EXPLAINING BRAINS BY SIMULATION

Doctoral Dissertation
Phil. Nat.

by

Wolfram Horstmann

Faculty of Biology
Bielefeld University

December 2003

Acknowledgements

I would like to thank all who know that I am grateful for their help.

This work was supported by grants from

University Network For MultiMedia of the State Government NRW, Germany

Federal Ministry of Education and Research, Germany

Printed on aging resistant paper in conformance with DIN ISO 9706
Gedruckt auf alterungsbeständigem Papier nach DIN ISO 9706

EXPLAINING BRAINS BY SIMULATION

Preface 7

1. **BRAINS** 11

1.1. **An explanatory framework** 11
1.1.1. Subject 12
1.1.2. Phenomena 13
1.1.3. Domains 14
1.1.3.1. *In between neural, cognitive and computational* 16
1.1.4. Disciplines 18
1.1.5. Levels of organization 19
1.1.5.1. *Explanation and the levels of organization: examples* 21
1.1.6. Dynamics and complexity 24
1.1.7. Mechanisms 27
1.1.8. Levels of explanation 29
1.1.9. Methodology 30
1.1.10. Control 33
1.1.11. Simulation 34
1.1.12. Summary and conclusion 36

1.2. **Explanations of brain phenomena** 39
1.2.1. Introduction 39
1.2.2. Responsiveness 40
1.2.2.1. *Binary neuron* 42
1.2.2.2. *Electric activity* 42
1.2.2.3. *Molecules and electric activity* 46
1.2.2.4. *Temporal change and responsiveness* 49
1.2.2.5. *Conduction: Spatiotemporal change and responsiveness* 51
1.2.2.6. *Beyond responsiveness* 54
1.2.3. Spikes 55
1.2.3.1. *Spikes as electric activity* 57
1.2.3.2. *Spikes as membrane events* 58
1.2.3.3. *Spikes as a formal event* 60
1.2.3.4. *Propagation of spikes* 62
1.2.4. The classical neuron 63
1.2.5. Synapse 68
1.2.6. Networks 77
1.2.6.1. *An example: Motion detection* 78
1.2.7. Behavior 84
1.2.7.1. *An example: Phototaxis* 86
1.2.7.2. *Further behavioral issues* 89
1.2.8. Learning 91
1.2.8.1. *Conditioning* 91
1.2.8.2. *Formal explanation of learning* 93

1.2.8.3. *Neural explanation of learning* 94
1.2.9. Summary and conclusion 97

1.3. **“Where to put information on brains!?”** 101
1.3.1. Introduction 101
1.3.2. A sketch of the brain sciences 103
1.3.3. Classification methods 105
1.3.4. Knowledge bases for brains 110
1.3.4.1. *Universal indexing* 110
1.3.4.2. *Thesauri* 111
1.3.4.3. *Glossaries* 113
1.3.4.4. *Evaluation of resources* 114
1.3.5. Conceptual frameworks for brains 116
1.3.5.1. *Textbooks* 117
1.3.6. A common scheme for the brain sciences 126
1.3.6.1. *Applying the common scheme* 129
1.3.6.2. *Problems* 131
1.3.7. Outlook 134
1.3.8. Conclusion 136

2. **SIMULATION** 139

2.1. **Simulations as media** 139
2.1.1. Introduction 139
2.1.2. Impediments for simulation standards 141
2.1.2.1. *Conceptual difficulty* 142
2.1.2.2. *Semantic variety* 145
2.1.2.3. *Complicated content* 147
2.1.3. Towards the core of simulation 148
2.1.3.1. *Simulation and media* 148
2.1.3.2. *Standardizing humans?* 150
2.1.3.3. *A theory of cognitive simulation* 152
2.1.4. Practical implications 153
2.1.4.1. *Intervention features* 154
2.1.4.2. *System features* 155
2.1.4.3. *Combined features* 156
2.1.4.4. *Things left aside* 156
2.1.5. Summary and conclusion 157

2.2. **Simulation in Science** 159
2.2.1. Scientific work 159
2.2.1.1. *A* s*cientific workflow.* 159
2.2.1.2. *Modeling and simulation* 161
2.2.1.3. *Mental simulation in scientific work* 166
2.2.2. Domains of scientific work 169
2.2.2.1. *From natural situations to natural preparations* 174

2.2.2.2. *From natural to artificial* 177
2.2.2.3. *Generating the artificial* 180
2.2.3. A generic simulation scheme 181
2.2.4. Strong and weak Simulation 187
2.2.5. Summary 189
2.2.6. Implications: “Simulation, media and explanation” 191

3. EXPLANATION 195

3.1. **Simulation as cognition** 196
3.1.1. Cognitive theories 198
3.1.1.1. *Knowledge* 198
3.1.1.2. *Learning* 198
3.1.2. Mental models 202
3.1.3. Mental simulation 206
3.1.3.1. *Representation* 207
3.1.3.2. *A conceptual system* 209
3.1.4. Simulation in action 214
3.1.5. Summary and conclusion 218

3.2. **Explaining brains by simulation: Design for a case study** 221
3.2.1. Introduction 222
3.2.1.1. *General task* 223
3.2.1.2. *Learning strategies* 226
3.2.1.3. *The model* 227
3.2.1.4. *Learning the vehicle* 229
3.2.1.5. *Error correction* 229
3.2.1.6. *Distinguishing strategies* 230
3.2.2. Material and methods 235
3.2.2.1. *Experimental procedure* 235
3.2.2.2. *Model implementation* 240
3.2.2.3. *Propulsion of the vehicle*. 240
3.2.2.4. *Model simulation* 242
3.2.2.5. *Selection of the vehicle configurations* 243
3.2.2.6. *Data analysis* 244
3.2.3. Discussion 246
3.2.3.1. *From performance to mental processes* 246
3.2.3.2. *Efficiency of mental simulation* 248
3.2.3.3. *Outlook* 251

EXPLAINING BRAINS BY SIMULATION – A SUMMARY 253

REFERENCES 259

FIGURES

fig. 1: Frog and fly. 39
fig. 2: Experimental setups. 41
fig. 3: RC-circuits. 43
fig. 4: "The container situation". 46
fig. 5: Molecular aspects of membranes. 48
fig. 6: Spike. 57
fig. 7: Outline of the 'classical' neuron. 64
fig. 8: Plasticity in Aplysia. 74
fig. 9: Motion Detector 79
fig. 10: Eye-Blink Circuit 96
fig. 11: A sketch of the brain sciences. 105
fig. 12: Typical architectures of neural networks. 122
fig. 13: A sketch of the Information Processing Approach. 123
fig. 14: 'Levels of organization' as extractable from textbooks of the Neurosciences. 125
fig. 15: A common scheme in the neural-cognitive-computational domain. 128
fig. 16: Categorization into the common scheme. 130
fig. 17: The concept of simulation. 142
fig. 18: Simulation cycle as human-computer interaction. 151
fig. 19: Characteristic of a model neuron. 161
fig. 20: Domains of scientific work. 172
fig. 21: The modeling relation. 182
fig. 22: The modeling relations in scientific work. 186
fig. 23: Concepts on molar knowledge structures. 199
fig. 24: The task. 225
fig. 25: Model of the vehicle. 227
fig. 26: Behavior of a specific vehicle configuration (overview). 242
fig. 27: Distribution of errors for a given vehicle configuration. 243
fig. 28: Learning curves. 245

TABLES

table 1: A collection of disciplines around the brain 104
table 2. Structure of different approaches to classification. 106
table 3: Concept positions in multiple directories. 108
table 4. Aptitude of several resources to represent ... 114
table 5. Summaries of the analyzed textbooks. 119
table 6. Summary of the analysis of textbooks. 121
table 7: Examples of criteria characterizing simulations. 154
table 8: Schematic overview of the task. 237
table 9: Overview of the chosen vehicles. 244

EXPLAINING BRAINS BY SIMULATION

Preface

This study examines the role of simulation in explaining brain phenomena. Consider a teacher in a biology class explaining color vision in bees to students. The role of simulation in this concrete situation can be manifold: The teacher can use an educational computer simulation to demonstrate, say, different stages of image processing of color vision. Long before the teacher stands in the classroom, brain scientists might have already applied scientific computer simulations in order to find the explanation that the teacher uses. And in the next exam, the students in the class should be able to mentally simulate color vision in order to answer the question the teacher posed. These cases indicate that the roles of simulation can be very different from each other and the notions of simulation underlying these roles are apparently diverse. The objective of this study is, thus, to describe the various roles of simulation in explanations of brain phenomena and to ask whether there is one generic notion of simulation that reconcile the various roles.

Which line of approach is best suited to find such a generic simulation scheme? First, as a kind of demand analysis, we need to know what is so special about brains that investigations and explanations of brains have to be based on simulation. This is provided in the first chapter 'Brains'. It will be argued that dynamics and complexity of brains are the most challenging task of explaining brains and that simulations are a key to unlock dynamics and complexity. Then, with a general understanding of the need of simulations for explaining brains, different forms of simulation can be analyzed. This is done as a kind of task analysis at the beginning of the second chapter ('Simulation') that focuses on scientific work as a good practice scenario for explanations that are based on simulation. By comparing different forms of simulation in science, the issue of simulation will be finally narrowed down to a generic simulation scheme that is proposed to be prevalent in all notions of simulation. This is provided at the end of the second chapter. When we know why we need simulation for explaining brains (chapter 1) and how the mechanism of simulation works (chapter 2), it is still unclear how to assess the role of simulation for explanations? Therefore explanation will be approached as a cognitive phenomenon involving perception, reasoning, learning etc. Simulation will be described as a mechanism that directs cognitive processes in order to generate an explanation. This approach to simulation allows a general assessment of the proposed simulation scheme with respect to major theoretical frameworks in cognitive science. This theoretical localization of simulation as an explanatory tool is done at the beginning of the third chapter ('Explanation'). At the end of the third chapter, it will be demonstrated that the conception of simulation as a cognitive phenomenon even makes possible to track the explanatory value of simulation experimentally.

Altogether, it will become clear that – contrary to the common notion that simulation is somewhere outside in a computer or some other device – most of it is inside our head. Explaining brains by simulation is primarily done by our brains.

Reading Advice

This text is definitely "in between" – between theory and praxis, science and humanities, brain and behavior, content and method, nature and model … So, who can read this text? The main groups I had in mind while writing were people interested in 'brains', e.g. affiliates of neuroscience, behavioral science and cognitive science, people interested in 'explanation', e.g. philosophers of science, but also teachers or journalists and, finally, people interested in 'simulations', e.g. computer scientists as well as theorists and practitioners from the educational domain. Meeting the needs of each group is a balancing act between being trivial and being unintelligible. So, I would like to please in advance to show some leniency towards sections that slip off in the one or the other direction. The reader is invited to autonomously adjust the appropriate reading focus due to prior knowledge. However, it is a particular concern of this study to satisfy the needs of different groups of readers. Therefore, I will explain the purpose of the major sections in the **Reading Advice** *paragraphs that precede the sections.*

For matching the needs of different reader groups, the general sections are supplemented with four 'Specials' that analyze specific questions in depth: two are a "Brain-Special" one is a "Simulation-Special" and one is an "Explanation-Special". 'Specials' can be read as solitary sections that do not depend crucially on the other sections. Whether or not a 'Special' is important to a reader depends on the specific interests and the prior knowledge. The shortest trajectory through the text – recommended only for experts of simulation and brains – leaves out all 'Specials'. Novices in the brain sciences are recommended to read the first Brain-Special because it can be seen as an introduction to brain sciences. For brain experts it has predominantly repetitive character. The second Brain-Special assesses the state of theoretical integration within the different disciplines contributing to the brain sciences by analyzing existing resources of information on brains such as databases, thesauri or textbooks. It provides all reader groups with a deeper understanding of specific strategies and problems of explaining brains. The Simulation-Special provides an introduction and critical discussion of common notions of simulation, also with respect to technical conceptions. Thus, it is interesting to novices in the field of simulation as an introduction – and also for all those simulation experts who sense that the concept of simulation is somewhat fuzzy. The Explanation-Special is a demonstration of how the notion of simulation proposed in the preceding sections can be assessed experimentally. The following table provides an overview of how a section applies to a specific reader group. Whether or not a section is actually to be read can best be decided directly before the respective section where the specific **Reading Advice** *is provided.*

Section	page	Type	Brain Scientist	Computer Scientist	Cognitive Scientist
1. Brains					
1.1 An explanatory framework	11	General	–	–	–
1.2 Explanations of brain phenomena	39	Brain-Special	repetition	introduction	introduction
1.3 Where to put information on brains!?	101	Brain-Special	deepening	deepening	deepening
2. Simulation					
2.1 Simulations as media	139	Simulation-Spec.	introduction	deepening	introduction
2.2 Simulation in science	159	General	–	–	–
3. Explanation					
3.1 Simulation as cognition	196	General	–	–	–
3.2 Case study	221	Explanation-Spec.	introduction	deepening	deepening

Have fun!

for Lena

1. BRAINS

The term 'Brains' stands for the subject, the phenomenon to be explained. Of course, the term has a metaphorical meaning and comprises the complete nervous system. The plural 'Brain*s*' indicates that there is not only one brain, i.e. the human brain to be explained, but that simpler nervous systems equally contribute to an understanding of brain function. 'Neuroscience', as an alternative term for naming the subject, bears a notion far too restricted to describe the respective scientific field since, for instance, Cybernetics, Neuroinformatics, Artificial Intelligence or Cognitive Psychology each play a considerable role in brain studies. However, 'Neuroscience' is the term most often used to denote the scientific field and I will do so as well when I speak of the scientific field (not the subject!) – keeping in mind that other disciplines, otherwise deserving their own credits, are implied.

Reading Advice

From the various possibilities of introducing brains, I chose an approach that will provide readers new to the brains sciences with a solid knowledge base for the following chapters "Simulation" and "Explanation". But, at the same time, this approach should unveil some implicit assumptions of brain sciences that might be even new to brain experts. In order to understand this approach it might help to think of a naïve observer who has attended many oral exams on neuroscience, now trying to sum up what candidates and examiners were talking about. The chapter 'Brains' contains three large sections. First, an explanatory framework in the form of basic concepts and terms will be provided. This is important to all readers because terminological conventions are provided that help to be clear throughout the text about what is actually at issue. Specific explanations of brain phenomena, mainly for non-experts in the brains sciences, will be presented in the follow-up section.

1.1. An explanatory framework

Behavior and cognition in animals and humans are thought to be controlled, realized and, thus, explainable by way of brain mechanisms. But brains are no easy things to explain. Complexity and dynamics of brain mechanisms encrypt them to a hardly readable book. However, the difficulties of explaining brains are unlikely to be caused by some magic properties of the brain. Rather, they might result from a lacking skill of understanding dynamics and complexity as such. If so, strategies to cope with dynamics and complexity should help to better understand brain phenomena. Dynamics can be controlled, for example, by defining discrete phases (serialization) and complexity can be broken down to constituent elements (decomposition). Such systematic procedures applied to brain phenomena shall help to decrypt some parts of the 'brain-book'.

