planning first. Algorithm 8.4 shows how the inverse model provides a means for planning motor commands given particular action effects. First, additional external input $\gamma_{planning}$ is provided to the features representing the action effects to achieve (line 2). Activation is then propagated through the network (line 3) and those motor command features in $\mathcal{V}$ that gained in activation are retrieved (line 4).

Once motor command features have been retrieved, they can be used to execute the corresponding action. Action execution proceeds by adding additional external input $\gamma_{execution}$ to the features representing that action. After activation within the recurrent network spanned...
Algorithm 8.4: INVERSEMODEL($\mathcal{N}, w, t, x(\cdot), a(t), F, \gamma_{\text{planning}}, D, E, \Omega$)

Data:
- Set of network units $\mathcal{N}$ with action effect features $\mathcal{U} \subset \mathcal{N}$ and motor command units $\mathcal{V} \subset \mathcal{N}$.
- Weight matrix $w$ of size $|\mathcal{N}| \times |\mathcal{N}|$.
- Current time $t$.
- External input $x_i(t')$ to all units $i \in \mathcal{N}$ for all $t' \in [t_0, t_1]$.
- Current activation $a_i(t)$ of all units $i \in \mathcal{N}$ at time $t$.
- Features $F \subseteq \mathcal{U}$ representing the action effect to achieve.
- External input $\gamma_{\text{planning}}$ to add to the input of the features in $F$.
- Parameters $D$ and $E$ for activation propagation in the recurrent network.
- Threshold $\Omega$ applied by threshold units.

Result:
- Motor command features $F' \subseteq \mathcal{V}$ that are activated after input has been provided to the action effect features in $F$.
- Updated activation $a_i'(t)$ of all units $i \in \mathcal{N}$ at time $t$.
- Updated external input $x(\cdot)$.

begin
1 $x_i(t') \leftarrow x_i(t') + \gamma_{\text{planning}}$ for all $t' \geq t, i \in F$
2 $a_i'(t) \leftarrow \text{UPDATENETWORK}(\mathcal{N}, w, x(t), a(t), D, E, \Omega)$
3 $F' \leftarrow \{i \in \mathcal{U}|a_i'(t) > a_i(t)\}$
4 return $F', a'(t), x(\cdot)$

by the units in $\mathcal{V}$ has settled, the motor command represented by the current activation of those units is executed. We call this operation EXECUTE($\mathcal{N}, w, t, x(\cdot), a(t), F, \gamma_{\text{execution}}, D, E$), where $\mathcal{N}, w, t, x(\cdot), a(t), D, E$ have the meaning as above and $F$ is the set of motor command features in $\mathcal{V}$ to provide with additional external input $\gamma_{\text{execution}}$.

8.1.5 The Intentional Level

The intentional level is driven by a simple modified interpreter of the BDI model (Rao and Georgeff, 1995). The BDI model was briefly mentioned in Section 3.2.2 and described in more detail in Section 6.2. Algorithm 8.5 provides a high-level overview of the interpreter loop. In every time tick (line 2), first newly available options are retrieved, in our case those goals of the agent that are unselected (line 3). Subsequently, the agent deliberates over these options to retrieve a set of goals that are to be selected and pursued actively (line 4). In our case, all unselected goals that are not currently achieved by an action referenced by any intention of the agent are selected. All of these goals are then marked as selected (line 5). For each of these goals that are not entailed by the knowledge base yet (line 6), means-end reasoning is employed to resolve an action that can achieve this goal (line 7). This potentially creates a new intention. We describe means-end reasoning in detail below. After means-end reasoning, the agent identifies any intention that it is the only agent of (line 8) and whose action is a primitive one, i.e. refers to a motor command (line 10). For such an action, first the
Chapter 8. An Implementation of the Joint Action Architecture

set of network units it is associated with are retrieved (line 11). Then, the action is executed (lines 12). No further actions are going to be executed in the same time tick (line 13). Note that while an agent can plan for coactors, it does not actually act on intentions that it is not the sole actor of. Finally, external events (e.g. percepts) are received (line 14) and impossible and successful attitudes are dropped (line 15). We describe the receipt of perceptions when we discuss the instantiation of this model for the representation of the Simon task in Section 8.2. The dropping of attitudes is a rather complex house-keeping task that we do not describe in detail here.

Algorithm 8.6 sketches the means-end reasoning procedure. The procedure first searches for a known action that could fulfill the given goal (lines 2 and 3). If this is successful, a new intention to execute this action is created (line 4). If this action is a complex action (line 5), its subgoals that are neither part of the agent’s goals yet nor entailed by the knowledge base are determined (line 6). These subgoals are then added to the goal set of the agent (line 7) and each of these goals is represented in associative memory (line 9). The procedure can then be terminated (line 10). If no existing action can be used to achieve the goal, the literals from the goal expression and their complements are extracted (line 11). Each such literal $\varphi$ is retained if goal $g$ would be entailed by the belief base $KB$ revised with literal $\varphi$ (line 12). For each of these literals its associated features are retrieved if any (line 14). Any non-empty feature set (line 15) is used as a goal for input to the inverse model in order to retrieve a set of newly activated features representing an appropriate motor command (line 16). If such a motor command is found (line 17), a high-level action representation (line 18) and an intention are constructed (line 19). Then, the intention is represented in associative memory (line 20).

8.2 Evaluation

In this section, we describe how the model presented in the previous section is instantiated to represent the different conditions of the Simon task. We first describe the general setup of the model, e.g. which memory elements we require. Then we exemplify how the model dynamics would play out given the model is setup to reflect a particular task condition. Finally, we show that the model can be setup to qualitatively reproduce the empirical results obtained from the various conditions of the Simon task.

8.2.1 Setup

Associative and symbolic memory provide representations of stimulus features (green color, red color, left position, right position) and of the effects and motor commands of the two available actions (push left and push right). Figure 8.1 displays the mapping between
8.2. Evaluation

Algorithm 8.5: \textsc{Loop}(\mathcal{KB}, G, I, A, \mathcal{N}, w, t_0, t_1, x(\cdot), a(\cdot), \gamma_{\text{planning}}, \gamma_{\text{execution}},
\gamma_{\text{stimulus}}, \gamma_{\text{representation}}, \Gamma, D, E, \Omega)

Data:
- The agent’s propositional knowledge base \( \mathcal{KB} \).
- The set of goals of the agent \( G \).
- The set of intentions of the agent \( I \).
- The set of actions available to the agent \( A \).
- Set of network units \( \mathcal{N} \).
- Weight matrix \( w \) of size \(|\mathcal{N}| \times |\mathcal{N}|\).
- Start time \( t_0 \), end time \( t_1 \).
- External input \( x_i(t') \) to all units \( i \in \mathcal{N} \) for all \( t' \in [t_0,t_1] \).
- Current activation \( a_i(t) \) of all units \( i \in \mathcal{N} \) at time \( t \).
- Additional external input \( \gamma_{\text{planning}} \) added to action effect features when employing the inverse model.
- Additional external input \( \gamma_{\text{execution}} \) added to motor command units when executing an action.
- Additional external input \( \gamma_{\text{stimulus}} \) added to stimulus features when perceiving a stimulus.
- Additional external input \( \gamma_{\text{representation}} \) to add to features due to the representation of associated symbols.
- Threshold of feature activations \( \Gamma \) for features to be considered for the activation of symbols after stimulus perception.
- Parameters \( D \) and \( E \) for activation propagation in the recurrent network.
- Threshold \( \Omega \) applied by threshold units.