1.1.1. Subject

Brains are involved in and involve a variety of phenomena: anatomical, electric, chemical, behavioral, cognitive, computational etc. Accordingly, various approaches exist to explain brains. Their treatment would be an effort far beyond the scope of this study. Thus, in order to limit the scope, this text will not pose the (ontological) question: “What is the brain as such?” but only the (epistemological) question “What contributes to explanations of brains?” An appropriate situation that reflects the approach taken in this section is that of a neutral third person visiting an oral exam on a neuroscience course in order to find formulations of the criteria by which the examiner assesses the consistency of the candidate’s utterances.

So, what is to be explained? A general explanatory target of brain studies is the question how a certain *function* is realized by the brain, e.g. color vision, temperature sensation, spatial memory etc. Functions as such often are not directly observable, but become overt as a means for explaining a *behavior* of a given system: a frog flicking a fly, a bee’s preference for, say, a blue flower, an iguana looking for a sunny rock, a rat’s improved orientation in a formerly unknown maze etc. Behavior usually relates to an environment in a functional (often called “adaptive”) manner. Here, for the time being, the scope of the brain studies end: *Brain studies seek to explain how brain phenomena make possible adaptive behavior*. Unfortunately, the whole story is not that simple: the system showing the behavior and its environment is not always very concrete as in the case of a frog flicking a fly. On the contrary, the factors determining function are frequently far from the actual animal behavior, e.g. a gene’s expression probability in a synapse, a neuron’s electric activity in an isolated brain slice, the performance of a human in a reaction time task or a trait of a brain-like artifact such as a simulation or a robot. These examples illustrate that principally any low-level (micro-) phenomenon potentially contributing to valid explanations of brains might be taken as a building block of an explanation. Thus, one difficulty in explaining brains is to arrange low-level phenomena in a way that produces a sound explanation of the global, high-level phenomenon (e.g. behavior). Additionally, low-level phenomena are themselves typically not exclusively hosted in a single explanation. Synaptic transmission, for example, is a low-level phenomenon that plays a role in many explanations of behavior. Not a general process of synaptic transmission is applicable in any explanation, but the specific ‘configuration’ of synaptic transmission has to be found. Thus, another prevalent difficulty in explaining brains is to specify the concrete instance of a mechanism in order to explain the global phenomenon. In sum, relating low-level phenomena to adaptive behavior is a general explanatory strategy in brain studies.

1.1.2. Phenomena

What do brain scientists observe? When brains are freed from their bodily shell, the first thing observed are anatomical phenomena, i.e. the size, form, color etc. Since the naked eye does not reveal enough detail, often additional methods such as manual isolation of brain structures and section techniques are applied. Brain tissue does not provide sufficient contrast and is, therefore, most of the time artificially stained with histological methods. The final brain preparation yields expressive pictures of microstructures, such as cells, assemblies, areas, regions etc. Anatomical techniques were systematically developed and improved since major breakthroughs of histological staining during the turn to the past century (see Ramón y Cajal 1906; Golgi 1906). Today, sometimes even more instructive pictures can be taken without opening the body, e.g. with fMRI (functional Magnetic Resonance Imaging) techniques (see fMRI Data Center 2003). Anatomical phenomena relate to the detailed spatial organization of brains. Gross physical phenomena such as density, weight etc. are descriptive, but seldom contribute to explanations of brain function.

Electric phenomena were discovered very early (see Galvani 1791) and, unlike gross physical phenomena, they yielded powerful approaches to explanations of brain function ever since. At the beginning, muscles were stimulated electrically and contractions could be observed. Today, even electric phenomena relating to abstract concepts such as ‘decisions’ can be observed by imaging techniques and can be stimulated by micro-electrodes (see e.g. Platt 2002). Electric phenomena are considered to be most closely coupled to functions of the brain, i.e. it is thought that the functional organization can adequately be explained by studying electric phenomena.

It was shown that electric phenomena can also be understood as chemical phenomena (see e.g. Nernst 1888). The responsiveness of single neurons is frequently termed an electrochemical phenomenon since a given electric activity is determined by the relative concentrations of ions solved inside and outside of the cell. The deeper understanding of electrochemical coupling also yielded new insights in a very special phenomenon – the action potential, often called spike (see Hodgkin & Huxley 1952). This is noteworthy since spikes are impulses that are actively generated by neurons and allow brains to encode a continuously modulating signal (e.g. a pure tone registered by sensory cells in the ear) into a series of discrete events. Such a translation process is a basic computational phenomenon. Spikes are important for understanding neural codes.

The computational perspective on brains (see e.g. Churchland & Sejnowski 1992) yields outstanding explanations of complex and dynamic phenomena such as humans extracting a single faint voice from the noisy babble on a cocktail party (see von der Malsburg & Schneider 1986) or desert ants finding their way home after an excursion by utilizing sun-position (see Wehner 1994; Lambrinos *et al.* 2000). Computational phenomena are abstract in nature, but make sense immediately if seen in a behavioral context.

Explaining behavioral phenomena – in the sense of the behavior of a frog or a fly – necessitates an integration of most phenomena described before. Explaining behavior, therefore, represents one of the most ultimate tasks in the venture of explaining brains. A specific sub-domain of behavioral phenomena are cognitive phenomena for which integrative explanations more and more succeed (see e.g. Gazzaniga 2002 for a general account). The task of explaining cognitive phenomena as brain phenomena suffers from the complexity that unfolds when trying to break them down to brain mechanisms. They comprise a vast amount of constitutive brain functions. Therefore, cognitive phenomena call for an ambitious integration of the different phenomena over space, time and complexity.

In sum, an explanation of a functional brain can involve anatomical, electrical, chemical, molecular, computational, behavioral and cognitive phenomena.

1.1.3. Domains

Classes of phenomena form domains. Differentiating between all different phenomena in any instance of explanations is demanding and sometimes not necessary. But some differences are striking! For example, the brain is most often conceived as a biological, 'wet' matter and less often as a computer or a cognition carrier. The 'wet' notion resides in the neural domain of brain studies. The neural domain subsumes all phenomena that are typically studied by (natural) sciences, i.e. physics, chemistry and biology. They form a coherent domain since it is widely assumed that explanations for one phenomenon are translatable in terms of explanations of another phenomenon, for instance as electric activity in a neuron is explainable in terms of ionic changes at the neuronal membrane.

As already indicated above, those phenomena that somehow fall out of the neural domain are computational and cognitive phenomena. Cognitive phenomena are thought to rely on brains, but the concrete mechanisms are difficult to tackle. Cognitive phenomena, e.g. traced by reaction-time experiments (a behavioral phenomenon!) are typically explained in cognitive terms (self contained in psychological theories) that are not directly translatable in neural terms. Yet, in single cases or to varying degrees such translations succeed. Beyond this problem of theoretical integration, the adequacy of a neural description of cognitive phenomena may vary from case to case and is still an unresolved issue (see also 1.2.7.2). In general, explaining cognitive phenomena might be conceived as a benchmark test for integrative studies on brain function.

Computational phenomena play a special role in brain studies. It is implicitly assumed (and very often explicitly demonstrated) that brain functions can also be realized in computer programs or machines. This characteristic decouples computational phenomena from the neural domain by definition. This does not imply that computational phenomena are not translatable in neural terms. But it implies that neural implementation is not a necessary condition for a computational phenomenon to be observed. Contrary to the translation of electric in chemical phenomena – where electric phenomena and chemical phenomena are mutually conditional – computational phenomena and neural phenomena might be observed solitarily and consequently form separate domains. Computational phenomena are remote from the neural domain, for the basic notion is that computational phenomena are independent from the concrete form of implementation.

In summary, we arrive at a tri-part distinction of domains in brain studies in which mechanisms explaining adaptive behavior can reside in: a neural, a cognitive and a computational domain.

1.1.3.1. *In between neural, cognitive and computational*

The relation between the neural, cognitive and computational domain, however, has been subject of many philosophical discussions that shall not be neglected completely here. The discussion on the relation between the neural and the cognitive ('mental') has often been caused by the subject of reduction. What is this problem about? It is widely accepted that mental processes rest on neural processes. Why then should we distinguish a mental domain from a neural domain if we can reduce all mental phenomena to neural phenomena? A polarizing account is put forward as "eliminative materialism" (see Churchland 1981; Churchland 1986) that predicts a total disappearance of explanations involving mental phenomena since they will all finally be 'translated' into more basic neural ('material') phenomena. However, such reduction of (explanations of) mental processes to neural processes is itself challenged by the view that coexistence and division of scientific labor in the respective disciplines are valuable alternatives. For example, Bechtel et. al. (2001a) propose that different disciplines concern different phenomena on different levels on the natural hierarchy. A supporting argument for this view can be developed if a concrete final form of such reductionist explanations is envisaged: for reducing high level phenomena to low-level phenomena, considerably more mechanistic explanations have to be taken into account. Each mechanism comprises further elements, activities and certain causal organizations. The principle of reduction can be applied virtually endlessly to arbitrary low-level domains. At a certain point of reduction, the consideration of more and more low-level mechanisms renders the whole venture uneconomical. This point is reached, for example, when the purpose of the venture does not profit anymore from the consideration of low-level explanations and, at the same time, high-level explanations provide the same functionality. For example, if spatial attention (the high level phenomenon) was the explanandum, an explanans might be based on cable theory (a low-level approach based on biophysical properties of single neurons, see also 1.2.2). But the explanans might also be based on a purely functional, phenomenological model without any reference to low-level phenomena. The aptitude of the explanans is determined by the application context. Consider a robot as the target platform for the spatial attention mechanism: the mechanism in the robot can be based on both low-level cable model and high-level functional model. The aptitude of the model would then be assessed by the criterion whether or not the model eventually realizes the desired behavior of the robot. If the behavior can be realized with both models equally well, the more economical model will be preferred. Since high-level models are usually more parsimonious (e.g. in terms of participating elements), they will usually compete low-level models. Thus, reductive low-level (e.g. neural) explanations are not generally superior – depending on the application context, functional high-level (e.g. cognitive) explanations might be superior in terms of economy. However, theorists could object that, even though economical constraints work in science, they usually (and fortunately) do not completely determine scientific inquiry. Therefore, the call for a unified account of a given phenomenon (such as spatial attention) would presumably not fall silent until all possible reductions are made.

The economical argument might be too weak too lastingly satisfy deeply theoretical issues, but low-level explanations might also be inferior for another reason: explanations involving mental processes might be more appropriate because neural explanations might show to be so complex that they exceed the cognitive capacity of scientists and recipients. Cruse (2001) puts forward a closely related argument. He compares the levels of description in the Neurosciences (e.g. molecular or behavioral) to different programming languages (e.g. object-oriented vs. assembled) and concludes that a choice between them is not a question of truth but a question of aptitude. This choice goes along with the limited perceptual abilities that hinder the experience of the complete reality of a system.

Similar to the problematic relation between the neural domain and the cognitive domain, the neural domain and computational domain also have border disputes. On the one hand, there are well-founded programmatic accounts of what computational neuroscience should be about (see e.g. Schwartz 1990; Churchland & Sejnowski 1992). On the other hand, critical accounts put forward that the computational domain is not related in a useful manner to the neural domain. Grush (see 1997) provides an analytical account of the relation between algorithm and environment, computation and representation and suspects a ‘semantic challenge’ to computational neuroscience. He assumes that there is a principal difficulty in showing that computations in the brain do something meaningful, analogous to the problems of integrating semantics in information theory. However, with 'informational semantics' (Dretske 1981) or 'biosemantics' (Millikan 1984; 1989) constructive attempts can also be found. Another critical argument against computation is put forward by Daugman (1990). Within the different approaches of explaining brains with metaphors (i.e. hydraulic, electric, computational etc.) he makes out computation as one of the weakest metaphors because evidence for brains carrying out logical operations is scarce. However, most opponents of computational neuroscience presume that computational neuroscientists seek for computations as the brain's central purpose. Even though this might be true in some cases, one might concede other computational neuroscientists that they see computations just as operations necessary to achieve meaningful representations – not as the representations as such. In this sense, computations describe the transfer functions that lead to meaningful representations and computational neuroscience provides an exact language and methodological framework for this.

In spite of the problems indicated above, it seems helpful to differentiate between the neural, the cognitive and the computational domain in order to respect all phenomena relevant for explanations of brains.

1.1.4. Disciplines

The different phenomena and domains in brain studies are object of a variety of scientific disciplines. Beside the 'classic' disciplines, i.e. physics, chemistry and biology, the younger disciplines psychology and computer science are most relevant for brain studies. Beyond these, numerous small special disciplines emerged within the last decades. Each special discipline is focused differentially on the neural, cognitive or computational domain. However, as special disciplines (e.g. cognitive neuroscience) have successors in 'special-special' disciplines (e.g. cognitive neuropharmacology), the relevance of special disciplines increasingly becomes a matter of taste. Thus, special disciplines shall not be reviewed in detail here. A short, subjective sketch of disciplines in the brain sciences and their orientation in the neural, cognitive and computational domain is given later (see also 1.3.2 and fig. 11).

1.1.5. Levels of organization

Brain phenomena can be located at various levels: spikes at the neuronal level, synaptic noise at the sub-cellular level, memory on the systems level etc. The notion of levels is very common in brain studies and important to understand and to explain brain phenomena (see e.g. Churchland & Sejnowski 1992). Moving from one level to the next (e.g. from sub-cellular to cellular) means moving from a lower level to a higher level. Thus, the notion of levels entails that the brain's organization is hierarchical in nature. How can this hierarchical organization be described? The simplest form of hierarchical organization is anatomical, i.e. refers to the spatial organization. A 'zoom' through a spatial organization can firmly be anchored at the cellular (neuronal) level. The neuronal level contains, for instance, membranes and synapses. Zooming inward (getting smaller, moving downward the hierarchy) opens the sub-cellular level with ion channels in membranes and vesicles in synapses. Below unfolds the molecular level with membrane channels being proteins and vesicles containing transmitters. The electric level, then, refers to ions moving through membrane proteins or electric interactions causing conformation changes of proteins during transmitter binding. Lower levels than the electric level, e.g. interactions of electromagnetic fields should not be neglected, but are scarcely considered in this study. It should also be noted that there is no distinct chemical level of organization in this study. The reason is that, although chemical phenomena play a crucial role in low-level mechanisms, they are readily contained by the electric and the molecular level of organization.

Zooming outward from the neuronal level (getting larger, moving upward in the hierarchy) reveals a variety of networks: Small networks (circuits), constituted by few neurons, can build a reflex arc or act as an ensemble. Large networks (modules) can build a functional column in the human cortex, a ganglion in an arthropod or regions such as a retina. 'Systems' denote the intermediate level between networks and behavior, but the term 'system' is not well defined. It relates to such different issues as sensory systems (e.g. visual, auditory), motor, memory or transmitter systems. However, systems (as a collective term for intermediate levels of organization) denote the last level before the neural level – and therewith the brain and the whole nervous system – is left behind. Passing the organismic level, finally, reveals the view on the (behaving) individual in its environment.