1 begin
2 \hspace{1em} for \( t = t_0 \) to \( t_1 \) do
3 \hspace{2em} \( O \leftarrow \{g \in G|\neg \text{SELECTED}(g)\} \)
4 \hspace{2em} \( S \leftarrow \{g \in O|\exists e \in I(\text{ACHIEVES}(KB, \text{ACTION}(e), g))\} \)
5 \hspace{2em} \( S \leftarrow \{\langle \mathcal{H}, \text{True} \rangle|\langle \mathcal{H}, \text{False} \rangle \in (S \cap (G^A \cup G^I)) \cup \{\langle \mathcal{H}, \text{True} \rangle|\langle \mathcal{H}, \text{False} \rangle \in (S \cap G^P)\} \)
6 \hspace{2em} for \( g \in \{g \in S|KB \not\models g\} \) do
7 \hspace{3em} \( I \leftarrow \text{MeansEndReasoning}(KB, G, I, A, \mathcal{N}, w, t, x(\cdot), a(\cdot), g,
\gamma_{\text{planning}}, \gamma_{\text{representation}}, D, E, \Omega) \)
8 \hspace{2em} for \( e \in \{e \in I|\text{AGENTS}(e) = \{me\}\} \) do
9 \hspace{3em} \( \alpha \leftarrow \text{ACTION}(e) \)
10 \hspace{3em} if \( \alpha \in A^P \) then
11 \hspace{4em} \( F \leftarrow \text{ACTIONFEATURES}(\alpha) \)
12 \hspace{4em} \( \text{EXECUTE}(\mathcal{N}, w, t, x(\cdot), a(\cdot), F, \gamma_{\text{execution}}, D, E) \)
13 \hspace{3em} \text{break} \)
14 \hspace{2em} \( \text{GETEXTERNALEVENTS}(KB, \mathcal{N}, w, t, x(\cdot), a(\cdot), \gamma_{\text{stimulus}}, \Gamma, D, E, \Omega) \)
15 \hspace{2em} \( \text{DROPATTITUDES}(KB, G, I) \)

the elements of both memories and the interaction between features in associative mem-
Chapter 8. An Implementation of the Joint Action Architecture

Algorithm 8.6: MEANSENDREASONING(KB, G, I, A, N, w, t, x, (.), a, t), g, γplanning, γrepresentation, D, E, Ω)

Data:
- The agent’s propositional knowledge base KB.
- The set of goals of the agent G.
- The set of intentions of the agent I.
- The set of actions available to the agent A.
- Set of network units N.
- Weight matrix w of size |N| × |N|.
- Current time t.
- External input x_i(t') to all units i ∈ N for all t' ∈ [t_0, t_1].
- Current activation a_i(t) of all units i ∈ N at time t.
- The goal to achieve g.
- Additional external input γplanning used by the inverse model.
- Additional external input γrepresentation to add to features due to the representation of associated symbols.
- Parameters D and E for activation propagation in the recurrent network.
- Threshold Ω applied by threshold units.

begin

for α ∈ A do

if ACHIEVES(KB, α, g) then

I ← I ∪ {⟨α, AGENTS(g)⟩}

if α ∈ A^G then

G_{new} ← {g ∈ SUBGOALS(α)|g \notin G ∧ (g ∈ G^A → KB \models g)}

G ← G ∪ G_{new}

for g ∈ G_{new} do

REPRESENTGOAL(g, N, w, x(t), a(t), D, E, γrepresentation)

return

L ← LITERAL(g) ∪ {¬ϕ|ϕ ∈ LITERAL(g)}

L ← {ϕ ∈ L(KB \models ¬ϕ) ∪ {ϕ} | ¬ϕ}

for ϕ ∈ L do

F ← LITERALTOFEATURES(ϕ)

if F ≠ \emptyset then

F', a(t), x(·) ← INVERSEMODEL(N, w, t, x(·), a(t), F, γplanning, D, E, Ω)

if F' ≠ \emptyset then

α ← ⟨ϕ, F'⟩; A ← A ∪ {α}

e ← ⟨α, AGENTS(g)⟩; I ← I ∪ {e}

REPRESENTINTENTION(e, N, w, x(t), a(t), D, E, γrepresentation)

return

ory. The propositional variables considered for the knowledge base are Σ = {Red-Stimulus, Green-Stimulus, Left-Stimulus, Right-Stimulus, PushedLeft, PushedRight}. We set initially:

KB = {¬Red-Stimulus, ¬Green-Stimulus, ¬Left-Stimulus, ¬Right-Stimulus, ¬PushedLeft, ¬PushedRight}
8.2. Evaluation

Note that the high-level action representations *push-left* and *push-right* are not available to the intentional level a priori. In Figure 8.1, we include them only for clarity.

We assume the set of stimulus and action effect features \( \mathcal{U} = \{ d, n, l, r, s, u_1, u_2, v_1, v_2 \} \). Unit \( d \) represents red color, unit \( n \) green color, unit \( l \) left position, unit \( r \) right position. Unit \( s \) represents a stimulus and together with any of the units \( d, n, l, r, s \) represents a stimulus attribute. For example, the set \( \{ s, d \} \) represents a red stimulus. The set \( \{ l, u_1, u_2 \} \) represents the effect of perceiving the left button being pushed. The set \( \{ r, v_1, v_2 \} \) represents the effect of perceiving the right button being pushed. Note the overlap between the representations of stimuli and action effects because units \( l \) and \( r \) are both part of a set of features representing action effects and a set of features representing a stimulus attribute. Via threshold units \( \mathcal{T} = \{ q_1, q_2 \} \), both action effect feature sets forward activation to features in the recurrent network spanned by the units in \( \mathcal{V} = \{ x_1, x_2, x_3, y_1, y_2, y_3, z_1, \ldots, z_m \} \). The additional units \( z_i \) represent that there is an entire eco-system of motor command units involved in the process determining which motor command is to be activated. We assume the following fixed link weights:

\[
\begin{align*}
    w_{ij} &= \begin{cases} 
        1 & \text{if } i = q_1 \land j \in \{ l, u_1, u_2 \} \\
        1 & \text{if } i = q_2 \land j \in \{ r, v_1, v_2 \} \\
        1 & \text{if } i \in \{ x_1, x_2, x_3 \} \land j = q_1 \\
        1 & \text{if } i \in \{ y_1, y_2, y_3 \} \land j = q_2 \\
        0 & \text{otherwise.}
    \end{cases}
\end{align*}
\]
Chapter 8. An Implementation of the Joint Action Architecture

Algorithm 8.7: PriorLearning($\mathcal{N}, w, m, D, E, \eta, k$)

**Data:**
- Set of network units $\mathcal{N}$ with motor command units $\mathcal{V} \subset \mathcal{N}$.
- Weight matrix $w$ of size $|\mathcal{N}| \times |\mathcal{N}|$.
- Number $m$ of units $z_1, \ldots, z_m$ in $\mathcal{V}$.
- Parameters $D$ and $E$ for activation propagation in the recurrent network.
- Learning rate $\eta$.
- Number of iterations $k$ for which to learn the patterns.

```
1 begin
2     $x_i(0) \leftarrow 1$ for all $i \in \{x_1, x_2, x_3, z_1, \ldots, z_m/2\}$
3     $x_i(0) \leftarrow -1$ for all $i \in (\mathcal{V} \setminus \{x_1, x_2, x_3, z_1, \ldots, z_m/2\})$
4     $x_i(0) \leftarrow 0$ for all $i \in \mathcal{N} \setminus \mathcal{V}$
5     $x'_i(0) \leftarrow 1$ for all $i \in \{y_1, y_2, y_3, z_m/2+1, \ldots, z_m\}$
6     $x'_i(0) \leftarrow -1$ for all $i \in (\mathcal{V} \setminus \{y_1, y_2, y_3, z_m/2+1, \ldots, z_m\})$
7     $x'_i(0) \leftarrow 0$ for all $i \in \mathcal{N} \setminus \mathcal{V}$
8     for $r = 1$ to $k$ do
9         $a_i(0) \leftarrow 0$ for all $i \in \mathcal{N}$
10        $a(0), s(0), t, x(0), a(0), D, E, \varepsilon \leftarrow$ UpdateRecurrent($\mathcal{N}, \mathcal{V}, w, t, x(0), a(0), D, E, \varepsilon$)
11        LearnRecurrent($\mathcal{N}, \mathcal{V}, w, \eta, a(0), x(0), s(0)$)
12        $a_i(0) \leftarrow 0$ for all $i \in \mathcal{N}$
13        $a(0), s'(0), t, x'(0), a(0), D, E, \varepsilon \leftarrow$ UpdateRecurrent($\mathcal{N}, \mathcal{V}, w, t, x'(0), a(0), D, E, \varepsilon$)
14        LearnRecurrent($\mathcal{N}, \mathcal{V}, w, \eta, a'(0), x'(0), s'(0)$)

end
```

That is, initially only the links from action effect units to threshold units and the links from threshold units to motor command units as sketched in Figure 8.1 have weight 1.