The spatial levels described so far largely correspond to the different phenomena introduced above – especially the phenomena subsumed in the neural domain (see also 1.1.2). However, space is just one dimension. The intuitive appeal of the spatial dimension often distracts from another basic dimension, namely time. The temporal dimension is usually less intelligible because it does not provide clear-cut categories: categories cannot be formed as they can be formed in the spatial dimension because time is usually not ‘contained’ in a way brains contain neurons and a synapses contain vesicles. The temporal domain extends on a continuous scale. Only a coarse differentiation between ‘timescales’ therefore is common practice, for example neural time around milliseconds, ‘real-time’ (behavioral) around seconds, lifetime around days and years and evolutionary time above that. Commonly, descriptions in the temporal domain are relative, i.e. simply achieved by a distinction between before and after a given event, for example, a presynaptic event at time $x(t)$ vs. a postsynaptic event at time $x(t+1)$. This event-relative thinking reveals ‘changes’ of phenomena, e.g. a change from a large presynaptic to a small postsynaptic signal. A pure temporal analysis of these changes yields a phenomenological description of the relation $y = x(t) - x(t+1)$, e.g. a value for the difference between before and after. However, the purely relational temporal analysis does not give any clue *why* something changed. This represents a fundamental difference between the relational notion of the temporal domain and the hierarchical notion of the spatial domain. Concerning the spatial dimension, a simple answer for the question “Why do you think is N larger than x?” can be offered, namely “Because N contains x!”. The hierarchical notion of the spatial domain is, literally, self-contained (such as a card house). This does not hold for the temporal domain. If it is asked “Why do you think is t later than $t+1$?” the corresponding answer “Because t relates to $t+1$” doesn’t make sense. A relation as such has no causal implication, while containment as such suffices to explain a difference in size! Explaining temporal changes requires hypotheses on cause and effect that determine the respective event. Thus, the focus on the temporal domain almost inevitably introduces causality.

Explaining the difference between the postsynaptic and the presynaptic signal can take several forms that involve temporal and spatial aspects to varying degrees. A pure temporal-relational answer would be “the signal before synaptic transmission is larger than the signal after *synaptic* transmission”. Purely computational answers are also phenomenological “the postsynaptic signal was damped (by factor of *x*)". An abstract causal-temporal answer would be "the postsynaptic signal is damped by a resistance". Causal answers would imply a mechanism for the abstract cause "the postsynaptic signal is damped by a synaptic resistance". A causal explanation would imply mechanisms on different hierarchical levels, i.e. in terms of the amount of released transmitter, receptor bindings etc. This shift from the temporal dimension to the spatial dimension becomes necessary because the temporal dimension alone only allows relational or phenomenological statements. The combination of spatial with temporal aspects introduces causal considerations. Space, time and causality, then, make up the explanatory framework for brain studies: the *levels of organization* (see also 1.3.2 and fig. 14).

1.1.5.1. *Explanation and the levels of organization: examples*

The levels of organization are the common explanatory framework for brain studies. The general idea is straightforward: Elements have spatial and temporal properties and are organized by causal relations. Spatial properties are hierarchically organized (..., sub-cellular, cellular, network, ...) and temporal properties typically show a before-after relation. If this were all, explaining brains would be a fairly simple matter. Whether this turns out to be true, shall be discussed on a concrete case.

Consider the example of the *resting potential* that is the voltage difference of the neuron's inside relative to its outside. A popular starting point for explanations of the resting potential is this: Ions playing a role for the resting potential are distributed unequally on both sides of the membrane, i.e. on one side of the neuronal membrane are more ions of a given species per unit volume of water than on the other side, say A^+B^- dominate inside and X^+Y^- outside. Concentration gradients tend to level out by way of ions moving from the less concentrated side to the more concentrated side. However, if there is no way for the ions to pass the membrane there will be no ionic current. These concentration gradients, by itself, have no effect on the charge of the membrane as long as positively and negatively charged ions are equally distributed (as many A^+ as B^- inside and as many X^+ as Y^- outside), i.e. there is no surplus of positive or negative charges on either side of the membrane. But the neuronal membrane is selectively leaky (permeable): only (small) ions of one charge type (say A^+) can pass, while the (large) ions of the other charge type (say B^-) are prevented to pass by the limited size of the openings of the membrane (channels). As a consequence, all ion species not permeating through the membrane can be ignored (only A^+ has to be considered). Now, the permeating ion species tends to move towards the lesser-concentrated side. As it moves, a charge is transported to the other side: the membrane is charged. (When A^+ moves outside the outside is more positive than the inside, or, put in another way, the inside is charged negatively relative to the outside because A^+ left a negative counterpart B^- behind.) The emerging charge prevents the concentrations from leveling out. Why that? The more ions move to the other side, the more the membrane is charged. Charges tend to level out as concentration gradients do. So, the ions are driven in the opposite direction of the concentration gradient (A^+ is needed inside for making the inside more positive again). Two forces, an electric and a chemical, act against each other. In the resulting situation, some ions move through some channels outside in order to level out concentration, while others move inside in order to level out the electric potential. The resting potential, then, is the dynamic equilibrium that turns up when driving forces of the concentration gradient and the electric potential are equal.

In this explanation of the resting potential, only two organizational levels are directly involved: the sub-cellular (the membrane) and the electric (charged ions). The organizational levels are distinct with respect to their temporal and spatial properties: the membrane is larger and less dynamic in space than ions. Accordingly, the organizational levels are distinct with respect to their causal organization: the behavior of charged ions can be explained with concepts like electromagnetic interactions, Brownian motion etc., i.e. concepts that are to be found in a physics textbook. The membrane is only explainable by referring to classes of macromolecules (not to be found in physics textbooks but in textbooks on organic chemistry or molecular biology). But, although only the electric and the sub-cellular level are directly involved in the explanation, all other levels are implicitly assumed as premises of the levels of organizations framework. For instance, the cell must be closed because a patch of membrane swimming isolated in a solution would not cause gradients and, therewith, significant ion fluxes. This sounds trivial, but is an essential premise at the cellular level. A premise at the molecular level is, for example, that the membrane's property of being semi-permeable is somehow brought about at some other level. Explicating all premises on all levels (and all level transitions) is not necessary to explain a phenomenon (i.e. the resting potential). But the explanation of the actual phenomenon is leaned against the backdrop of the levels-of-organization framework by taking for granted that all explanatory 'snippets' fit together, and that all levels are interconnected. It is assumed that everything belongs to the same organization[1].

This situation entails the possibility to extend explanations through various levels. For example, the membrane's property of being semi-permeable is realized by channels as protein-based structures having certain electrically determined conformational states that allow a selective permeation of ions. The gaze can wander through the levels of organization with various degrees of freedom. Which level is to be considered and which level is left aside is primarily determined by the actual question posed. There are no straight rules for designing an answer to the question posed. In the case of the resting potential considered above, the phenomenon is largely explained by the interactions of activities on two hierarchical levels that are *not* directly connected in the hierarchy, namely the electric level (ions) and the sub-cellular level (membrane). In a perfectly systematic explanation that applies a straight spatiotemporal framework, the molecular level should be in between the electric and sub-cellular level. Yet, such explanations are functional. This indicates the "forgiving" character of the levels of organization as an explanatory framework. Moreover, spatial, temporal, computational or causal aspects are 'blended' in explanations on brain phenomena. Consider the following sentence: *An action potential elicited by an air puff is synaptically*

[1] It should be noted, though, that there are flaws and pitfalls in the idea about the levels of organization. If not applied thoughtfully, the levels of organization will do more harm than good. The intuitive power enables authors to allege relations that are neither evidenced nor explained. For example, the maintenance of the electrochemical equilibrium at the neuronal membrane ('resting potential') is typically explained by the sodium-potassium pump – an exchange mechanism of charges across the membrane. When charges move across the membrane, say, as a result of synaptic transmission, students frequently assume that the pump has to bring these charges immediately back to where they came from because the cell would otherwise run quickly out of charges. They do not realize that the overall concentrations are only marginally affected by displacement of only few charges, while the actual membrane potential is significantly changed. Textbook accounts frequently appeal to the organizational principles (i.e. 'potential maintained by pump') and neglect the concrete spatiotemporal parameters, which would give an idea about the actual quantities. These incautious analogies and the resulting categorical errors have a considerable probability of occurrence in this explanatory framework. It is generally assumed that high level phenomena (e.g. resting potential) are reducible to low-level phenomena (e.g. moving charges), and, hence, rules governing the low-level will also be present in the high level – disregarding the possibility that the level transition itself will render low-level rules invalid. It will not be claimed here, though, that authors should give a full account of all the phenomena on a higher level in terms of the lower level. This would make many neuroscientific notions unintelligible. Rather, it shall be made clear that the idea about the levels of organization is not an exact theoretical framework but an optimistic and fruitful rationale.

transmitted from the sensory cell to the next neuron. This simple sentence reveals any of the explanatory aspects introduced so far: spatial (next neuron) temporal (first air puff, then action potential, then synaptic transmission) computational (transmitted), causal (elicited by an air puff). These examples shall illustrate that there is no strict prescription for explanations in the levels of organization. Usually, the elements and activities necessary to explain a phenomenon are not explained level by level and step by step, but constructed situatively according to the question posed. However, there is a general assumption in the background that it is conceivable and it is in general possible to explain each phenomenon level by level and step by step. Thus, it might be naïve to believe that explaining brains in the levels-of-organization framework is a fairly simple matter, but the framework definitely helps to simplify explanations by providing a conceptual (and lingual) 'stage' for brain phenomena.

1.1.6. Dynamics and complexity

Fairly simple explanations on brain phenomena can be developed in the levels-of-organization framework. But sometimes, if not often, explanations of brain phenomena are not at all thought to be easy. Given that these difficulties are not a result of the wrong explanatory strategy, what then? Brain phenomena are difficult to understand because they are dynamic and complex. For representing dynamics and complexity in explanations they must be also conceivable as intrinsic properties of the levels-of-organization framework. Consequently, a descriptive account of how dynamics and complexity are conceivable in the terms introduced so far shall be provided.

A subjective notion of complexity becomes immediately evident when the various levels of organization, the elements and activities and the possible causal relations are tried to be conceived altogether. It is simply not possible, because the capacity of mental processing is limited to a few elements and activities at a time (see also 2.2.2). This subjective notion already implies that complexity results from a specific property of the levels of organization, namely to be compositional. This means that an element can be composed of other elements and can itself be part of a composition. *Compositionality* refers primarily to the spatial organization. For example, a neuron is composed of sub-cellular compartments and composes networks. Additionally, changes in time, i.e. activities are to be considered. Activities relate to temporal organization and make possible that elements can also be dislocated and, thus, spatial levels can be transcended. For example, ions moving through channels can transcend sub-cellular, cellular and networks levels. *Change* in time and space make out the spatiotemporal organization. Finally, elements and activities are causally related. Ions move from this to that side of the membrane because other ions are on this or that side of the membrane etc. Causal organization allows us to relate elements and activities to almost any other element and activity. As long as the systems under scrutiny have a small number of elements, activities and causal relations, all possible relations might be traceable. But the brain is not small. The large number of possible relations between elements, activities and causal relations – the *combinatorial power* – is the third ingredient of complexity.

Complexity is determined by change, compositionality and combinatorial power of constitutive elements. There are numerous accounts for a formal and quantitative description of complexity (see e.g. Gell-Mann 1995). But for assessing the difficulties complexity causes in understanding brain phenomena, these accounts are not immediately relevant. Most important in the present context is that complexity can cause difficulties in understanding brain phenomena since it makes the (possible) explanation exceed the limited capacity of observers.

Brain phenomena are typically dynamic – they change while they change[2]. How can dynamics be described in the levels-of-organization framework? In the above description of complexity it was stated that for small systems it could be true that all relations between elements, activities and causal relations are traceable. Then, the state of the system at a specific instance of time t can be described completely. The same shall be true for the next instance of time t_1 and all the following states until the final state t_i where the analysis ends. In an overall view, the trajectory through the state space of the system can have a specific form that can be described by a rule. For example, the trajectory of a frog's tongue towards a specific point (e.g. at a tree trunk) could be characterized by a function that depends exclusively on the final coordinates of the tongue tip. Expressed verbally, the rule that specifies the trajectory is just: "Propel with a velocity *v* to point *ABC*". There is

[2] The notion of dynamics meant here is temporal. There is, however, a notion of dynamics prevalent in physics that refers to force and kinetics. This notion is not meant, here.

no need to consider an intermediate state of the system in the specification of the trajectory between the state at t_1 and the state at t_i. This is a temporal change – but it is not dynamics. Now consider a fly escaping the frog's tongue! (The motion control system of the fly is assumed here to be so simple that it is completely describable.) First, consider the frog's tongue tip to be fixed on a given position (e.g. because it froze on a iced tree trunk after propulsion). The motion control system of the fly could specify the maneuver prior to the actual execution. This would equal an application of a rule, again. However, another (and evidently biologically more plausible) alternative is that the trajectory of the fly is specified continuously during the flight. A given sensory input at t_1 depends on the position of the tongue at t_1 (how large the frog's tongue is on the retina) and specifies the motor output at t_1 that determines the position at t_2 that determines the sensory input at t_2 that determines… Each state of the system includes a response that in turn influences the next state. Such a phenomenon of motion control is dynamic.

Everybody has an intuitive understanding of dynamics as a temporal phenomenon. Watching a film, for instance, is a dynamic process. But, though intuitive, it is not trivial as analytical accounts demonstrate. For example, why do we see a helicopter moving on the screen when we follow it with our eyes and the retinal position stays the same, whereas we see a coffee cup standing still when we move our eyes and the retinal position changes? (A simple answer is that humans process retinal images inside a general, spatial reference frame that 'subtracts' and compensates eye movements.) This example illustrates that accepting and handling dynamics is something else than understanding dynamics. Dynamics are more difficult to trace when the dynamic processes (such as the continuously processing motion control system of the fly) interacts with other changes (such as a straight moving frog's tongue) or other dynamic processes (such as an elastically fluttering frog's tongue). Dynamics can be made even more difficult by regarding two inter-dependent dynamic systems, for example, if the frog could adjust its tongue trajectory during propulsion in relation to an observed escape maneuver of the fly that can adjust its maneuver with relation to the new tongue trajectory etc.

A concrete problem in understanding dynamics is, as already indicated above, that no simple rule can be applied to the participating elements. The rule itself must contain change, e.g. the enlargement of the tongue in the retinal image of the fly. The change can have a qualitative character ("becomes larger" or "becomes smaller"), but often also quantitative measures are necessary ("becomes larger with an extension of *x*"). Thus, dynamics frequently call for quantitative thinking. But beside the specific comprehension problems introduced by quantitative estimates (see e.g. Ploetzner & Van Lehn 1997) the difficulty of understanding dynamics is comparable to that of understanding complexity: the capacity for processing is too limited for taking into account all the case differentiations that would be necessary to assess dynamics. Since each state determines the next state, dynamics enforce a consideration of multiple intermediate states of the system (as opposed to a merely 'changing' system that can be understood by comparing the initial and the final state). Causally relating elements and activities in an earlier phase to elements and activities in later phases might fail because the mental system was meanwhile loaded with intermediate states so that the earlier state is simply not present anymore for comparison ('forgotten'). Thus, particularly the capacity of relating elements in a temporal context might cause problems in understanding. Dynamics point at the limited temporal extension of representational capacity as an impediment for understanding brain phenomena.

Dynamics and complexity are interrelated. Both rest on change and both push understanding to the limit by the cognitive load they cause. That these interrelations may amplify each other in causing comprehension problems should be evident. And both are characteristic properties of brain phenomena. So, there is no way out of tackling the comprehension problems caused by dynamics and complexity if we want to understand brains. Since they both are limited by capacity, a promising strategy to handle them should be reduction. Chopping of large systems in smaller systems (decomposition) and chopping of large processes in phases (serialization) seems the most valuable way. Orientation inside the resulting explanatory chops can be backed by the levels-of-organization framework that should help to put them together again. In this sense, complexity and dynamics of a given system might not be fully explained (and experienced) in one instance of time, but in a journey through consecutive phases of explanation.