Subsequently, the weights of the links between units in $\mathcal{V}$ are learnt as shown in Algorithm 8.7. We first set all activations to 0 and then create the following two patterns: The first pattern $x_i(0)$ provides an input of 1 to the units $x_j$ ($j = 1, 2, 3$) and half of the $m$ additional units $z_j$ ($j = 1, \ldots, m$) in $\mathcal{V}$, and an input of −1 to all other units in $\mathcal{V}$ (line 2 to 4). The second pattern provides an input of 1 to the units $y_j$ ($j = 1, 2, 3$) and half of the $m$ additional units $z_j$ ($j = m/2 + 1, \ldots, m$) in $\mathcal{V}$, and an input of −1 to all other units in $\mathcal{V}$ (line 5 to 7). We fix $m = 12$ in the following. The first pattern represents the motor command of pushing the left button, the second pattern represents the motor command of pushing the right button. Each motor command is negatively associated with the other one. Both patterns are subsequently learnt for $k$ times (line 8) by the network (lines 9 to 14). This prior learning creates strong associations between the features of the same motor command. Input to part of a motor command causes the network to propagate activation to the other units of that motor command, in effect completing the previously learnt pattern. Providing input to both motor commands leads to the respective feature sets competing for activation so that the activation pattern of the units in $\mathcal{V}$ requires some time to settle to a stable state.
As discussed in the previous chapter, we make the following assumption regarding the activation propagation from action effect features in $U$ to motor command features in $V$: Only together, attitude representation and stimulus perception generate enough activation on the features representing $\text{Pushed-Left}$ ($\text{Pushed-Right}$) so that this activation is propagated to the features representing $\text{push-left}$ ($\text{push-right}$). This is implemented by the appropriate threshold $\Omega$ of the threshold units $q_1$ and $q_2$. For example, representing a goal which refers to $\text{Pushed-Left}$ is not sufficient for activation to be propagated to the features of $\text{push-left}$. Likewise, activating the $\text{left}$ feature, for example, is not sufficient to activate the features of $\text{push-left}$. We do assume that activation due to action planning and execution is sufficient to overcome this threshold. We shall see the ramifications of these assumptions shortly.

### 8.2.2 Sample Trial

Recall that the following conditions of the Simon task can be distinguished: There is an individual and a social task. In the individual task, there is a no-conflict and an action-conflict condition. In the social task there is additionally a task-conflict and a both-conflicts condition. All of these conditions can be manipulated so that only one part of the task is actually represented by the participant (no corepresentation) or both parts of the task are represented (corepresentation). Corepresentation is the default condition. In the individual Simon task, corepresentation can be prevented by not instructing the participant to carry out one part of the task. That part of the task is therefore not corepresented. In the social Simon task, corepresentation can be prevented in the same way and thereby yielding the same condition as in the individual task. Alternatively, corepresentation in the social task can be prevented by recruiting higher-level mechanisms such as the social relationship between the coactors as done by Hommel et al. (2009). In the social Simon task we do not distinguish both types of manipulation and call them both conditions “without corepresentation”. This is because both have the same result: the absence of corepresentation or at least the discarding of the representation relating to the other half of the task.

We consider the action-conflict condition with corepresentation in the social Simon task as our example. To setup the agent for this condition, a top goal $g_1 \in \mathcal{G}^p$ with $g_1 = \langle \text{SimonTask}, H, \text{False} \rangle$ is provided where $\text{SimonTask}$ is a complex action and $H = \{\text{you}, \text{me}\}$ is the set of agents responsible for the task. The agent is equipped with the complex action $\text{SimonTask} \in \mathcal{A}^C$, which can be performed to achieve the top goal. We have $\text{SimonTask} = \langle \text{True}, \{g_2, g_3\} \rangle$ with $g_2, g_3 \in \mathcal{G}^M$:

\[
g_2 = \langle \text{Green-Stimulus} \leftrightarrow \text{Pushed-Left}, \{\text{you}\}, \text{False} \rangle \quad (8.6)
\]

\[
g_3 = \langle \text{Red-Stimulus} \leftrightarrow \text{Pushed-Right}, \{\text{me}\}, \text{False} \rangle. \quad (8.7)
\]
Chapter 8. An Implementation of the Joint Action Architecture

The first subgoal means that if and only if the stimulus is green, the left button is to be pushed by the other agent. The second subgoal means that if and only if the stimulus is red, the right button is to be pushed by this agent. Hence, we have initially $G = \{g_1\}$ and $A = \{\text{SimonTask}\}$. An experimental trial in the action-conflict condition with this setup proceeds as follows:

1. The agent selects the top goal $g_1$ so that $\text{SELECTED}(g_1) = \text{True}$. Using means-end reasoning, the agent creates an intention $\langle \text{SimonTask}, \{\text{you}, \text{me}\} \rangle$ to perform the $\text{SimonTask}$ action with the other participant. This leads to the creation and selection of the above mentioned subgoals. We assume corepresentation here, so the subgoal referring to the other participant is indeed represented. Consequently, we have $G = \{g_1, g_2, g_3\}$.

2. Due to the representation of subgoals, input is added to the corresponding features of symbols (i.e. $\text{Green-Stimulus}$, $\text{Red-Stimulus}$, $\text{Pushed-Left}$, $\text{Pushed-Right}$) according to the $\text{REPRESENTGOAL}$ operator. This is what corresponds to action and task corepresentation considered by Sebanz et al. (2005). Both the other participant’s task (encoded in the subgoal) and action (via action effects) is corepresented. However, no activation is propagated to motor command features yet because threshold units do not receive sufficient input to be activated.

3. After this initial setup, a red stimulus is presented on the left side. This triggers the $\text{PERCEIVESTIMULUS}$ operator during the $\text{GETEXTERNALEVENTS}$ stage in the interpreter loop (Algorithm 8.5). This causes additional input on the features $s$, $l$, and $d$. Now the feature set representing $\text{Pushed-Left}$ has sufficient input to have activation leak over into the features representing the $\text{push-left}$ action. Also, corresponding propositions ($\text{Red-Stimulus}$ and $\text{Left-Stimulus}$) are then made true in symbolic memory.

4. Now, the subgoal $g_3$ needs to be maintained because the proposition $\text{Red-Stimulus}$ is true and $\text{Pushed-Right}$ is false.

5. Means-end reasoning fails to retrieve an existing action that would maintain the subgoal $g_3$. Therefore, means-end reasoning employs the inverse model to obtain an action $\langle \text{Pushed-Right}, F \rangle$ referring to motor command units in the set $F \subset V$ that can achieve $\text{Pushed-Right}$. We call this action $\text{push-right}$ but emphasize that this action was not explicitly available to the intentional level before and its name is arbitrarily

\footnote{Note that a more expressive language for the representations in symbolic memory would allow a more elegant description.}
8.2. Evaluation

selected. When using the inverse model, additional input is provided to the feature sets representing Pushed-Right and push-right so that their activation is increased.

6. An intention \langle push-right, \{me\} \rangle is created to execute push-right, which invokes the REPRESENTINTENTION operator that further adds to the input of the features representing that action.

7. Because the intention \langle push-right, \{me\} \rangle refers to this agent and the action push-right is a primitive action, the agent invokes the EXECUTE operator to perform that action. This causes additional input to be provided to the features representing push-right. After the recurrent motor command network has settled into a stable activation pattern, the motor command is performed.

8. Perceptual input of pushing the button then triggers the PERCEIVESIMULUS operator. This causes another increase in the input to the features representing Pushed-Right, so that this proposition is made true and the subgoal of this agent is deemed achieved.

8.2.3 Parameter Estimation

The following has been observed empirically:

- There is no Simon effect of any kind without corepresentation (Hommel et al., 2009; Sebanz et al., 2003).\(^3\)

- There is a Simon effect in the action-conflict condition with corepresentation and the effect is of similar size in the individual and the social task (Sebanz et al., 2003).

- The Simon effect in the task-conflict condition is larger than the one in the action-conflict condition (Sebanz et al., 2005).

- The Simon effect in the both-conflicts condition is larger than the sum of the Simon effects in the task-conflict and the action-conflict condition (Sebanz et al., 2005).

To validate our model against these qualitative observations, we take the time required by activation to settle in step 7 in the procedure above as an estimate of the response time and hence of the Simon effect. This in line with Sebanz et al. (2005) who suggest that the Simon effect is due to an action selection conflict after actions have been activated. Any input on the action not to be executed (push left for the right-hand side participant) increases this settling time. In the action-conflict condition, the left-side stimulus adds onto the input of the feature

\(^3\)Note that Sebanz et al. (2005) do report a small effect in this condition in the reproduction of their previous experiments (Sebanz et al., 2003). However, they hypothesize that this might be due to carryover effects from previous trials or due to another type of compatibility effect.
Chapter 8. An Implementation of the Joint Action Architecture

$l$, which is shared with the features representing Pushed-Left and hence provides further input to the features of push-left. In the task-conflict condition, means-end reasoning for the other participant’s task via the inverse model adds onto the input of the features representing the push-left motor command. Representing different experimental conditions is a matter of providing different goals and a different stimulus (see above and Section 7.2).

While we call this section “Parameter Estimation”, we do not perform any sophisticated form of parameter estimation. The reason for this is that different parameters for the model cannot be compared quantitatively with respect to the match they yield with the qualitative empirical observations described above. A parameter either yields a model that reproduces the relationships between the Simon effect in different conditions or not. Therefore, our goal is only to find any parameter for which these relationships are reflected by the different conditions of the Simon task in our model.

The parameter space is high-dimensional and continuous and therefore an exhaustive enumeration of all possible parameters is impossible. We opt for generating parameters randomly from the parameter space and test whether these generated parameters lead to a model that is a fit with regard to the qualitative observations described above. We note that some parameters can be discarded immediately because they cause the model to disobey the rules of activation propagation from action effect units in $U$ to motor command units $V$ as identified above. Let us call those parameters invalid and all other parameters valid. We have generated enough parameters to collect 1000 valid parameters and have found 9 of these to reproduce the desired effect. Table 8.1 shows the individual entries of one such parameter (this one was crafted manually though). Note that here $D$ and $E$ were set in accordance with previous work (Rumelhart et al., 1986; Smith and DeCoster, 1998). The learning rate $\eta$ was chosen to be relatively small, which is in line with typical setups of connectionist models (see discussion in Chapter 5). A more rigorous parameter estimation could be conducted with the technique used in Chapter 5, for example.

We conclude that this simulation model is able to reproduce qualitatively the empirical observations obtained from the different conditions of the Simon task. Hence, this model and the architecture it is based on are promising objects for further investigations into computational models of human joint action.

8.3 Related Work

In this section, we discuss how the joint action architecture proposed in the last chapter and its concrete implementation presented in this chapter relate to other formal models of joint action. Recall the main requirements we identified in Section 7.3:

- The encoding of stimuli and actions in terms of a shared set of features.
### 8.3. Related Work

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{\text{planning}}$</td>
<td>0.4</td>
</tr>
<tr>
<td>$\gamma_{\text{execution}}$</td>
<td>2.0</td>
</tr>
<tr>
<td>$\gamma_{\text{stimulus}}$</td>
<td>0.3</td>
</tr>
<tr>
<td>$\gamma_{\text{representation}}$</td>
<td>0.2</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>0.2</td>
</tr>
<tr>
<td>$D$</td>
<td>0.15</td>
</tr>
<tr>
<td>$E$</td>
<td>0.15</td>
</tr>
<tr>
<td>$\eta$</td>
<td>0.1</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 8.1: A sample parameter setting for which the model reproduces the empirical observations on the Simon effect in different conditions of the Simon task qualitatively.

- An associative memory that hosts these features and their relationships.
- Inverse and forward models for action planning implemented on top of associative memory.
- Practical reasoning based on symbolic representations of mental attitudes such as beliefs, goals, and (joint) intentions.
- A symbolic memory that holds these mental attitudes.
- The ability to describe joint tasks at the symbolic level.
- An interaction between associative and symbolic memory and the conditioning of associative memory by the symbolic representation of the joint task.

There is a vast array of formal models of human behavior. We discuss that work which comes closest to fulfilling these requirements and which features most prominently in the literature.

#### 8.3.1 MOSAIC

The Modular Selection and Identification for Control model (MOSAIC, Wolpert and Kawato, 1998) is a computational model of motor control. What makes MOSAIC interesting for our discussion is that it relies on forward and inverse models as our model does and that it has been considered as a model of social interactions.

The constituents of MOSAIC are modules each of which consists of a forward model, an inverse model, and a responsibility predictor. The current motor command is a mixture of the motor commands produced by all inverse models. The overall motor command is fed back into all forward models, which provide predictions of the next state based on this
motor command. The contribution of each inverse model to the overall motor command is weighted by the prediction accuracy of its associated forward model and the output of its responsibility predictor. A responsibility predictor estimates the responsibility (weight) for its associated forward and inverse model from context information only. The system is given a Bayesian interpretation with the likelihood of each module being determined by the predictive accuracy of its forward model. The estimates produced by all responsibility predictors provide a prior probability distribution over all modules. The posterior distribution is determined by the resulting prediction accuracy of all forward models.

Forward models, inverse models, and responsibility predictors are function approximators and their parameters are learnt gradually with the experience of the agent. The learning rate for each forward and inverse model is proportional to the prediction accuracy of that forward model. Therefore, different forward models become “experts” for different situations and each inverse model acquires the capability to provide good motor commands in situations in which the associated forward model makes good predictions. Responsibility predictors are updated based on the difference between predicted responsibility (prior) and the resulting prediction accuracy of associated forward models (posterior).

MOSAIC has been implemented and its ability to perform in motor control tasks with changing conditions has been shown by simulation (Haruno et al., 2001). Two extensions of MOSAIC are particularly interesting from our perspective: considering other agents as part of the control systems, and considering the hierarchical organization of modules (Wolpert et al., 2003).

The first extension considers an agent to exert control over another agent’s hidden internal state by acting on or towards this other agent (e.g. by signaling or communicating). A change of the other agent’s internal state manifests itself in that agent’s reaction, which is perceived as sensory consequences by the acting agent. This essentially takes the other agent as part of the system being controlled in addition to the motor apparatus. Hence, internal models can be learnt to “control” another agent through actions of the motor apparatus, similar to learning to control one’s own motor apparatus. Wolpert et al. admit that controlling the state of another agent is disproportionately more difficult than controlling the own motor apparatus because the degrees of freedom of the controlled system is much larger. They suggest that the solution to this problem is the similarity of brains across people, or in their model the similarity of MOSAICs across agents.

If an actor and an observer share similar MOSAICs, the observer can efficiently map any observed behavior onto responsibilities of its own modules that would reproduce the observed behavior. This allows an observer to make use of its own motor control system to make sense of the other agent’s behavior. Consider an agent that observes the behavior of another agent. Each Inverse model of the observer’s MOSAIC determines a motor command that it deems
to reproduce the observed behavior of the actor. The motor command produced by each inverse model is used as an input to the associated forward model. Based on the received motor command, each forward model predicts the subsequent state of the actor. Comparing prediction to the actually observed state yields the prediction accuracy for each module. Prediction accuracies determine the responsibility of each module over the time frame of the observed behavior. In essence, this yields a symbolic stream of module responsibilities that represents the observed behavior. This stream can be stored more efficiently than a representation of the entire observed behavior. This symbolic representation can also be reused later, for example for imitation by invoking modules with the responsibilities encoded in the extracted stream. Obviously, the efficiency of this mechanism depends on the similarity between the MOSAICs of actor and observer.

The second extension considers a hierarchical organization of MOSAIC modules. Higher-level modules learn to orchestrate lower level ones given the current situation and any desired behavior. Higher-level modules produce prior probabilities for the selection of lower-level modules and in return receive responsibilities (posterior probabilities) as input from these lower-level modules. Higher levels learn increasingly abstract control behaviors with those at the top of the hierarchy representing high-level goals. When observing the behavior of another agent, responsibilities for modules at all levels of the hierarchy can be determined. This allows the extraction of complex behaviors including both low-level behavior as well as high-level goals.

Wolpert and Kawato (1998) consider that the function approximators representing inverse models, forward models, and responsibility predictors can be implemented by neural networks (i.e. a connectionist architecture). In that, MOSAIC can fulfill our requirement that actions be encoded in terms of features within an associative memory that enables inverse and forward models. Wolpert et al. (2003) discuss the representation of higher-level goals in a hierarchy of modules. Yet this does not allow the higher-level symbolic reasoning that we have identified as a necessary component of a complete architecture of joint action. In particular, it is not clear how for example the task description of the Simon task would be represented in MOSAIC. From this follows that there is no symbolic memory other than a storage for responsibility streams. Wolpert et al. do not elaborate on this storage capability.