1.1.7. Mechanisms

Explanations of brain phenomena must be consistent in order to serve their purpose. What are the characteristics of a consistent explanation for a brain phenomenon? A typical consistent explanation of brain phenomena involves a 'mechanism' (see Machamer *et al.* 2001). For example, the question: "How do I detect a air puff?" calls for a mechanism to be answered. A possible answer is: "Your skin contains mechanoreceptors. The wind deflects a hair (upper part) on the skin relative to the base (lower part) that is anchored in the skin and stretches the neuronal membrane. This opens channels for charged substances (ions) that can now pass the membrane. The resulting electric potential is processed from that moment on as a representation of the onset of the air puff." The mechanism is composed of nested elements and activities (curled brackets indicate beginning or end of another element):

{air puff | you {skin {mechanoreceptor {upper part | lower part {membrane {channels {ions}}}}}}}.

Causal relations, according to which elements and activities behave, are added, e.g. causal relations about mechanical deflection, membrane potential, diffusing ions etc. In this way it is determined how elements and activities are temporally and causally related to each other and, therewith, how changes (i.e. an air puff) cascade through the elements and activities (the mechanoreceptor, membrane, ionic currents etc.). The result of the cascade is the global causal relation (e.g. between you and the air puff) and serves as an answer to the question posed (e.g. mechanism of air puff detection).

How exactly does a mechanism 'connect' elements, activities and causal relations over several levels of organization? For example, (chemical) synaptic transmission is the mechanism by which a given signal can be transmitted from one cell to another cell. Synaptic transmission is much too complex to be explained completely here, but the principle is that the spatial gap between neurons cannot be easily overcome by electric signals So the electric signal is transformed into a chemical signal, namely a 'cloud' of transmitter molecules that is released as a consequence of the incoming electric activity at the presynaptic membrane and moves along its concentration gradient towards the postsynaptic membrane. Here it is translated back to an electric activity. Many elements, activities and causal relations are present in this excerpt of an explanation of synaptic transmission. On the cellular level of organization are the two cells, on the sub-cellular the two membranes, on the molecular the transmitter, on the electric the electric signal. The elements are predominantly bound to each other by causal relations. The cause that makes the transmitter move, for example, is the concentration gradient (diffusion). The cause 'diffusion' transcends a single spatial level in that it arranges elements from different levels of organization, namely electric activity, transmitter, membrane, cell in a common framework – just as a stage direction for actors and requisites in a screenplay. In this sense, a mechanism is a script that produces an explanation. Mechanisms are specific configurations of elements, activities and causal relations at different levels of organization. But a mechanism not only describes one state of a system – it also implies the rules that specify the succession of (at least) two states. Usually, mechanisms have a larger scope than just two states of a specific system, but are general causal patterns. This is why they are discussed to take over the role that natural laws have in physical explanations in biological explanations (see e.g. Sober 1997). Mechanisms are concrete explanatory strategies for brain phenomena. Because of their ability to align state transitions, they can help to handle dynamics and complexity.

1.1.8. Levels of explanation

Beside the levels of organization that represent the natural hierarchy or ontology used in neuroscientific explanations, there are also levels of explanation that can be distinguished (see Machamer *et al.* 2001). Three levels appear most relevant for answering a question or for explaining a phenomenon. The purely mechanistic explanation refers to the isolated question or phenomenon (level 0), e.g. detection in the wind-puff example. The mechanism can be put in a broader context, e.g. adaptivity of wind detection in terms of detecting falling trunks. This contextual aspect of the explanation yields a function of the mechanism and is accompanied by a shift of the focus on a higher level (level +1). Finally, the constitutive (reductive) aspect of explanation finally refers to further mechanisms the actual mechanism is grounded on (level -1), e.g. explanations on how the air particles hit the surface of the hair, how the density of the lipid layer of the membrane is influenced by the deflection etc. It should be noted that the actual mechanism (the explanandum) is distributed over several levels of organization and can imply several other mechanisms that themselves can be distributed over several levels of organization. This implies that the relation of the levels of organization and the levels of explanation (isolated, contextual and constitutive) is *not straightforward*. The levels of explanation represent the actual focus of an explainer in the levels of organization. As explained above, an explanatory focus involves elements and activities from various levels of organization (e.g. electric and sub-cellular). Moving down one level of explanation does not necessarily mean to move down one level of organization – it might be none or more than one. For example, it might be a detail on an intermediate level of organization that is additionally considered (e.g. the role of calcium for the electrochemical coupling in synaptic transmission). It is improbable however, that the level of explanation increases as the level of organization decreases and vice versa.

Since there are already enough levels to keep track of, it might prove helpful to use the terms mechanistic, functional and reductive *modes* of explanation for isolated (level 0), contextual (level +1) and constitutive (level -1) explanation, respectively. But it is important to keep the difference between the mode of explanation and levels of organization in mind because they explain why the levels of organization sometimes are not applied systematically, i.e. why certain levels are considered relevant for a phenomenon, while others are neglected.

1.1.9. Methodology

Brain scientists apply a specific explanatory framework: brain studies reveal anatomical, electrical, chemical, molecular, behavioral, computational and cognitive phenomena. A phenomenon is typically explained in terms of a mechanism that relates elements and activities by causal relations. The phenomenon resides in a natural ontology (the levels of organization) that is determined by spatial, temporal and causal dimensions. The levels of organization most relevant for explaining brain phenomena are electric, molecular, sub-cellular, cellular, network, systems, organismic, environmental. The mechanical, functional and reductive modes of explanation determine the actual focus set on a phenomenon.

This explanatory framework is operationalized by way of experimental procedures. The phenomena that stand at the beginning of an explanation are basically observations that a scientist makes by applying a certain method, e.g. anatomical (surgical and histological), electrophysiological, imaging techniques etc. These methods yield specifications of the elements and activities that are thought to play a role for the explanation of the phenomenon, i.e. the phenomenon is decomposed into elements and activities and localized (cf. Bechtel & Richardson 1993) at the levels of organization. Consider a scientist – for the sake of simple language the scientist is defined to be female – developing a hypothesis about a neural circuit that determines the spatial and temporal coordinates of the frog's tongue so that it hits the fly. She tries to find out how elements and activities are causally related for participating functionally inside a mechanism. (Usually, hypotheses refer to a single causal relation between two elements or activities, whereas mechanisms comprise a number of causal relations between several elements or activities.) The hypothesis is used to make a prediction about the acting of the elements. The prediction is tested in the experiment. Therefore, the scientist has to specify an experimental design that shall operate the mechanism in such a way that the action of a single causal relation is revealed. For example, if the hypothesis is that the frog's catching mechanism works most precisely for stimuli with a velocity of *1m/s*, she designs artificial stimuli that trigger a propulsion of the frog's tongue and varies systematically parameters of the movement (velocity, direction, distance) during presentation. She measures the precision of the frog's tongue propulsion (e.g. the hit rate). For corroborating her hypothesis, she has not only to show that there is an optimum for 1m/s, but also that it is actually the velocity and not the direction or distance that determines the optimum. The experiment confirms or falsifies the prediction and, therewith, the hypothesis. This is the basic hypothesis-design-experiment-evaluation cycle (see also 2.2.1).

In practice, evaluation (confirmation and falsification) implies complicated procedures. The experimental method provides data. Since a single data item (e.g. one hit) can be pure chance further data items are collected. Comparison of these data items usually reveals certain differences (variance effects). Since these differences may jeopardize the validity of the statement to be made, many data items are collected that quantify the differences. Unless it is shown that variance effects are not the result of an additional causal relation (that was supposed to be ruled out by design) the data will not yield reliable results. For example, if it turned out that the hit rates of the frog's tongue are additionally systematically dependent on temperature (and the temperature was not controlled in the experiments), temperature contributes to variance and the results are less reliable than results that would be obtained with a controlled temperature. In order to assess reliability, data items are statistically analyzed. The analysis finally yields the confirmation or falsification of the hypothesis. Applying the knowledge gained in this experiment, subsequent studies can inform about the correct causal relation (in the case of falsification) or reveal further causal relations (in the case of confirmation). The result of the experiment is an evidence, i.e. a very specific causal relation that was tested with a very specific experimental procedure. The evidence has to be evaluated conceptually, i.e. interpreted in terms of its significance for a broader context (contextual, functional explanation).

It should be noted that the experiment described above referred to a single causal relation as opposed to a mechanism that usually implies multiple causal relations acting on multiple levels of organization (see also 1.1.5). (In the frog's tongue example, only the stimulus parameters were used to characterize the overall performance of the system. Imagine the neural levels that have to be taken into account for a mechanistic analysis…) Since each causal relation can influence any other casual relation present in the mechanism, the expense necessary to reveal a single causal relation in an experimental procedure would be intensified dramatically if a complete mechanism is considered. Each interaction has to be testable (statistically) with data gathered under a separate condition where only one factor is varied, while all others are kept constant (*ceteris paribus*).

Consider the example of the three stimulus qualities velocity, direction, distance being a, b, c ('¬' means logical 'not' and 'v' means logical 'and'). With only one factor a single condition is needed (control condition omitted), for two factors at least two conditions are needed: (a v ¬b), (¬a v b). For three factors, the single factors have to be isolated: (a v ¬b v ¬c), (¬a v b v ¬c), (¬a v ¬b v c). But also combinations have to be tested in order to determine the type of interaction: (a v b v ¬c), (¬a v b v c), (a v ¬b v c), (a v b v c). These constraints of experimental work are present in any science and are no major problem for experimentally well feasible systems. But in brain studies, the combinatorial effects of multiple factors constrain the operability of experiments substantially: Frequently, the experimental expense in brain studies for examining a single causal relation is pushing the limit and most of the time it is hard enough to isolate that very one causal relation in order to collect sufficient data for a neat statistical analysis. In the frog's tongue example, a neural analysis could involve electrophysiological recordings that imply a technically highly demanding experimental setup. Instead of explaining this demand in detail, just some time ranges shall be provided for illustration: developing and tuning the setup might take months, obtaining data of a single cell for a single condition might take a day or more. Each condition should include, say 20 trials. For a reliable statistical analysis minimally five cells of the same type in different frogs have to be tested. Taken together, examining three factors systematically could take several months. These are serious practical (and economical) constraints for a brain study. But even worse cases can occur: the experiment can fail altogether because every propulsion of the frog's tongue causes vibrations in the frog that cause the microelectrodes to slip out of the cell the activity of which was to be measured. Thus, for lack of ways isolating a causal relation or controlling the mechanism adequately or for lack of an appropriate technique, testing a hypothesis experimentally sometimes seems impossible.

1.1.10. Control

Experimental feasibility is a chronic limiting factor of brain studies. This circumstance results from problems controlling either the experimental techniques or the experimental preparation (e.g. the brain), taken together the experimental setup. These control problems arise from the general sensitivity of the setup that causes variance effects in the data. Where to draw the line between experimental techniques and experimental preparation? As already explicated above, the position of a microelectrode (depth, angle etc.) may vary over experimental trials and influences recording quality that might eventually determine measured signal amplitudes. Thus, sensitivity of the experimental techniques can manifest itself as artifacts in the data. Unless such variance effects are excluded, uncertainty about the validity of the data will persist. Therefore brain scientists have to monitor and quantify these effects. Control problems are introduced also by the experimental preparation. The sensitivity of the experimental techniques has its correspondence in the sensitivity of the preparation. For example, electric signals in brains are usually small (low energy consumption is usually advantageous for biological systems) and, therefore, always prone to interferences: 'noise' is a permanent troublemaker for experimenters in the brain sciences. Moreover, the more complex (and the more dynamic...) the underlying mechanism is, the more sensitive to variance effects it becomes and the more difficult it is to design the experiment and its procedures so that it isolates a single causal relation (see above). The reason is, again, the complicated task of predicting outcomes of multiple causal interactions that might be the reasons for variance effects.

1.1.11. Simulation

The methodology of brain studies is afflicted with control problems. Simulation (in the following paragraphs only computer simulation is addressed) is integrated in the methodology of brain studies in order to solve or handle control problems. How is this conceivable? Simulations are an “imitative representation of the functioning of one system or process by means of the functioning of another” (Merriam-Webster 2003). Put in slightly other words, simulations are an inherently functional representation (“a running model”) of a mechanism. As already explained, a mechanism is considered here as a specific configuration of causal relations between elements and activities that can be studied in an experiment. Simulations usually shall help to solve or handle problems of both the experimental techniques and experimental design. In the example of the frog’s tongue propulsion a scientist might want to simulate a model of the neural circuit that determines the spatiotemporal coordinates of the tongue tip because the scientist wants to test stimuli for the experimental design in the simulation before the much more costly electrophysiological experiments with the natural preparation (the frog) can be performed. A simulation represents the strategy to operate a mechanism in a way that informs the scientist evidently about a causal relation by revealing an observable (measurable) behavior of constituent elements (see also 1.1.7).

The reasons for carrying out simulations are to be found in the desiderative understanding of the mechanism that demands a ‘better’ (more controllable) form of representation. Since it mimics the natural preparation in the experimental setup, it can be called an artificial preparation. For this artificial preparation to work, the simulation must contain the elements and causal relations that are considered relevant for the mechanism. But contrary to the experimental setup, the simulation is substantially reduced. Only those elements and causal relations are considered that are assumed to be indispensable for the mechanism to work. This reduction of complexity allows a systematic testing of single causal relations, while additional ‘distracting’ causal relations can be ruled out. This systematic testing is often not possible in the experimental setup since the elements and casual relations cannot be arbitrarily knocked out in brain preparations. Thus, reduction of complexity is one of the main strengths of simulations as opposed to experimental setups with natural preparations.

However, not only the mechanism itself is reduced, but also the experimental techniques and procedures that are applied to operate the mechanism. The control problem introduced by the experimental technique has a simple solution: all measuring techniques are substituted by simulation techniques and the natural preparation is substituted by the simulation. In this way, the experimental situation is streamlined towards a direct confrontation of scientist and mechanism. Cognitive experimental procedures supersede physical (hardware) experimental procedures. Whether this substitution effectively pays off, requires a case-to-case evaluation since computers and software themselves introduce new problems and novelty effects that could render the whole venture of simulation uneconomic.

Simulations allow systematic testing of the effects of a single causal relation. But simulation is no end in itself. Simulation happens against the background of solving and handling control problems that come up when the effects of a given causal relation cannot (or only under disproportional great expenses) be evaluated in an experiment with a natural preparation. Simulation is a kind of hypothesis testing. Within the hypothesis-design-prediction-testing cycle (see also 2.2.1) simulations serve several functions: they allow for a better evaluation of the hypothesis, i.e. an assessment of the role a single causal relation plays inside a mechanism (e.g. velocity sensitivity of the neural circuit determining the spatiotemporal coordinates of the frog's tongue propulsion). This is particularly important for a hypothesis that implies multiple, mutually interdependent causal relations (e.g. isolating velocity from direction). By eliciting a formerly not considered crucial factor (e.g. angular velocity), simulations can help to find new experimental designs, i.e. innovative ways of driving the underlying mechanism adequately so that it unveils its principles. On this view, the function of simulation in brain studies is primarily to improve the scientist's understanding of a given brain phenomenon that was formerly locked in complexity, i.e. not accessible for the different reasons mentioned above. An improved understanding, then, yields improved explanations, yields new hypotheses and new experimental designs.