Overall, MOSAIC is a more concrete and complete description of motor control than offered by our architecture. In particular, we have not elaborated on the factor of time in our description. In our simulation model actions are instantaneous. Also, concrete learning algorithms have been specified for MOSAIC while we have not discussed learning in any detail. However, our open specification appears to allow MOSAIC to be adopted as the implementation of our non-intentional level. As required by us, MOSAIC enables action simulation to predict another agent’s subsequent state from an observation of its behavior.
Yet forward models in MOSAIC are not assumed to provide additional information to inverse models directly. MOSAIC enables actions to be hierarchically organized at the non-intentional level. We have only considered a hierarchical organization of behavior at the intentional level. However, we concede that the ability to organize behavior hierarchically at a lower level is likely to be required to avoid unnecessarily expensive intentional reasoning.

### 8.3.2 SCM

The Shared Circuits Model (SCM, Hurley, 2008) is a description of motor control at an intermediary level between neural-level mechanisms and mechanisms of higher-level reasoning. One of the goals of SCM, however, is to predict certain neural-level mechanisms and to explain how its mechanisms enable higher-level ones, in particular those for joint action such as imitation and mind-reading. In that, SCM is cast at a similar level to MOSAIC but does not go into the detail of neural implementations to the extent that MOSAIC does. SCM is relevant to our discussion because our motor control mechanism at the non-intentional level borrows from SCM and because Hurley discusses how SCM gives rise to mechanisms of joint action.

Hurley describes SCM in terms of five layers that gradually build upon each other. While we adopt the same approach to the description of SCM here, we emphasize that the appropriate description of a layer is to consider it as an extension of previous layers rather than an autonomous entity.

The first layer consists of a basic control system with an inverse model (called comparator) that receives a target state and external input and produces a motor command as output. The inverse model compares the current input to the target and adjusts the motor command accordingly. The motor command modifies the motor apparatus, which causes an update to the perceptual feedback provided as input to the inverse model. At this level, no distinction is made between input that is due to the agent’s actions (reafferent) and input that is due to external causes (exafferent).

The second layer adds a forward model that receives a copy of the motor command and feeds this back to the input signal. As discussed earlier, a forward model can provide simulated feedback that is more readily available than real feedback from the controlled system. An effective use of the forward model requires that the system be able to switch between forwarding simulated and real feedback to the inverse model. By comparing simulated with real feedback, the system is now able to distinguish reafferent from exafferent input. This could enable the higher-level distinction between actions of the self and environmental effects. However, there is not yet any possibility to distinguish actions of the self from actions of others. Hurley suggests that the forward model would give rise to associations between object affordances and motor commands as implemented by canonical neurons.
8.3. Related Work

Canonical neurons relate to mirror neurons in that they have been found to be active during the execution of an action as well as the perception of objects that afford this action (e.g. Jeannerod et al., 1995; Murata et al., 1997).

The third layer adds the ability to exercise the forward model in the other direction, that is, from effect to cause instead of cause to effect. Hurley calls this process “mirroring”. When the effects of a motor command appear as input, mirroring ensures that this motor command is activated. Mirroring entails the activation of motor commands through observing the actions of another agent. Such activations can lead to overt copying of the actor’s behavior. At this level, mirroring does not distinguish actions of the self from actions of other agents. Recall that in our model the inverse model implements mirroring. In SCM, the inverse model is separate from mirroring but Hurley notes that both might share the mirror neuron system as part of their neural-level implementation.

The fourth layer introduces the possibility of output inhibition and the monitoring of inhibition. This capability allows the system to use the forward model of the second layer offline in order to predict the outcome of different possible actions rather than providing predictions for actual actions. Another consequence is the distinction between one’s own actions and other’s actions activated due to mirroring: The latter are inhibited, the former are not. This enables action understanding because the system extracts instrumental representations (motor commands) from observed actions and can distinguish these from its own actions.

The fifth and last layer adds the possibility to specify counterfactual input offline and to monitor whether the input currently received is offline or not. This allows for simulating others’ possible actions, which provides information about the causes and further effects of these possible actions. Together with the effects of own possible actions, this can give rise to basic forms of strategic deliberation.

Hurley indicates that multiple instances of SCM could operate in sequence, hierarchies, or networks to implement complex behaviors in a way similar to MOSAIC. For example, in a sequence of SCM modules one module’s target would be another module’s means.

Both MOSAIC and SCM explain how higher-level mechanisms of motor control, including some characteristic to joint action, can emerge from lower-level mechanisms. However, while MOSAIC is cast as a formal model close to the neural level with the intention to implement and test the model, SCM is described conceptually at a slightly higher level. MOSAIC is mainly applied to tease out precisely how the motor control system can be efficient in complex (social) environments. SCM, in contrast, is described at a higher level than the neural one. SCM is used to explain in general how prerequisites for higher-level mechanisms of mind-reading such as self/other and actual/possible distinction come about and which general neural level mechanisms are required by SCM.
Our architecture draws on a model of motor control that is close in its general setup to SCM. In fact, the arrangement of inverse and forward model in our architecture is inherited from SCM, in particular considering the forward model to provide simulated feedback to the inverse model. A difference is that we consider the inverse model to be responsible for mirroring and planning, while Hurley distinguishes these two functions. However, she states that both functions might share part of their neural-level implementation, i.e. the mirror neuron system. We also consider output inhibition as an important part of our architecture but we do not discuss taking input offline. It appears therefore that at a conceptual level, our non-intentional level is generally compatible with SCM, which allows our non-intentional level to share part of Hurley’s analysis. Furthermore, SCM suggests a neural-level associative memory, which at a higher level can be perceived as an associative memory over features like those we presume.

SCM is a model of motor control and thus does not consider any higher-level reasoning. Hurley suggests that full-scale mind-reading and tracking of multiple agents in joint action likely requires language-based capabilities of reasoning on top. However, these are not explicated. Also, it does not become clear from Hurley’s discussion which mechanism controls output inhibition. In our previous discussion we have taken this to be part of the responsibility of the intentional level.

8.3.3 HiTEC

HiTEC (Haazebroek et al., 2011) is a cognitive architecture of the interplay between perception and action based on the theory of event coding (TEC, Hommel et al., 2001), which we mentioned briefly in the previous chapter. HiTEC is interesting from our perspective because it presumes that stimuli and action effects are represented by a shared set of features as we do and because it has been used to reproduce the Simon task.

We first describe TEC in more detail than in the previous chapter and then discuss HiTEC. According to Hommel et al. (2001), TEC is a theory of the coupling between perception and action that relies on mechanisms of “late perception” and “early action”. TEC assumes that stimuli as well as action effects are encoded in a shared set of distal features, which code for abstractions from proximal features that underlie “earlier perception” and “later action”. Perception as well as action planning cause the activation of event codes, which are compositions of features that represent events (stimuli or action effects). Because stimuli and action effects share the same representational domain, perception and action planning can interfere with each other. So far, this is in line with our model of associative memory.

TEC departs from our model of associative memory in the following ways. First, TEC does not include any features that code for motor commands. Only the planning of actions in terms of their effects is considered but not the translation of this representation into a
motor command. Second, therefore TEC cannot be a complete model of motor control in the sense of MOSAIC or SCM and it does not claim to be. Third, representing an event does not only consist of the activation of its respective features but also in their temporary integration. This temporary integration can be implemented, for example, by the temporal synchronization of the activation of participating features. TEC is agnostic about the particular mechanism providing for feature integration. Integration of features to a particular event makes them unavailable for the representation of other events. This is the reason for interference between mechanisms working on different events whose features overlap. When a feature is preactivated during task preparation, it has a competitive advantage over other features in the competition for feature integration after stimuli perception. Hence, an agent’s attention is biased towards task-relevant information, an effect that also is afforded by our architecture and model.

HiTEC is based on a partly-connected recurrent network with similar mechanisms to the ones we employed in this chapter. Units are partitioned into 3 pools (so called levels): the sensory-motor level, the feature level, and the task level. In general, connections between levels are bidirectional and exhibitory. At the sensory-motor level, units represent proximal attributes of sensory perceptions or motor codes. At the feature level, units represent distal features that are compositions of proximal attributes. A feature unit can be connected both with sensory units and with motor units at the sensory-motor level. At the task level, units represent task rules. A task rule connects both with features representing stimulus attributes and features representing responses, i.e. features that are connected with motor units. By integrating stimulus and response features, a task rule corresponds to an event in the sense introduced above. In each level, units are arranged into so called maps. The units of the same map represent incompatible entities, e.g. left and right location at the feature level. Links between units in the same map are inhibitory.

Activation flows between units from the same and different levels and units compete for activation. When an action is to be planned, the features representing the desired effects are activated and activation is allowed to flow through the network and to motor units. Once any motor unit’s activation exceeds a certain threshold, that motor unit is activated and the motor command executed. The number of cycles that activation needs to be propagated through the network for a motor command to be executed is considered as a measure for response time. A stimulus can activate task rules via features that code for stimulus attributes. Activated task rules then propagate activation to features that code for responses, which then leads to an activation of motor units and the execution of a motor command. The network of units is specified by hand including most of the connections between units. Only the associations between features and motor units are learnt.
When the individual Simon task is represented, task-irrelevant stimulus attributes can lead to activation being propagated to incorrect motor units via feature units. This additional activation to the incorrect motor unit slows down the process of settling to the motor unit that represents the correct action. This amounts to a Simon effect. Hence, this model is able to reproduce the results of the no-conflict and action-conflict conditions in the individual Simon task with corepresentation.

HiTEC does not account for any mechanism that controls the creation of task rules and their associations with features. So far these entities are set up manually. Therefore, the model cannot account for any higher-level mechanisms that control, for example, whether corepresentation affords or not. Also, the model does not provide any mechanisms for the planning for another agent, as observed in the social Simon task. In fact, the representation of a part of the task is determined by the decision to set up a corresponding task rule or not. However, the association between features and motor codes is independent from this decision. Therefore, even without corepresentation, a task-irrelevant stimulus attribute can add to the activation of an incorrect action. Hence, HiTEC also predicts a Simon effect in the action-conflict condition without corepresentation. As we have described earlier in this chapter, this is not what has been observed.

In contrast to TEC, HiTEC considers “early perception” and “late action” in the sensory-motor level. This is necessary to represent how perceptual input adds to the activation of features and how the activation of features leads to the activation and execution of motor commands. However, HiTEC stays faithful to TEC with respect to its omission of more complex mechanisms of motor control as considered by MOSAIC and SCM. Motor commands in HiTEC are not activated because they are explicitly planned for. Motor commands are activated when they are referenced by activated task rules or by features that overlap with perceptual input. The lack of higher-level reasoning allows only for simple task descriptions in terms of stimulus-response rules, which contrasts with the representational power of our intentional level. HiTEC considers the integration of features into events only in terms of task rules, which, however, are specified manually. Also, associations between features and the events (task rules) they constitute are permanent and not temporary. Competition does not only occur between events (task rules) but also between contradictory/incompatible features or sensory or motor codes. This competition is the cause of the Simon effect in HiTEC. In our model, in contrast, competition only occurs after motor commands have been activated, representing the action selection conflict hypothesized by Sebanz et al. (2005).

Overall, what our model mainly shares with HiTEC and TEC is the common coding of stimuli and action effects in terms of a shared set of features. HiTEC has been implemented to reproduce empirical work, which provides evidence in favor of TEC. In fact, there is no
way TEC can be validated—a problem that our architecture faces as well—because it is not precisely specified. However, evidence in favor of these conceptual models can be obtained by creating and testing concrete instantiations.

8.3.4 DiPRA

The Distributed Practical Reasoning Architecture (DiPRA, Pezzulo, 2009) is a two-level agent architecture in which a higher-level practical reasoning module interacts with a lower-level motor control module. DiPRA is primarily intended to be a robot architecture. However, since DiPRA draws on models and theories from psychology and philosophy in a similar way we do, it is relevant to our discussion here. In particular, the two levels of DiPRA correspond closely to our intentional and non-intentional level.

Following Bratman et al. (1988), the higher, intentional level of DiPRA performs practical reasoning based on beliefs, goals, intentions, plans, and actions. The relations between beliefs, goals, and plans are encoded by a fuzzy cognitive map. Each such mental attitude has an activity level that is forwarded to all the other attitudes it has a relationship with. For example, a belief that is part of a goal has a relationship with this goal and therefore contributes to that goal’s activity level. Since the propagation of activity levels proceeds in parallel, multiple possibly competing goals and plans can be active at the same time. Therefore, deliberation about which goal to pursue and means-end reasoning about which plan to adopt is inherently interleaved. At any point in time, the goal with the highest activity level is the currently pursued one and the plan with the highest activity level that achieves this goal is the current intention.

Actions at the intentional level correspond to schemas at the lower, sensorimotor level. A schema is basically the representation of a behavior. Each schema consists of an inverse model and a forward model implemented on top of a recurrent connectionist network. Perceptual schemas control the sensors of the agent. Motor schemas control the actuators of the agent and receive perceptual input from perceptual schemas. As mental attitudes at the intentional level have an activity level, so do schemas. An intentionally scheduled plan can evoke actions during its execution and thereby cause the execution of corresponding schemas, which adds to their activity level. The activity level of a schema also increases if it is relevant to the management of primitive drives such as hunger or fear. Additionally, the activity level of a schema increases with increasing prediction accuracy of its forward model. Hence, schemas are more active the more urgent their execution is and the more suitable they are for the current situation. The output of all schemas is merged to determine the current motor command. However, each schema contributes to this motor command only in proportion to its current activity level.
Beliefs are grounded in schemas in that they constantly monitor the activity of relevant schemas and set their value based on the activity of these schemas. Pezzulo (2009) provides the example of a belief “the door is open”. The value of this belief can be inferred from a high activity level of a motor schema for passing through doors. However, beliefs can also determine their value by simulating relevant schemas. In the example above, the agent can infer the value of the belief “the door is open” by simulating the schema for passing through the door and determining the success of that simulation. Simulation takes input and output of a schema offline and uses its inverse and forward model to do prediction. Simulation allows a belief to be determined without actually executing any motor command.

Computational resources in DiPRA are bounded and all entities (goals, plans, beliefs, and schemas) compete for these resources. Entities with low activity levels receive less resources, which means for example in the case of a belief that this belief can check its current value less often. Effectively, the activity of an entity represents the relevance of this entity to the current context. Because entities receive access to computational resources in proportion to their activity level, the agent’s reasoning and acting is inherently focused on relevant information and behavior. This represents a kind of attention mechanism. However, despite attention and the commitment to a particular course of action, the agent stays reactive to any development that might require immediate action. In such a case, relevant schemas will increase in their activity level even if they were not planned for intentionally.

The overall structure of DiPRA is reminiscent of our architecture. However, DiPRA goes further in specifying in more detail the processes happening at and between the different levels. In that, the intention behind DiPRA is a slightly different one. In our architecture description, we sought to outline the basic mechanisms required to reproduce empirical results on human behavior in joint action. Consequently, our architecture is a framework for different instantiations that interpret its basic components in different ways, yet obeying its generic structure. Our intention was not to provide a model that can be readily implemented. In contrast, this is the goal of DiPRA. And in fact, Pezzulo (2009) is able to demonstrate that this model leads to effective robot behavior more so than models that include only either of its two levels.

Because of its different purpose, DiPRA also does not fulfill all of those requirements we identified in the previous chapter and outlined at the beginning of this section. Let us go through these requirements one by one. A schema does encode stimuli and action effects in an associative memory, which enables an inverse and a forward model. However, action effects and stimuli do not share the same subset of features. Also, actions or schemas are not explicitly controlled in terms of their effects. The inverse model of a schema is learnt a priori to associate certain behavior with a particular context or situation. During execution, the schema’s inverse model receives the current context as input and produces a motor command.
8.3. Related Work

In that, a schema knows which control to exert in a given context but it does not explicitly know how to control for the achievement of a particular goal state. Also, schemas cannot be activated directly by an overlap between their features and the features of a stimulus. As discussed above, the activity level of a schema increases only due to activation from plans, input from drives, and accurate predictions of the forward model. Therefore, we cannot see how a Simon effect could arise within DiPRA. Likewise, the observation of another actor performing an action would not lead to any schema activation in the agent. Therefore, DiPRA does not directly provide any mechanism for action understanding, imitation, or mimicry.

Obviously, practical reasoning based on symbols is an integral part of DiPRA. However, DiPRA does not consider the description of joint tasks because joint action is not on its agenda. What our architecture does share with DiPRA is the distinction between an upper level that employs practical reasoning and a lower level for motor control, both of which interact with each other.