In brain studies, simulations play an important role in explaining complex and dynamic brain phenomena that are hardly accessible in experimental setups with natural (biological) preparations. But the function of simulations goes beyond this role as a scientific tool. The computer programs simulations are based on can be applied in several other contexts. A prominent example is their application in robots: typically, simulations control the behavior in the robot, e.g. as a brain surrogate or a local control structure. Sometimes it is even argued that pure simulations (without hardware implementation) do not generate proper knowledge about brains because the constraints of the real (physical, mechanical) world cannot work in the simulation. Another field of application is found in bionics. Here, simulations that control signal processing are implemented in chips that are to be implanted in damaged organs such as the retina or the parts of the ear. Very successful, but more remote from the brain are applications originating from the field of mathematics and computer science. This research tradition, usually termed artificial neural networks or Neuroinformatics, has developed a method of computation that is inspired by brain function in that it makes use of formal 'neurons' that process information in a parallel and distributed fashion. This approach is well suited for setting up simulations of brain phenomena. However, a largely independent research tradition emerged because it was found that this approach to modeling was suitable to solve principally any formal problem and, additionally has specific strengths: beside their forgiving character (fault tolerance), artificial neural networks are primarily characterized by their ability to learn in which they outperform most other simulation approaches. Thus, it is no wonder that artificial neural networks are applied not only theoretically in science but also practically on a large scale in various contexts such as fingerprint detection or stock prediction.

1.1.12. Summary and conclusion

The brain, as a subject of explanations, serves as a kind of metaphorical center for all the functionality that is needed to understand adaptive behavior of animals. The analysis of brains reveals a variety of phenomena ranging from anatomic, electric and molecular to behavioral and environmental. Most of the phenomena belong to the neural domain, but some are better classified to reside in a computational domain or a cognitive domain. The organization behind the brain phenomena is determined by spatial, temporal and causal characteristics. It is in principle hierarchical so that phenomena on a higher level are thought to be dependent on phenomena on a lower level. Though principally continuous, the organization is frequently partitioned into levels that represent functional units themselves organized by specific causal contexts (e.g. electric, molecular, cellular etc.) and, consequently, referred to as the “levels of organization”. The spatial, temporal and causal interrelation between elements and activities in the organization is characterized by dynamics and complexity. An explanatory strategy that can be applied for the levels of organization and the inherent dynamics and complexity are mechanisms. Mechanisms ‘put to stage’ elements, activities and causal relations participating in the phenomenon in a directed manner and produce consistent explanations. Explanations can be in a mechanistic, functional or reductive mode depending on whether they seek to explain a phenomenon in isolation (level 0), in a context (level +1) or constitutively (level -1), respectively. The methods of producing evidence are manifold. Due to the complex and dynamic character in natural preparations of the brain, experiments are afflicted with control problems. As a specific methodological approach, simulations with artificial preparations yield possibilities for increased control of dynamics and complexity.

Reading Advice

The previous section introduced generally how brains are explained and provided a first idea of the role simulations plays in explaining brains: namely helping to control dynamics and complexity. For the (hurried) brain expert this might be enough to proceed directly with the second chapter on simulation. Other reader groups might have received only a rather abstract impression of what it means to explain brains and why we need simulation. Therefore, the next two sections are "Brain-Specials" that extend the presentation of brains – the next presuming a consensus on explanations of brains and the next but one starting with the variety of approaches to brains. In the next section, the first Brain-Special "Explanations of brain phenomena" provides concrete explanations of brain phenomena moving from low levels to high levels of organization. The focus will be on explanations on lower (cellular) levels because the consensus is by far greater than on higher levels. (Brain experts will therefore rather repeat than deepen their knowledge. They can, however, deepen their 'meta-cognition' on brains by focusing on explanatory strategies.) In the next but one section, the second Brain-Special "Where to put information on brains!?" starts with the variety of disciplinary approaches to brains and searches the commonalities between them. This is done by analyzing existing information resources such as databases, thesauri and textbooks in order to find a common 'ontology'. As a result, the state of theoretical integration in the brain sciences can be assessed.

1.2. Explanations of brain phenomena

1.2.1. Introduction

Brain studies seek to explain how brain phenomena realize adaptive behavior (see also 1.1.1). Ultimately, they explain a behavior of an organism in its environment, e.g. how a fly prevents collision with a frog's tongue (see fig. 1). Now, how exactly are explanations of brains designed? By applying a general explanatory framework (see also 1.1) to concrete brain phenomena such as 'spike', 'synapse' etc. specific explanations of brain phenomena are developed. The objective is not to explain each phenomenon thoroughly – rather each phenomenon introduces building blocks for explanations of brains that are to be applied in subsequent explanations of more complex phenomena. Thus, it will be demonstrated that explanatory building blocks can be applied as if there were an "explanatory construction kit" for brains. Since lower level phenomena are likely to be contained by higher level phenomena and, thus, are often presupposed in higher level phenomena, the focus will clearly be set to electric, molecular and cellular phenomena. Higher-level phenomena such as behavior, learning etc. will be discussed only briefly along the lines of exemplary cases.

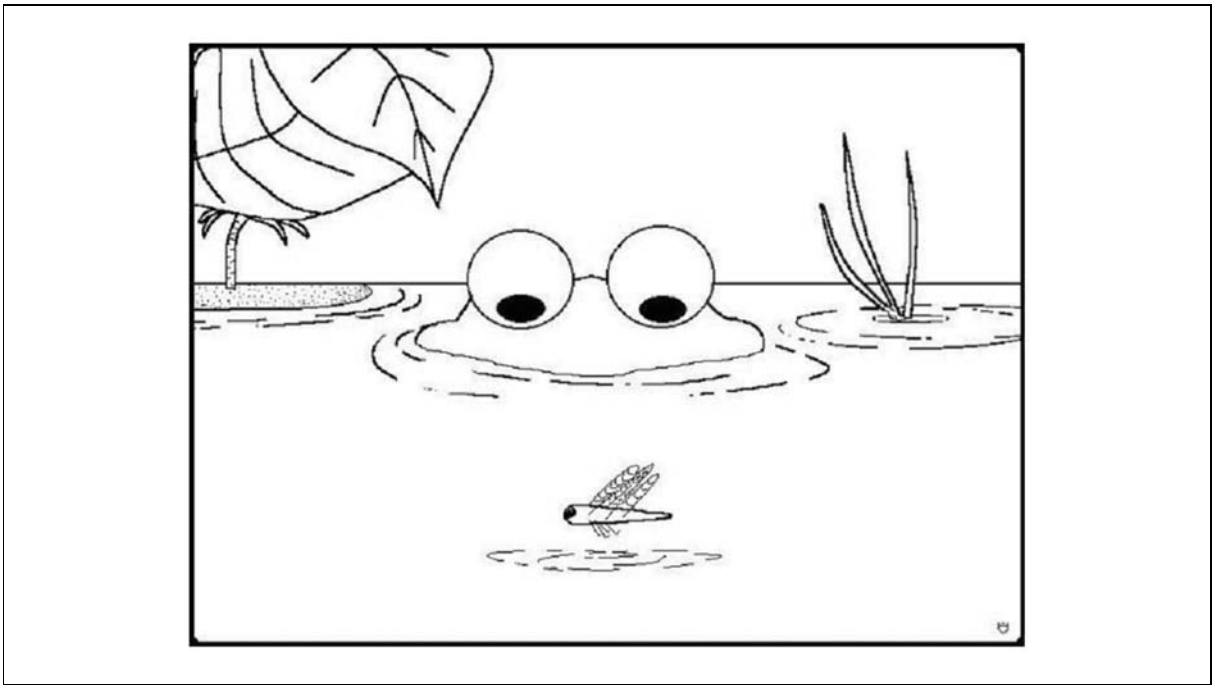


fig. 1: Frog and fly. © 2003 State of California (license permissive for reprint).

1.2.2. Responsiveness

Consider a fly flying over a pond. A frog sitting on the pond's edge detects the fly coming within reach of its tongue. The critical situation for the fly is given as the frog just flicks its tongue. What does the fly need to manage a successful escape maneuver? It needs receptors that can detect visual stimuli, a brain that can process the detection and compute the correct parameters for a behavior. Finally it needs a motor system that can realize the behavior. Nerve cells are able to represent specific aspects of the situation and process these representations in order to find the correct parameters of behavior. The ability of neurons to represent and process something is based on their responsiveness. The property of showing responses usually refers to the neuron as a whole and is thought to be the result of mechanisms acting on the neuronal membrane. (In the following, it will usually be attributed to the neuron – a neuronal state can equally well be attributed to a patch of membrane or to a neuronal circuit, though).

Responsiveness is such a basic property of neurons that it is easily passed over in silence. Therefore it shall be explained explicitly: a typical experimental setup for the observation of responsiveness is a neuron penetrated by a microelectrode that measures electric changes (potentials or currents) monitored on an oscilloscope (see fig. 2). Given that the experimental setup is tuned appropriately, the observations will be restricted to a straight line that represents the electric state of the neuron. The straight line as such, obviously, is no evidence for the ability to represent. For something to be observed that evidently reveals responsiveness, changes must be imposed. In a natural situation (i.e. the frog an the fly), these can be sensory stimuli (i.e. the frog's tongue). But here, for the sake of simplicity, assume that electric pulses are applied through a second, stimulating microelectrode. A change in the electric activity of the neuron should be observed as a deviation from a straight line. This effect of an applied pulse reveals evidence for the ability of the neuron to respond to externally imposed changes. Of course, this property is so basic that it can also be found in a liver cell, for instance, since most animal cells also show this simple form of electric activity. (The more advanced characteristics of neurons will be introduced later.) Nevertheless, this simple example indicates that the principle of representation by responsiveness usually presumes a specific change in the environment of the neuron (an electric pulse or the frog's tongue) that can be detected and represented by the neuron by changing from one state of activation to another state of activation.

Additionally, something to be activated generally implies that an element has to undergo a transition from one state to another state. Activation implies a notion of change, of before and after or, more generally, of cause and effect. The ability to represent refers to the property of the neuron to show an activity that corresponds (causally, temporally and spatially) to another activity. Of course, this description is so abstract that it hardly bears any meaning. But it should be noted that it enables the neuron to be conceived as an information or signal processor, as a representative unit. It is generally assumed in brain studies that a neuron is able to 'take up' changes of its environment by changing its state of activation and 'hold' these changes in that state, i.e. it is assumed that an external affair is represented as long as activation is present. If the activity of the neuron (or of a part of the neuronal membrane) in turn causes changes at a spatiotemporal neighbored neuron (or part of the membrane) the change is additionally *processed*. In a broader sense, changes in the environment (e.g. about a frog's tongue) can be transported through the nervous system (e.g. the fly's brain), can be computed and evaluated to generate an adequate behavior (e.g. preventing a collision).

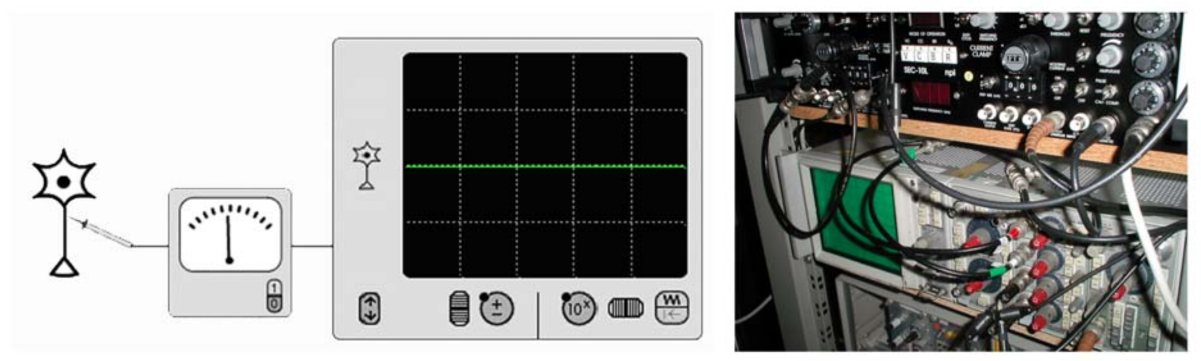


*fig. 2: **Experimental setups.** On the left side, a schematic representation of a simple experimental situation is shown. From left to right a neuron's electric activity is registered with a microelectrode, processed through metrology and visualized on an oscilloscope. © 2000 rubin, (permitted reprint). The photograph of metrology used in brain studies on the right side illustrates that laboratory setups are usually much more complicated than suggested in the scheme.*

1.2.2.1. *Binary neuron*

In the following, the notion of responsiveness is developed from very abstract descriptions to concrete, neuroscientific descriptions. An abstract notion of a neuron as a representational unit becomes possible just by assuming that neurons show responsiveness – and not by assuming underlying mechanisms. It should be noted that pure 'computationalists' (proponents of a computational approach) who do not bother about biological realization could enter at this point with their explanations of brain phenomena. Computationalists merely need a function to explain brain phenomena. The most simple form of describing responsiveness is a binary neuron, i.e. a neuron *n* switches between the activity states *0* and *1* depending on an activity of stimulus *s: n(s); {0/1}*.

1.2.2.2. *Electric activity*

Of course, neuroscientists seek to explain brain phenomena by taking into account the underlying mechanisms. So, what are underlying mechanisms of responsiveness? A common explanatory strategy coming into place is one that focuses on the smallest level of organization introduced above, namely the electric level. The neuron is conceived as an electric circuit. For describing the simple form of electric activity, as it was observable in the simple experimental setup, the circuit consists of a resistance and a capacitor (see fig. 3a). The neuron's state of activity is represented by the membrane potential *V*. As long as nothing else happens it stays constant. (This explains the straight line observable on the oscilloscope.) This property of the neuron can be conceived as a capacitor because it has the general ability (*capacitance C*) to separate positive and negative charges.

If charged particles were added by some electric source (be it a microelectrode fed by a power supply or a biological antiporter such as the sodium-potassium pump) to either side of the membrane, the charge is not immediately leveled out, but persists on that side of the membrane. Consequently, the inside is charged to an amount of *Q* relative to the outside of the neuron. The capacitor is charged. But the capacitive properties of the neuron are not perfect. The charge can leave the capacitor as a current of charges *I* – but only to the degree the neuron can conduct charges. The conductance *g* is given by the reverse of the resistance *R* of the neuron *g=1/R*. The resistance *R* hinders the current to flow straight from one side to the other. This is why discharging the capacitor does not happen instantaneously but delayed. The moving charges change the overall state of the neuron: the membrane potential *V*. As long as the capacitor discharges, the membrane potential changes. When the capacitor is discharged the 'dislocated' charges return to the other side in order to restore the electric equilibrium: the membrane potential returns to its initial value. These processes explain the delayed deviation from a straight line as a response to a current pulse. In this way, it is possible to explain responsiveness in an electronic circuit or, more concrete, in an electronic device consisting of building blocks such as resistors and power sources in a construction kit for kids. Up to this point, no further assumptions about underlying mechanisms have been made. The pure mathematical framework *n(s); {0|1}* that was applied in the case of functional description is merely exchanged with another framework, namely electricity. Again, the neuron as such is not taken into account. As might already have become clear from the very abstract character of the description, the electric framework raises concrete conceptual questions: What is resistance, what is current etc.?

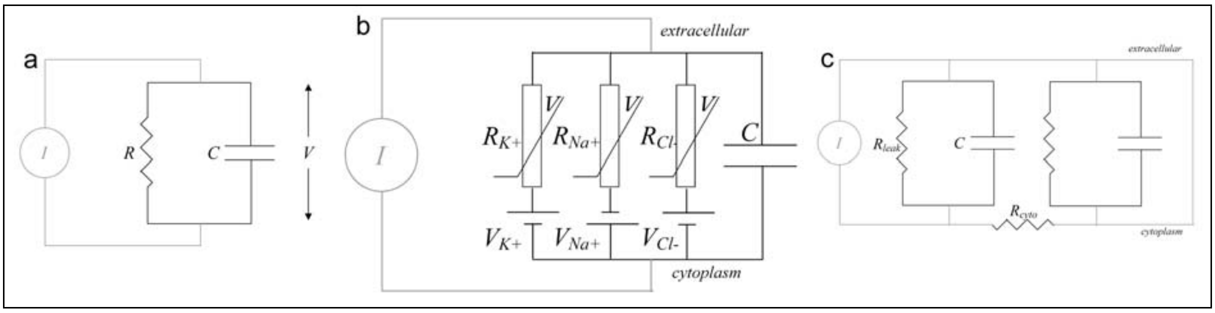


fig. 3: RC-circuits. (a) Simple RC-Circuit. A hypothetical power supply provides current I. *The potential* V *over the circuit changes as the current pulse* I *charges capacitance* C *and flows through resistance* R*. This electrical equivalent circuit is a simple model for electric phenomena at the neural membrane. (b) RC-circuit for spikes. Elaboration of the general form of the RC-circuit that can be used to explain spikes on the electric level of organization. The situation for each ion (potassium, sodium and chloride) is shown as a combination of a battery that corresponds to the electric gradient and a variable resistor (varistor) that corresponds to voltage dependent channels, respectively. (c) RC-circuit for passive spread of electric activity. Variation of the general form of the RC-Circuit that can be used to explain passive spread of electric activity. Form one point in space to another point. A new resistance R_{cyto} is introduced to account for the cytoplasmatic resistance that has to be overcome by electric activity in the lateral direction (along the membrane).*

A common further step in the explanatory strategy is to provide exactly these concepts. Consider the neuron as one side of a container filled with water that is separated by a membrane from the other side of the container ('the external world') that is also filled with water (see fig. 4). Now, soluble substances (salts) are put into the water. (As the salt dissolves in the water, it separates in positive and negative counterparts (ions). For the moment, it is only important to consider the net amounts of the substance – the solution is electrically neutral.) If the amount of these substances differs on either side, a difference of concentration results in the solution. This difference in concentration represents a chemical potential. Since the concentration differences tend to level out ('diffusion' resulting from the different electrostatic interactions of water and substances on both sides), the substances would tend to move towards the lesser-concentrated side, if the membrane were permeable for the substances. If the membrane were impermeable for the substance (but permeable for water), a chemical potential would arise that would cause water to pour through the membrane and level out the concentration difference (diffusion and osmosis). Now, consider the membrane to be permeable only for positive elements (ions). Since the concentration difference can be partly leveled out by positive elements, they move to the side of lesser concentration leaving their negative counterparts behind. The membrane begins to charge. At this point, an electric potential arises. Now, two forces act against each other: a chemical force pressing positive ions outward and an electric force dragging them to the inside. The positive ions will stop moving through the membrane when electric and chemical forces are equal – equilibrium potential is reached. This explains why the resting potential (the straight line observed on the oscilloscope in the simple experimental setup introduced above) is actually different from zero (negative). If there is only one substance involved, the membrane potential is the same as the equilibrium potential for that substance and is given as

$$E = (-R \cdot T \,/\, z \cdot F) \cdot \ln\, [SUBSTANCE_{outside}] \,/\, [SUBSTANCE_{inside}]$$

This is the Nernst equation that relates the electric forces to the chemical forces. R and F are physical constants (the Molar Gas Constant $8.314472\, J / K \cdot mol$ and the Faraday Constant $9.64853415 \cdot 10^4\, C / mol$, respectively). T is the absolute temperate (measured in degrees Kelvin) and z is the valence of the ions, i.e. if it is negative or positive and if it bears one or more charges. The Nernst equation was originally determined in experiments with batteries (Nernst 1888).

The container situation allows to develop a simple electrochemical account of the premises of responsiveness with an electrochemical construction kit for kids. But the simple electrochemistry introduced so far has to be completed by a description that represents the situation in natural neurons more adequately: not only one substance (one positive and one negative ion species) determines chemical forces, but more ions are involved. Consider a second container with another dissolved substance (another positive and the same negative ion species) with another ion specific equilibrium potential described by the Nernst equation. Now, pour the second container into the first container. The resulting potential depends on the equilibrium potentials of all ions. Ions may pass the membrane not equally easy – there is an ion specific permeability. Permeability has not been considered in the Nernst equation because only one ion species is taken into account. But if two or more differential permeabilities act in parallel, all must be considered because they affect the ion specific forces together with the corresponding equilibrium potential. Often, three ion species are regarded: the ions potassium K^+ (the positive ion species of the first container), sodium Na^+ (the positive ion species of the second container) and chloride Cl^- (the negative ion species in both containers). The overall state of the membrane determined by the ion specific permeabilities (P_{K+}, P_{Na+}, ...). The influence of chloride is small relative to those of sodium and potassium.

The same the Nernst equation did for the single substance situation – relating electric forces to chemical concentrations – does the so called Goldman-Hodgkin-Katz Constant Field equation, in short “Goldman equation” for the multiple ions situation.

$$E = -RT / zF \cdot (P_{K+} \cdot [K^+]_{out} + P_{Na+} \cdot [Na^+]_{out} + P_{Cl-} \cdot [Cl^-]_{out}) / (P_{K+} \cdot [K^+]_{in} + P_{Na+} \cdot [Na^+]_{in} + P_{Cl-} [Cl^-]_{in})$$

The Goldman equation considers more ion species than the Nernst equation. The potential depends on the ion specific permeability of the membrane and the ion specific electrochemical forces that are set up by the differential concentrations. The permeability defines the factor by which the electrochemical force of a given ion species influences the overall potential.

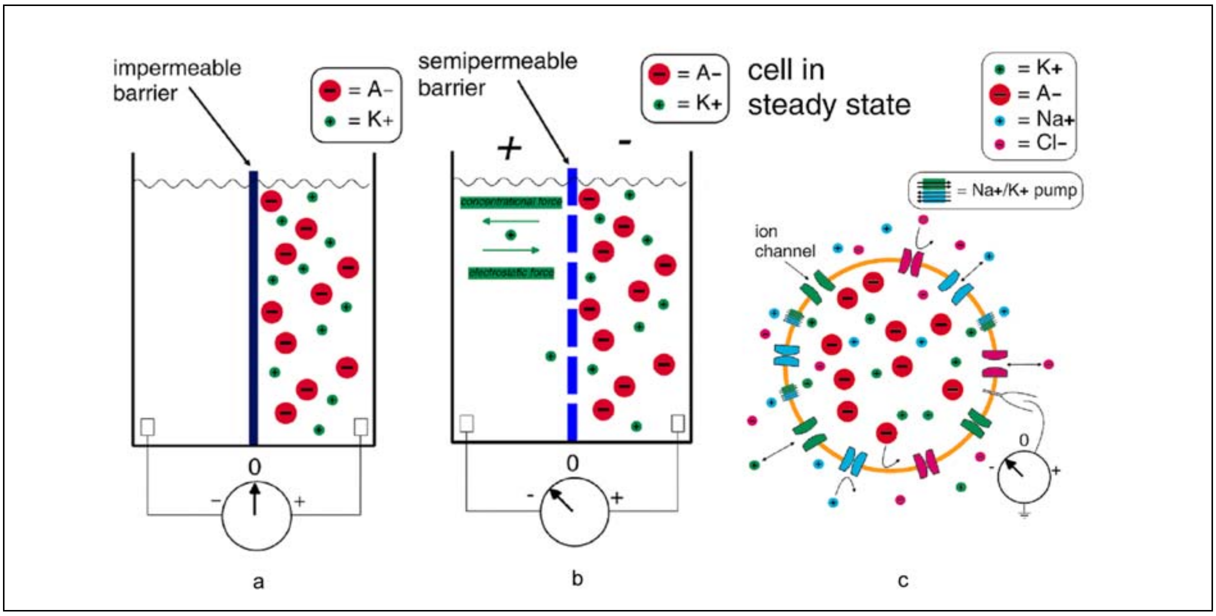


*fig. 4: **"The container situation"**. Typical approach to explaining the resting potential. **Left**: As long as the outside of the cell (left side of the container) is separated from the inside filled with equal amounts of negative ions A^- and positive ions K^+ (right side of the container) by an impermeable barrier, no charges can move and no potential is observable. **Middle**: When a semipermeable barrier is present, positive ions can move to the outside leaving their negative counterparts behind: a negative potential from inside to outside is observable. With the moving ions, a chemical potential from outside to inside arises. When electric and chemical potentials are the same, a dynamic equilibrium appears: the resting potential. **Right**: The resting potential in a model cell. More ion species and ion specific channels types are shown. © Shizgal & Oda csbn.concordia.ca (permissive license).*

1.2.2.3. *Molecules and electric activity*

The result of the explanation so far is a more detailed description of the resting potential. Step by step, a deeper understanding of the neuron as an electrochemical device was introduced. It should be noted that the electric level of organization was not yet left behind for entering the molecular level. (Of course, the transition between the electric and the molecular level is continuous rather than discrete: experts might have noticed, for example, that genuine molecules – e.g. proteins as opposed to atomic ions – might also carry charges. Anyhow, their role then is electric, i.e. that of a charge carrier – not as a molecule with its specific properties that are taken into account later.) The molecular level hosts explanations of major concepts that were only introduced as presupposed constraints so far. An instructive example of such a supposition is the selectively permeable membrane. What is its molecular basis? Generally, a cellular membrane is a phospholipid-bilayer (see fig. 5). The 'heads' of the phospholipids (phosphate) are on the outside of the membrane, i.e. one layer of heads points to the intracellular side, while the other layer of heads point to the extracellular space. The tails build the inner space of the membrane that is apolar and, thus, only permeable for apolar (e.g. non-charged) substances but impermeable for polar substances, i.e. polar molecules such as water and charged substances such as ions. This explains the membrane's ability to separate charges, i.e. its capacitance.

But what enables selective permeability? A neuronal membrane is equipped with channels. Basically, channels are large proteins that pervade the membrane. Channels permit specific ions to pass the membrane, while they block other ions. One type of channel is particularly important: the 'passive' channel that is always open. It is assumed to be selectively permeable for potassium. A simple explanation of this selectivity (a more precise explanation is beyond the scope) refers to the size of potassium. In a solution, potassium is smaller than sodium because it has smaller electrostatic powers ('electro-negativity') than sodium. The surrounding water molecules are not 'geared' in the same radius as a sodium ion, resulting in a smaller 'hydrate sheath'. Thus, the smaller solved potassium ions can pass a channel that sodium can not pass. Since potassium is the smallest positive hydrated ion with significant concentration around the membrane and is considered to pass the passive channel, it plays the crucial role in generating and maintaining the resting potential. For understanding this crucial role, the concrete concentrations at the membrane have to be taken into account. The concentrations of potassium inside the cell are high (relative to the outside). The concentration of sodium is high outside the cell (relative to the inside). The common explanation for this initial situation on the neuronal membrane is a molecular mechanism called "sodium-potassium pump". This special type of channel is thought to exchange three sodium ions from the inside with two potassium ions from the outside (see fig. 4, 5b, 5c). This works against the electric and chemical forces (potential and concentration gradient, respectively) and is realized as an active, energy-consuming process. In exactly the same way as introduced in the container situation, potassium tends to come out and sodium tends to leak in. The potassium selective channel admits a small potassium outward current, thereby leaving an excess of negative ions behind. The intracellular space begins to charge negative relative to the extracellular space. The electric forces get stronger as the chemical forces get weaker and equilibrium turns up.

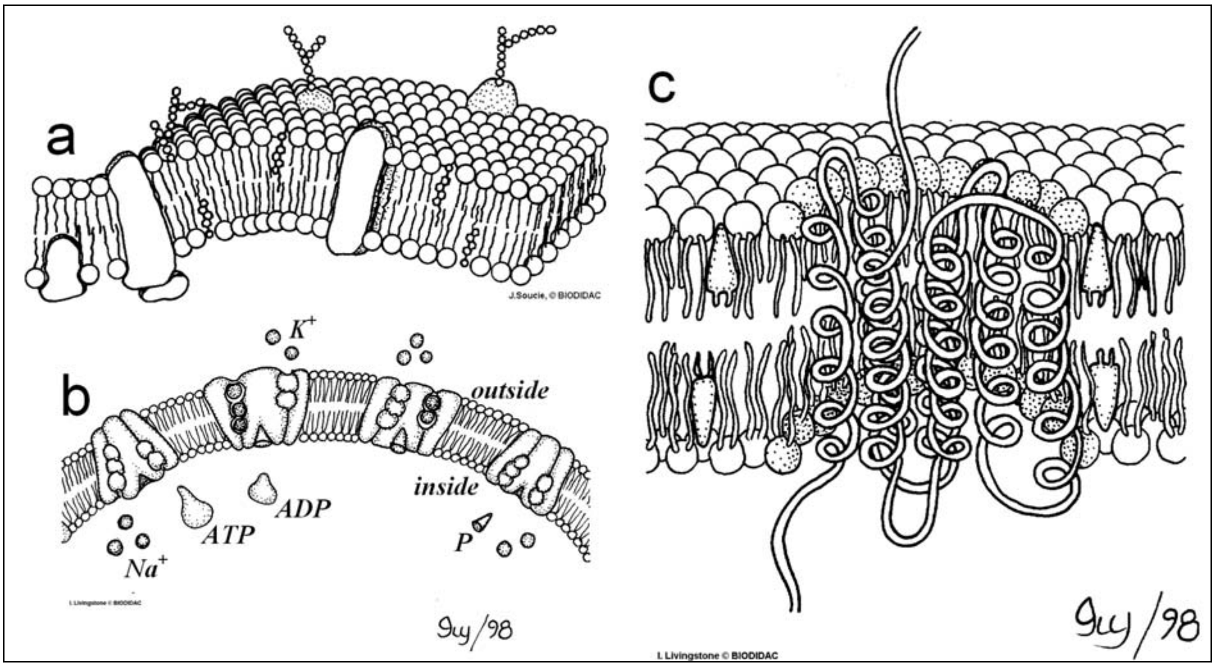


***fig. 5: Molecular aspects of membranes.** **a)** Fluid mosaic model. Phospholipid molecules consisting of hydrophilic phosphate part (small circles, outward) and hydrophobic lipid part (small lines, inward) constitute the phospholipid bi-layer. Proteins (larger 'chunks') are sunk in the membrane (e.g. receptors) or go through the membrane (e.g. channels). **b)** Sodium-potassium pump. The mechanism is illustrated as four stages from left to right: three sodium ions from the inside of the cell and a phosphate (coming from ATP -> ADP+P) bind to the protein. The conformation of the protein changes and the sodium ions are released to the extracellular space while two potassium ions bind on other receptive structures of the protein. The conformation changes again and phosphate as well as potassium is released to the intracellular space. **c)** Alpha-helical channel protein. For illustrative purposes it is shown that channel 'chunks' can also be presented as structured proteins that could be further decomposed into amino acids, molecules etc. (not shown).* © BIODIDAC *(license permissive for reprint).*

Channels explain why it is particularly important to consider potassium in explanations of general responsiveness because they provide an explanation for selective permeability (for potassium). Channels also explain the resistance of the membrane: as already stated above, resistance is the inverse of conductance, $R{\sim}1/g$. A given conductance is determined by the relative frequency of channels in the membrane and their efficacy. The more channels per area membrane and the more efficient the channels the more conductive it is. Current can be conceived as the actual amount of ions passing the membrane, thereby causing a change in membrane potential.