In comparison to MOSAIC and SCM, DiPRA puts less emphasis on the emergence of higher-level mechanisms from motor control, and does not consider the emergence of higher mechanisms of joint action. As described above, DiPRA's interpretation of an inverse model is also slightly different from those of MOSAIC and SCM. DiPRA does not share much with HiTEC since HiTEC is intended to represent TEC, whose main components—common coding of action effects and stimuli and the planning of actions in terms of their effects—are not accounted for by DiPRA.

8.3.5 Summary

In this section, we have discussed those models that share substantial resemblance with our architecture or are born out of similar intentions. However, in analyzing these models, we have found that none of them fulfills all of the requirements that we have identified in the previous chapter. We conclude that despite the variety of existing models, our architecture contributes a novel perspective to the computational modeling of joint action.

We have not discussed multi-level cognitive architectures such as ACT-R (Anderson, 1993) and CLARION (Sun, 2002). The reason is that these architectures were primarily developed to represent the cognitive mechanisms of individuals, and that their higher level typically does not enable practical reasoning in the sense of Bratman et al. (1988). We have also neglected models of motor control that remain mostly silent about the role of motor control models in joint actions, such as the emulation theory of representation (Grush, 2004) or HAMMER (Demiris and Khadhouri, 2006). However, we believe that the architectures discussed above do not only constitute a representative exposé but to the best of our knowledge cover those approaches that were developed with similar intentions to ours.
8.4 Discussion and Conclusions

This chapter closes our two-part discussion of the joint action architecture. In the last chapter we have described how the architecture is informed by existing empirical and theoretical research. In this chapter we have demonstrated how the architecture gives rise to an implementation that reproduces qualitatively the results of a central experiment in the study of human joint action (the Simon task and its derivatives). We are not aware of any other computational models that can reproduce all the derivatives of the Simon task. Without the context of the computational-conceptual model of the previous chapter, however, it would be unclear how the model presented in this chapter is grounded in empirical and theoretical research. This demonstrates the benefit of auxiliary models as hypothesized in Chapter 3. It also demonstrates the value of using the architecture as input to further models of joint action and thereby indirectly advances the study of cultural dynamics, thus contributing to RQ1.

Obviously, the computational complexity of this model surpasses the complexity of the model introduced in Chapter 4. A comparison with the models introduced in Chapter 5 and 6 depends on further details. Both of those other models can be related to one of the levels of the architecture discussed in this and the last chapter. We leave this task open for future work. At the same time, this model is a rather faithful representation of theoretical and empirical work on the details of joint action. What remains to be discussed are the limitations and shortcomings of our implementation model with respect to the architecture, and some more general limitations. These are the topics of the remainder of this section and chapter.

At the intentional level, we have not included any mechanism that provides for the proper management of mental attitudes. For example, we have not considered the handling of conflicting goals or intentions. We have also relied on a simple representation of symbolic knowledge that does not allow for the representation of temporal knowledge. Our model of joint intentions is a simple one as is our model of actions. The means-end reasoning process is a rather simple one and agents are not able to derive the responsibilities for subtasks from a first principle representation of the capabilities of the coactors. Instead, these are hard-wired into action plans, e.g. the SimonTask action. In our example scenario, the Simon task, these features were not required. However, their implementation would certainly increase the scope of phenomena this model could reproduce.

At the non-intentional level, we also lack a representation of temporal aspects. Actions are instantaneous so that there is no room for any actual motor control in which the inverse model gradually executes an action. Only part of associative memory undergoes a learning process and we remain silent about the learning of other parts. We have seen that HiTEC makes the same simplifications. Feature activation levels in our implementation are not assumed to attenuate over time, a feature that would be necessary to allow associative
8.4. Discussion and Conclusions

memory to encode new and different stimuli. Again, these simplifications do not prevent the model from fulfilling its purpose.

MOSAIC, SCM, and DiPRA account for the concurrent execution of multiple motor control modules and potentially the hierarchical or sequential arrangement of these models. We have not had this point on our agenda. Yet certainly the challenges of motor control can more easily be mastered by a modular architecture.

TEC explains how multiple objects and their stimuli are bound together temporarily. We have not considered the configuration in associative memory to a detail sufficient to discuss such a capability in our architecture. However, it is important for future work to consider how different concurrent stimuli might be distinguished.

A last important avenue for future work is the expansion of the preliminary parameter estimation we have conducted in this chapter and of course a proper validation of this model.
Chapter 9

Conclusions

This thesis has investigated the computational modeling of the transmission of cultural information and of the emergent cultural dynamics. The main background assumption has been that the study of cultural dynamics requires a two-component research program: (1) the study of cultural transmission during micro-level social interactions and its interrelation with macro-level properties of culture and nano-level mental representations and processes, and (2) the expansion and application of these results to the larger scale of societies.

We have argued that computational modeling can add to both parts of this agenda because computational models are well suited to represent the mechanisms underlying cultural transmission and dynamics at all levels of analysis (nano, micro, and macro). In the case of the first component of this agenda, computational modeling contributes in particular by supporting the refinement of models and theories developed in fields such as psychology and cognitive science. In the case of the second component, computational modeling promises to afford the capability to cope with the bi-directional interactions between culture as a macro-level property, the transmission of cultural information during micro-level social interactions, and the enabling nano-level mental representations and processes.

While there are many uses of models, two have dominated this thesis: The implications of computational models can be studied by simulations, and the development of computational models itself adds to the refinement of the models and theories they represent. We have also emphasized and relied on the benefit of two other types of models, which we have termed computational-conceptual and formal logics models. Both types of models can serve as auxiliary models for the translation of informal social-scientific theories and models into concrete computational representations. At the same time they allow precision to be added to the discussion on the theories and models they represent.

Given these observations, this thesis has presented computational models of cultural dynamics based on a particular theory of cultural transmission in micro-level interactions—
Chapter 9. Conclusions

the grounding model of cultural transmission by Kashima et al. (2008). The overarching research questions have been the following ones (see also Section 1.3):

- How can the grounding model of cultural transmission be translated into computational models of cultural transmission and dynamics?

- How can these computational models contribute to the refinement of the grounding model of cultural transmission and to the understanding of cultural dynamics?

This chapter is structured as follows. In Section 9.1, the contributions of this thesis are reviewed. In Section 9.2, limitations are outlined, and in Section 9.3, directions for future work are indicated. Finally, the thesis concludes with a few remarks in Section 9.4.

9.1 Contributions

The contributions of this thesis towards the first research question are a series of computational models of cultural transmission and dynamics that draw on the grounding model of cultural transmission. In particular, we have built on the premises and implications of the grounding model of cultural transmission identified in Section 2.2.3 to construct these models. The analysis of these models presented in the respective chapters is the contribution of this thesis towards the second research question. The specific contributions towards these two research questions are:

Chapter 4 has introduced a novel agent-based model of cultural dynamics as an extension of the well-known and well-studied Axelrod model of the dissemination of culture. The innovative aspect of our model is that it accounts for the co-evolution of cultural, social, and physical space. To the best of our knowledge, this has at least not been done within the context of the Axelrod model. The model draws on the grounding model of cultural transmission—in particular the consequence of successful and unsuccessful grounding on social relationships—to justify the mechanisms that underly the co-evolutionary dynamics.

The analysis of this model supports the view that global cultural homogenization can indeed go along an initial cultural heterogenization local to social communities. While the number of different cultures decreases constantly over time, the cultural diversity within communities experiences an initial increase. The initial local heterogenization is facilitated by a higher probability of interaction over longer distances, a situation representing the recent improvement of transportation and communication technologies. However, ultimately a number of cultures with local homogenization emerge. Global homogenization—represented by a lower number of surviving cultures—is favored by
large population sizes and densities, which reflects the situation of large cities as well as city states with large populations. Our analysis also suggests that migration can be a promoter of cultural segregation and global diversity when migration is not random but directed towards existing social contacts.

Chapter 5 has introduced an agent-based model of the communication of stereotype-relevant information. The novelty of this model lies in three distinguishing features: First, agents have a memory of the information that they and their audience associate with third-parties. Second, agents make use of this information when they communicate information about a third-party to their audience. Third, the storage of information about individuals and social groups is inherently integrated, thereby reflecting the nature of stereotypes as information that is about groups as much as individual members of these groups. The model is mainly built on an empirical study by Lyons and Kashima (2003) that can be considered a first investigation of the predictions made by the grounding model of cultural transmission.

By successfully calibrating and validating the model support is given to the mechanisms that the grounding model of cultural transmission offers as an explanation of the empirical data collected by Lyons and Kashima (2003). In addition, the model is put further to the test when larger agent populations are considered in simulations with agents obtaining stereotype-relevant information through first-hand experience as well as second-hand information exchange. This paves the way for the development of models of the diffusion of stereotype-relevant information that account for the composition and configuration of the underlying social network, i.e. the structure of social links as well as the common ground associated with social relationships, the group memberships of communicating agents, and the group memberships of the third party communicated about.

Chapter 6 has presented a comprehensive computational-conceptual account of the grounding model of cultural transmission, thus contributing to the refinement of that model. The presented model lends itself to serving as an auxiliary model to other efforts at modeling cultural dynamics. At the core of that chapter is a novel formal logics model of common ground that has been developed based on an analysis of relevant philosophical literature and on the postulates of the grounding model of cultural transmission. The model’s properties are analyzed to better understand the consequences of the choices made. In addition, we have described semi-formally how this model of common ground can be integrated into a formal model of joint activities and the grounding process. The interaction between these three components is a crucial element of the grounding model of cultural transmission. By accounting for the specific requirements
of the grounding model of cultural transmission, our model is distinguished from other computational formalizations of dialogue, common ground, and grounding. Therefore, the particular contribution we make is the adoption, integration, and application of different formalisms to describe the particular features of the grounding model. Our common ground formalization, which receives most attention in the Chapter, moreover, goes well beyond “only” contributing an application of an existing formalization. This formalization is truly novel in itself.

Chapter 7 has introduced a novel architecture of joint action that accounts for the interaction between higher- and lower-level coordination mechanisms. The architecture is presented as a computational-conceptual model, which can be used to inform other attempts at modeling joint action computationally. This architecture fits well with the other models in this thesis because a key assumption of the grounding model of cultural transmission is that cultural transmission is essentially a process of inter-personal alignment during joint activities. So far the work by Kashima and colleagues has focused on inter-personal alignment by symbolic communication. Chapter 7 adopts the perspective that this kind of alignment is supported by an alignment on other, lower levels as well. Our contribution is an analysis of how the interactions between higher- and lower-level alignment processes could be accounted for in computational models of cultural transmission. The discussion about the architecture and its relation to cultural transmission thereby goes beyond the consideration that has been given to joint actions in the context of the grounding model of cultural transmission. The architecture is mainly based on relevant literature from computer science, philosophy, psychology, and a particular experimental task that has revealed an interaction between higher- and lower-level coordination mechanisms.

Chapter 8 has provided an implementation of a concrete cognitive model based on this joint action architecture. The implementation is used to represent the experimental task mentioned above and to offer a preliminary evaluation of the implementation and the architecture. This contributes to the evidence in favor of the empirical adequacy of the architecture.

Throughout this thesis, we have gradually introduced models that are more faithful to the grounding model of cultural transmission than previous ones. Chapters 4 and 5 have introduced larger scale models of cultural dynamics with an idealized representation of the grounding model of cultural transmission. Chapter 6 has introduced a detailed model of the grounding model of cultural transmission. Chapters 7 and 8 have gone beyond the immediate scope of the grounding model of cultural transmission. In that, earlier chapters have mainly emphasized the exploration of the operational consequences of the grounding model...
model of cultural transmission, while later chapters have mainly aimed at contributing to its refinement. Earlier chapters have emphasized the role of the macro and micro level while later chapters have emphasized the role of the micro and nano level.

The following additional contributions have emerged as corollaries from the above mentioned work:

- This thesis has commented extensively on the value of computational models of cultural dynamics and on the process of creating such models. Chapter 3 has argued for agent-based modeling as an appropriate tool for the study of cultural dynamics but also emphasized the benefit of cognitive models. That chapter has highlighted the role of computational-conceptual and formal logics models as auxiliaries in the modeling process. The benefit of these types of models is underlined by the analysis of the computational-conceptual and formal logics models of Chapter 6 and by the role of the computational-conceptual model of Chapter 7 in its implementation in Chapter 8.

- The discussions in Chapters 2 and 6 have contributed to clarifying the notions of common ground and grounding beyond their immediate role in the grounding model of cultural transmission.

9.2 Limitations

We have analyzed the limitations of each presented model within the respective chapter. However, it is instructive to point out the major general limitations of the presented work.

First, any modeling of systems that involve humans is inherently difficult. The main reason is that human behavior is extremely complex, due to humans’ unique mental capabilities and the complex social processes enabled by these capabilities. As long as social scientists struggle with describing human behavior accurately, any model needs to incorporate assumptions that cannot be justified conclusively. Therefore, we have taken care in this thesis not to overstate the validity of our models and results. Computational modeling of social behavior in its current state can in most of the cases only be a supportive tool. However, we did argue that computational models are nonetheless beneficial for various reasons.

Second, the validation of models requires data and the more complex the system that is represented, the more data is required. For the validation of the models in this thesis, only a minimal amount of data was available. Hence, the models presented here draw their justification mainly from the theories and models they are based on and the process with which these theories and models were translated into the respective model. Therefore, we have sought to be clear about the origin of each model component, which has required a more extensive elaboration on background material than typical for computer science theses.
Third, on a more technical note, the implementations of models presented in this thesis do not include any graphical user interface. Hence, it would be difficult for non-technical domain experts to explore the implications of these models on their own. This is not a major shortcoming for the thesis itself but it impairs the “second-hand” benefit that can be drawn from it. However, we have published the source code of our models in conjunction with the publications produced from this thesis. This does enable the reproduction of our results and the extension of the work presented here.

9.3 Future Work

Some avenues for future work arise immediately from the limitations presented in the previous section. Other possibilities for future work with respect to the introduced models have been discussed extensively in the respective chapters. We comment here on a few other, more general, points.

The models presented in Chapters 4 and 5 are compatible in the sense that the information that is communicated between agents in both models is represented by simple numbers. It appears possible to import the more detailed model of common ground and grounding in Chapter 5 into the model in Chapter 4. This would allow agents to adjust their communication with each other based on their perceived common ground and likely lead to interesting dynamics when integrated with the co-evolution of cultural, social, and physical space.

The model presented in Chapter 6 provides a more comprehensive account of the intentional part of grounding than the architecture introduced in Chapter 7 does. Hence, it seems reasonable to extend the intentional level of the joint action architecture based on the results presented in Chapter 6. This would add precision to the architecture and enhance its capability for being implemented.

We also consider the further validation of the presented models, both based on additional inspection by domain experts and by the acquisition and use of empirical data, to be important. In particular, we anticipate that data collected from online social networking sites such as Facebook\(^1\) or Twitter\(^2\) could be used for this task. The advantages of this data are its massive amount and its representation of real social interactions that were not created artificially in lab settings.

From a more general point of view, the computational modeling of cultural dynamics could progress with the following achievements:

- A commonly accepted understanding of what culture actually is and formalizations of this understanding.

\(^1\)http://www.facebook.com
\(^2\)http://www.twitter.com
• An agreed and proven methodology for the development of computational models of social behavior and dynamics.

These observations certainly offer plenty of opportunity for further work.

9.4 Concluding Remarks

With increasingly multicultural societies an understanding of the mechanisms by which cultures are formed, changed, and maintained becomes increasingly important. If we are to mitigate the emergence of conflicts due to cultural differences, we need to gain an understanding of the basic mechanisms that underlie cultural dynamics. Therefore, this thesis addresses an issue that concerns all of us as fellow human beings.

We are fortunate that the increasing relevance of this topic goes along with an advanced understanding of the nano- and micro-level mechanisms of cultural transmission and with the realization that computational modeling is a suitable tool for scaling the study of social dynamics to the macro level. In effect, this thesis has jumped on both of these trains and opened a door to the computational modeling of cultural dynamics based on empirically justified nano- and micro-level descriptions of cultural transmission.
Bibliography


263
Bibliography


Bibliography


265


Bibliography


274


Bibliography


278


Bibliography


Author/s: PFAU, JENS

Title: Towards computational models of cultural dynamics based on the grounding model of cultural transmission

Date: 2012


Persistent Link: http://hdl.handle.net/11343/37827

File Description: Towards computational models of cultural dynamics based on the grounding model of cultural transmission

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.