In sum, introducing the molecular level contributes significant explanatory connections between the neuron and its electrochemical properties. In particular, it provides a spatial notion, a 'stage', for the electric level of organization. The electric circuit and the container can be supplemented by biological objects that are made up of molecular objects such as cells, membranes and channels. For assessing the explanatory strategy that was applied, it should be noted that the molecular level of organization has an intermediate position in the spatiotemporal dimensions that were introduced before. It provides spatial building blocks for the neuron and provides positions and trajectories for charges. However, considering the molecular level is evidently not necessary to understand the responsiveness of neurons. Nevertheless, molecular dynamics (MD) models (see e.g. Leach 2001 for a textbook), for example, that allow one to imagine conformational changes in the structure, and movement of ions across the channel, can surely be a valuable means for deepening the understanding of responsiveness.

Generally, it can be stated that responsiveness of neurons can be understood in terms of neurons, membranes, channels, ions, resistances, conductances, currents etc. The explanations so far were predominantly simple with respect to the information processing capabilities of neurons: responsiveness is the property of neurons to change their electric state if a stimulus is present: "no stimulus=straight line" and "stimulus=change"! Such a binary understanding of responsiveness is sufficient to represent the information processing capabilities considered so far. In a broad band of phenomena to be understood this notion of responsiveness is actually all that is needed. But what about more detailed observations?

1.2.2.4. *Temporal change and responsiveness*

The representational properties of neurons taken into account so far were restricted to changes from active to inactive. But the example of the RC-circuit (see also 1.2.2.2) already implied that neurons do not show a binary behavior, but show intermediate states (transitions). Reconsider the simple experimental situation: the membrane potential of a neuron is measured via a microelectrode and monitored on an oscilloscope (see fig. 2). Current pulses can be applied with a stimulating electrode. If temporal resolution of stimulus and response is sufficiently high in the oscilloscope, it will become visible that the response of the neuron is not instantaneous, but changes gradually, i.e. the current pulse is needle-like, but the membrane potential follows the upward stroke gradually. How can this time-course be explained?

In the general introduction of responsiveness, the following explanatory strategy was applied: the observation (the signal) was first described and then explained on increasing levels of organization – from electric to molecular… This procedure shall also be applied here: the phenomenological description (gradual change) is first to be explained in terms of on an electric circuit (without reference to the molecular level).

If a current pulse is fed into simple Resistor-Capacitor (RC) circuit (see fig. 3a) through a power supply (the stimulating microelectrode), capacitive current causes an almost instantaneous charging of the circuit: the withdrawal of electrons (positive pulse) or the addition of electrons (negative pulse) causes an electric field that orients the elements of the isolating layer of the capacitor due to their polarization (positive vs. negative) and lets elements in the conducting material around the capacitor arrange correspondingly. It should be noted that no charges (e.g. electrons) have to move in the RC-Circuit to bring about this situation. Subsequently, electrons move along the electric field building a slow current that drains through the resistor. The electric field force is reduced. The overall current I can be conceived as the sum of both current types, the capacitive current I_C and the resistive current I_R: $I = I_C + I_R$.

The principles of this explanation can then be applied straightforward to the molecular level of organization (thereby considering the intermediate chemical level.) The excess of charges introduced by the current pulse causes an electric field that spreads almost immediately through the extracellular and intracellular space. Charges are organized along the electric field and tend to level out the electric field gradient. But the isolating membrane stands in the way for charges to move and level out the electric gradient. This resistance can be overcome by potassium ions passing through the passive channels in the membrane. These ions build the resistive current. Only few ions relative to the total amount of ions in the solution are necessary to level out the electric field gradient between both sides of the membrane. Thus, the concentrations of the ions are not effectively changed by current pulses and can be easily leveled out by the sodium-potassium pump. The dynamics of the membrane potential is determined by the resistive current because it moves charges in the system – the capacitive current does not move charges across the membrane, but rather can be conceived as a reorganization of the actual charge budget. This reorganization does not take effect on the potential.

Both capacitance and resistance of the membrane determine what happens to the injected current: the capacitor transforms charges electric field gradients and the resistance hinders charges to move through (or along) the membrane. Therefore both factors prolong the duration of the membrane potential after a current pulse and, thus, determine the time constant $\tau = R \cdot C$. This time constant determines the time course of the membrane potential for a given current and resistance: $V_m(t) = I \cdot R\ (1 - e^{-t/\tau})$. The 'input value' I representing the injected current pulse does not equal the 'output value' V_m, but can be reached only to a certain degree that is indicated by the term that is subtracted from *1*. What 'is left' from the input current also depends on the resistance R of the membrane. The 'input value' is approached exponentially (determined by the e in brackets). It becomes larger with progressing time t, but only to the degree that τ admits, i.e. a large τ prolongs the time course of the membrane potential.

Considering the time course of responsiveness mediates an impression of how *gradual change* of membrane potential can be conceived. Furthermore, it provides a deeper understanding of the elements involved in responsiveness (e.g. ion, membrane, channel etc.), particularly because it enforces the definition of causal factors acting on the elements that help to define phases ('charging' vs. 'discharging'). Notably, explaining this time course of electric activity is possible by nearly completely omitting the molecular level of organization, although it might help to 'stage' elements and their causal relations (a membrane might be conceived more illustrative than a RC-circuit). The molecular level of organization will come into place again at other instances of this section.

1.2.2.5. *Conduction: Spatiotemporal change and responsiveness*

In a natural situation such as the one in which the fly attempts to prevent collision with a frog's tongue, a change of membrane potential is thought to function as a representation of the stimulus (e.g. the frog's tongue) that can be processed in the system. For something to be processed, it must be handed over to the next stage of processing. The next 'stage' of processing is reached on the next instance in space and the next instance in time. In other words, neurons were considered so far as being a single point in space – below, neurons will be conceived as having spatial extension.

Considering first the electric level of organization, a spatial extension can be introduced by plugging together two RC-circuits (see fig. 3b). The dynamics of the simple stimulus-response situation can be described in terms of a RC-circuit. The RC-circuit can be understood as a representation of a whole neuron but also as a representation of a patch of membrane. This combination of RC-circuits introduces a new resistance that current has to overcome – not *through* the membrane but *along* the membrane (the 'cable' from one RC-Circuit to the next.). Put in terms of the molecular level of organization, ions have to overcome the resistance of the cytoplasm.

It becomes evident that considering a single RC-circuit as a representation of the membrane treats the membrane as a single point in space – a thing hardly conceivable on the molecular or cellular level of organization. So let's first put things straight for the resistance through the membrane. The resistance through the membrane depends on the permeability of the membrane on a given surface area that can be conceived as the density of passive channels (see also 1.2.2.3) on a slice of the cell compartment. This leak resistance R_{leak} depends on the circumference of the slice $1 / R_{leak} \sim 2 \cdot \pi \cdot r$. The greater the diameter of this slice the smaller the resistance. If x slices are plugged together to a compartment of length x, the resistance through the membrane decreases (more channels=more leaks). The specific term r_{leak} quantitatively describes the decrement in resistance over length $R_{leak} = r_{leak} / x$.

Considering the length of the compartment introduces another new resistance *along* the membrane: the cytoplasmatic resistance. It can be conceived as the resistance that currents have to overcome when they flow along the inside of the neuronal membrane. (The outside is commonly neglected because it is often very large relative to the inside.) The opposite situation to that of the resistance through the membrane is given: the more slices are taken into account, the more resistance turns up ("more cytoplasm"). The specific term r_{cyto} quantitatively describes the increment factor of cytoplasmatic resistance over length $R_{cyto} = r_{cyto} \cdot x$.

Thus, the overall resistance of the compartment increases with length and decreases with diameter. Thick, round compartments have a low resistance, while thin, long compartments of the same volume have a high resistance. Consider a compartment that has the same leak resistance as cytoplasmatic resistance $R_{leak} = R_{cyto}$ and $r_{leak} / x= r_{cyto} \cdot x.$ Then the ratio of the two types of resistance determines the properties of the cell. This ratio is called the space parameter or length constant $\lambda = r_{cyto} / r_{leak}$. The greater the specific membrane resistance (through) and the smaller the internal resistance (along), the greater the length constant. Thus, the length constant describes the ease of a current traveling along a compartment. Consider a hose filled with a fluid (cytoplasm) that receives a shot of water (current pulse): r_{leak} relates (inversely) to the amount of holes (leaks) in the hose and r_{cyto} relates to the thickness of the fluid. The length constant determines how deep the shot of water can penetrate into the hose before ebbing.

Coming back to responsiveness, the amount of water in the hose can be conceived as a response. The residual of an initial response at a given distance x can be determined by $V_{residual}(x) = V_{initial} \cdot e^{-x/\lambda}$. A complete geometrical description of a simple compartmented neuron, including all diameters and specific resistances, provides a model of the neuron in which arbitrary responses imposed on the neuron can be traced spatially and temporally. This approach allowing to relate the dynamics of the neuron to spatial organization is called ‘compartmental modeling’ (see Rall 1989 or Walter 1999 for a textbook).

In the preceding sections, an explanation of a phenomenon (e.g. responsiveness in a neuron) was developed from the lower electric level of organization (refraining from molecular terms) towards the higher, molecular level. The explanations provided in this section on spatiotemporal change of activity applied the same procedure, but referred to molecular aspects from the beginning (as the term ‘*cytoplasmatic* resistance’ already indicates). One reason is that concepts referring to the electric level (e.g. resistance) were directly related to concrete molecular organization of the biological substrate, e.g. the spatial extension of neuronal structures. This mix of levels of organization is typical for brain studies and differs considerably from a pure electric account of responsiveness that treats the neuron as an electric building block. By mixing the levels of organization, the neuron is treated as a element of individual geometry that ‘deserves’ a specific approach of description, e.g. cytoplasmatic resistances.

1.2.2.6. *Beyond responsiveness*

Several further mechanisms contributing to the generation, maintenance and conductance of a given activity can be considered. However, a fairly detailed account of responsiveness has been introduced so far. This account applied several concepts referring to the electric level, but rested on very few assumptions on the neuronal membrane, i.e. phenomena on the molecular level of organization. Merely a phospholipid-bi-layer with a passive channel (supported by an abstract sodium-potassium pump) was introduced. This account is also called the "passive membrane" because it just explains 'passive' responses and does not imply still to be shown 'active' changes of the input signal introduced by further neural mechanisms.

However, if studied experimentally, even a simple passive membrane would reveal phenomena that are not explainable with the mechanisms introduced so far. In every simple experimental situation (see fig. 2), it would be observable – if amplification factor and temporal resolution were maximized – that fluctuations pervade the neural signal. Given that these fluctuations are not artifacts introduced by the experimental setup, they are the result of activities in the neuron not considered so far. Since it is assumed that these activities do not contribute significantly to responsiveness, it is termed 'neuronal noise'. This phenomenon points at the possibility of a more differentiated understanding of responsiveness, even if only a passive membrane is considered. The sources of neuronal noise may be manifold. A prominent and illustrative explanation can be given on the electric level. Consider the differential concentrations of ions on both sides of the membrane as described by the Goldman equation. It is assumed in the equation that the equilibrium potentials for the different ions and, thus, the membrane potential stay constant over time. Even for non-physicists, it must be a strange notion that individual ions 'wait' and stand still all the time on either side of the membrane until that very moment when an electric field potential (generated by an activity) arises and drags it to the other side. It is known from everyday observations from elements in solution (such as sugar in tee) that these are in continuous movement. These movements can be traced back to Brownian motion. Generally the membrane potential is a dynamic equilibrium. Dynamics imply that individual ions change position (and even sides) and the membrane leaks continuously. This causes small fluctuations in the membrane potential that can be conceived as neuronal noise. (Other ‘stochastic’ processes such as those in more complicated, so far unconsidered, channel types are even more influential, but are too complicated to be explained at this instance.)

How can noise be added to the different versions of the explanations of responsiveness introduced in this section? For a phenomenological description of noise, it is possible to add fluctuations in the form of a stochastic (random) function that simply adds a value chosen arbitrarily from a small range of values on the signal at each instant in time. An electrochemical account would add stochastic qualities (probability functions) to the ion concentrations in the Goldman equation. A dynamic Goldman equation would be a complicated system of differential equations. The explanatory value of this detail is of inferior significance for most brain phenomena. (This, again, makes clear that the choice of mechanisms that are to be considered in an explanation of a phenomenon depend crucially on question posed.) For understanding the principles of representation (e.g. of the frog’s tongue) and for understanding simple stimulus response relations, noise is not essential. However, it definitely gains significance in other issues (see e.g. Shadlen & Newsome 1994).

1.2.3. Spikes

Neurons are responsive. They can represent a change of the environment by a change in their membrane potential (an activity). For an activity to serve a function (e.g. representing a frog’s tongue), it has to be embedded in a larger context. The questions are: Where does the activity come from and where does it go to? The question where it comes from shall be faded out for the moment. (For the sake of simplicity it can be assumed that the stimulus is injected in the cell by an electrode in an appropriate experimental setup.) Where it goes to is the question that shall be answered first. For transporting the representation bearing activity to another place, the neighboring membrane patch has to take it over. On the electric level of organization this can be explained in terms of RC-circuits in series (see also 1.2.2.2 and fig. 3b). Unfortunately, the membrane is anything but an ideal conductor for currents; it rather resembles a leaky hose. In the cases considered so far, the activity seeps away in the extracellular space only after a short distance (relative to the overall size of the neuron). On the electric level of organization, several possibilities for improving this situation can be found:

(A) Decreasing the resistance between two RC-elements (resistance along the membrane) so that current passes RC-elements more easily. (B) Increasing the resistance in a single RC-element (resistance through the membrane) so that the leak current is reduced. (C) Amplifying the activity. All of these possibilities are realized in neurons: (A`) Since the resistance along the membrane increases with diameter of the cell compartment, the neuronal structures that shall propagate an activity (axons and dendrites) usually are long and thin cables rather than round and thick spheres. Alternatively, the resistance along the membrane could be reduced by changing the properties of the cytoplasm that, however, is not without reason as it is (it has to serve other functions than propagation of electric signals). (B`) Reducing the resistance through the membrane can be achieved by reducing leak currents and, equivalently, by improving the capacitive properties of the membrane by increasing its isolating properties. This is found in many neurons in the form of an isolating structure around the propagating structures – due to the major constituent called ‘myelin’ sheath. This measure improves propagation of electric signals considerably. (C`) Amplification of the activity can be realized by channels that facilitate certain ion species to move across the membrane by changing their conformational state. Such channels are called ‘active’ for demarcating them from ‘passive’ (always open) channels that determine the passive membrane properties. Membranes with active channels show a variety of properties that passive membranes do not show. Accordingly, a lot of different types of active potentials supposedly have to be distinguished. But one type of active potential is spread across all animal species and extraordinary conspicuous when neural activity is analyzed. Sometimes it is called ‘action potential’ due to its active, energy consuming character; sometimes it is called ‘spike’ due to its short, needle-like form. For the sake of brevity the latter term is preferred in the following.

Phenomenologically, spikes differ in their form from other electric membrane events. Compared to most activities on passive membranes, spikes are fast and large potential changes (see fig. 6). Observed on a large timescale (a few seconds on an oscilloscope) spikes look like peaks on a straight line, a finer temporal resolution (a few milliseconds) reveals a recurring characteristic time-course. First, a gradual upward (positive) change in membrane potential (initial depolarization) can be observed. Then an abrupt steep ascent (depolarization) begins. Near the maximum, the ascend becomes less steep and ends in the maximum amplitude. The descend (repolarization) begins:

only after short duration, the descend obtains nearly the same steepness as the ascend had, even though the descend is usually less steep in average. Typically, the membrane potential falls to a value below (more negative than) the initial value (usually the resting potential), which, afterwards, is slowly approached again (afterhyperpolarization).

1.2.3.1. *Spikes as electric activity*

How can the time-course of a spike be brought about in an equivalent circuit? How can it be understood on the electric level of organization? It is impossible to explain this time-course with a simple RC-circuit (see also 1.2.2.2 and fig. 3). Simple RC-circuits also show a rising phase and a decay phase when stimulated appropriately, but they cannot account for the specific time-course of the spike and they cannot account for amplification. Explaining the time-course of spikes necessitates an extended RC-circuit with at least two different resistors and voltage sources. The initial depolarization can still be explained with a simple RC-circuit: if a current

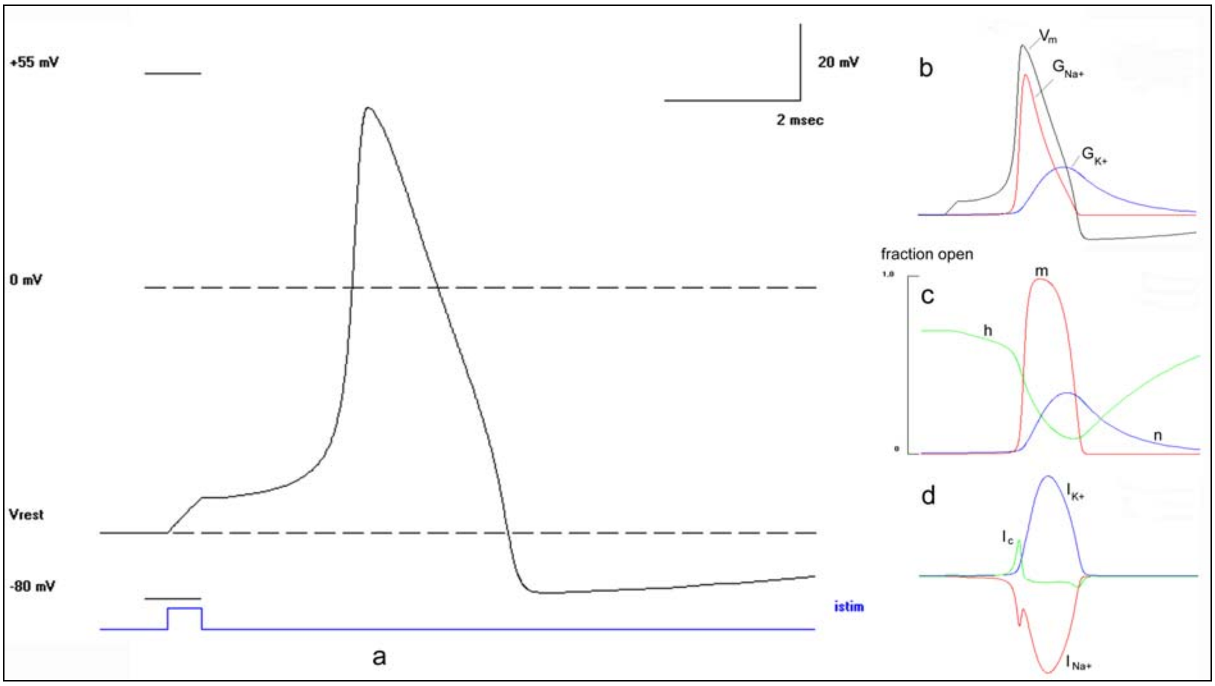


fig. 6: Spike. ***a)*** *Course of membrane potential (abscissa) over time (ordinate). After stimulation (i_{stim}), the membrane potential (V_{rest}) decreases from resting state, (depolarization), turns and increases towards negative values (repolarization). The characteristic afterhyperpolarization ('undershoot') passes into the resting state again.* ***b)*** *Conductance changes of potassium (G_{K+}) and sodium (G_{Na+}) during spike (V_m).* ***c)*** *Channel kinetics of sodium activation (m), sodium inactivation (h) and potassium activation (n) during action potential representing the fraction of open channels.* ***d)*** *Current course during a spike: capacitive (I_C), sodium (I_{Na+}) and potassium (I_{K+}). © Steven A. Siegelbaum 1994, APSIM v1.0.*

pulse is applied, first capacitive current flows. Subsequently, resistive current flows according to given conductances building the initial depolarization. But the transition from the plane to the steep part of the ascend can only be explained by introducing a new mechanism. One mechanism that can explain the steepness is a voltage dependent resistor ('varistor'), i.e. a resistor that changes its conductance with changes in voltage (see fig. 3c). In electric engineering, varistors are applied for preventing surge current to damage the system because the resistance increases for stronger currents. But here the varistor does exactly the opposite: the resistance is decreased with higher values so that more current can flow. This explains why the ascend of the positive potential becomes steeper. But what makes the potential descend again? Consider a second RC-circuit that contains a capacitor that is oppositely polarized. The current making up the steep ascend of the spike charges the second capacitor. This second RC-circuit also has a varistor that has a slower 'reaction time' than the first one. Thus, its increase in conductance lags behind. As the conductance of the second varistor increases, both currents work in the opposite direction and as the conductance of the first varistor decreases and the conductance of the second varistor increases, the overall potential does not rise anymore, but slowly turns and then descends again. (The afterhyperpolarization can be conceived as an effect of the prolonged increased conductance of the second varistor.) The two varistors explain the time-course of a spike. But how about amplification? To account for amplification, it can be assumed that the circuit is continuously charged by a voltage source and, thus, resembles a battery. The resting conductance of the varistors is so low that no current flows for discharging the battery, but it is 'unleashed' when the varistors are activated. In this sense, the spike is a local event at a large battery.

1.2.3.2. *Spikes as membrane events*

So far, an explanation on the electric level of organization was provided. On the molecular level of organization, most of the mechanisms introduced in the equivalent circuit also hold. In the resting state, the concentration gradients across the membrane correspond to a large battery that has stored chemical energy in abundance. Outside is sodium excess; inside is potassium excess leading to opposite concentration gradients. Potassium can leave the cell in limited amounts (leak conductance), but, as it pours out, it leaves behind a shortage of positive charges so that an electric gradient is generated. The equilibrium of all forces is the resting potential that is described by the Goldman equation (see also 1.2.2.2). This explains a single battery. When a positive current pulse is applied to the cell, it can be conceived as an injection of positive charges (or a withdrawal of negative charges). As a result of this current pulse, an excess of positive charges is inside the cell. The depolarization activates sodium specific channels that have 'sensors' for positive charges. The conductance for sodium increases and a sodium inward current is brought about. This explains the initial depolarization. But the phenomenological description of the spike's time-course now shows the transition from a rather plane ascend to the steep ascend. What happens? In the equivalent circuit, the first varistor was used for realizing this phase of the time-course. The analogues of varistors at the membrane are the voltage dependent channels. The initial depolarization first affects channels that are specific for sodium. In a multi-stage process, positive charges that come near the 'sensors' of the channels cause conformational changes of the protein. This change is the transition from close to open: sodium can pass the membrane along its concentration gradient from outside to inside (inward sodium current). The initial depolarization is amplified by the increased conductance that releases power saved in the concentration gradients (batteries) and still triggers further, 'neighboring' voltage dependent sodium channels.

The turning phase from ascend to descend can be explained by three other factors that affect the situation. First, as the membrane patch depolarizes, the sodium equilibrium potential is approached. The electrochemical force acting on sodium ions decreases and the current is reduced. It should be noted that these processes occur primarily locally, at a given membrane patch. (The potential, the electric forces, the concentrations etc. are not significantly changed for the whole cell - only at a given location.) The second limiting factor for the sodium current (and the cause for the turn of the potential) is an intrinsic property of the channel: after being opened, the voltage dependent channel enters a state of inactivation before it can be activated and thereafter opened again. The change from active to inactive occurs autonomously, independent of the voltage. Thus, after a certain period of time many channels are inactive. Third, another type of voltage dependent channel (the second varistor) comes into play: the voltage dependent potassium channel. It also 'senses' depolarizations, but its responsiveness lags behind that of the voltage dependent sodium channels so that the conductance increment for potassium of the membrane patch appears slightly delayed. (The leak conductance for potassium is very small relative to the conductance during a spike.) A potassium outward current along the chemical gradient can not only explain the turn of the potential, it also explains the repolarization of the membrane. The afterhyperpolarization can be explained by the still increased potassium conductance that is higher than the leak ('resting') conductance. Thus, the membrane potential comes closer to the potassium equilibrium potential than in the resting state. Ion species other than sodium and potassium play a minor role in the generation of the spike and are therefore neglected here.

As it was already stated for the explanation of responsiveness (see also 1.2.2), considering the molecular level of organization provides a 'stage' for the elements and activities participating in the explanation of the phenomenon. Ions can be imagined to be located around the membrane, move through channels etc. The molecular version of explaining spike generation also offers mechanisms that cannot be found easily on the electric level of organization. For instance, the voltage dependent channel is easier explained mechanistically than the varistor as the equivalent on the electric level of organization. In order to explain a varistor mechanistically, it has to be introduced that it is made of a mix of ceramic and metallic substances with an irregular structure that changes its conductance for electric charges as a response to currents in a complicated manner. By contrast, understanding the 'sensor'-mechanism of the voltage dependent channel in the membrane is no more difficult than understanding the function of a number of doors.

1.2.3.3. *Spikes as a formal event*

The last version of explaining spike generation offered here shall provide a formal description. The ionic (resistive) current determines the membrane potential. The overall ionic current can be described as the sum of all ionic current types $I_{ionic} = I_{Na+} + I_{K+} (+ I_{leak})$. A given current type depends on the conductance of the membrane g and the electrochemical forces acting on them $V_{actual} - E_{ion}$, i.e. the 'farer' the membrane potential 'keeps away' a given ion species from its equilibrium potential the stronger are the forces acting on that species $I_{ion} = g_{ion} (V_{actual} - E_{ion})$. It should be kept in mind that this is different for the different ion species: potassium tends to move outward, but only weakly since it is very close to its equilibrium potential E_{K+} (due to leak conductance), whereas sodium tends inward with great electrochemical force because it is far from its equilibrium potential E_{Na+}. The dynamics of conductance realized by the voltage dependence of channels is described by certain parameters that were determined experimentally. For example, the conductance of the membrane patch for sodium is described by two different parameters that are necessary to account for activation *m* and inactivation : $g_{Na+} \cdot m^3 h$. Potassium is slightly simpler in that it only needs a single parameter for activation *n*: $g_{K+} \cdot n^4$. Each parameter represents a time constant for a channel state and therewith provides a description of the time-course of the actual fraction of channels being open. The exponents can be memorized as the channel's gating elements, i.e. elements that have to be activated for opening the channel: the voltage dependent sodium channel is thought two have 3 (m^3) gating elements, the potassium channel 4 (n^4). Beyond the ionic current, capacitive current determines the overall current: $I_m = I_c + I_{ionic}$. The capacitive current can be described as the product of change of voltage in time and the actual capacity of the membrane patch $I_c = C \cdot dV_m / dt$. Finally, adding the leak currents yields an equation of the current that determines the time course of the membrane potential. The general forms $I_m = I_c + I_{ionic}$ and $I_{ionic} = I_{Na+} + I_{K+}$ $(+ I_{leak})$ result in:

$$I_m = C_m \cdot dV_m / dt + g_{Na+}\, m^3 \cdot h\, (V - V_{Na+}) + g_{K+} \cdot n^4\, (V - V_{K+}) + g_{leak} \cdot (V - V_{leak})$$

This so-called Hodgkin-Huxley equation contains the dependence on the membrane's capacitance on the actual change of potential, on the conductances as well as on channel kinetics and on the forces brought about by ion specific equilibrium potentials. The 'batteries' are represented in the different equilibrium potentials of sodium and potassium.

A vast number of novel phenomena can be obtained by combining the interactions of spike-related mechanisms introduced so far. Just to give one example, a novel phenomenon in the temporal domain called refractoriness can be obtained by applying two stimuli in short succession ('double pulse'). Refractoriness refers to the period of time in which it is not at all or only to a limited extent possible to generate a spike. That inter-stimulus-interval in which the second stimulus does not yield a spike is called absolute refractory period. That inter-stimulus-interval that yields an action potential with reduced amplitude is called relative refractory period which determines the minimal inter-spike-interval. Refractoriness can be easily explained in terms of molecular mechanisms as that period of time after a spike in which a great fraction of voltage dependent sodium channels are in the state of inactivation.

1.2.3.4. *Propagation of spikes*

So far, the generation of a spike was explained as a local event, i.e. with respect to a given membrane patch. But with respect to the whole neuron, spikes function to solve the problems introduced by the poorly conducting membrane: spikes transmit signals over large distances. The principle of propagation can be described with elaborated RC-circuits for an active membrane patch plugged in series. The neighboring membrane patch is activated in just the same manner as a current pulse through a stimulating electrode activated the initial membrane patch. However, a novel phenomenon is introduced by the properties of voltage dependent sodium channels. If the membrane patch of a given location has generated a spike, the massive presence of inactivation states of the channels implies that no second spike can be generated on the same patch of membrane for the absolute refractory period. This is functional for signal transmission since refractoriness prevents that a given activity causes 'ping-pong' spikes between neighboring membrane patches. But how exactly travels a spike from position *X* to position *Y*? First, current goes the way of least resistance. The elongated form of the propagating structures alone causes a higher resistance through than along the membrane. Thus, activity prefers to travel *along* the axon. But still two directions are left!? In fact, if a pulse were applied in the middle of an axon a spike would be propagated in both directions. This case is improbable to occur in natural situations, however, since spikes usually are not elicited in the middle of an axon – the natural situation usually assumed is that a spike is elicited at the beginning of the axon at 'the spike initialization zone'. Beyond this place (in opposite direction to the axon), the density of voltage dependent sodium channels is assumed to be too low for the generation of a spike. If a spike travels down an axon, it leaves behind a trace of inactivated channels that prevents the spike from being propagated back to its initial location. Thus, the propagation of the spike has a one-way direction.

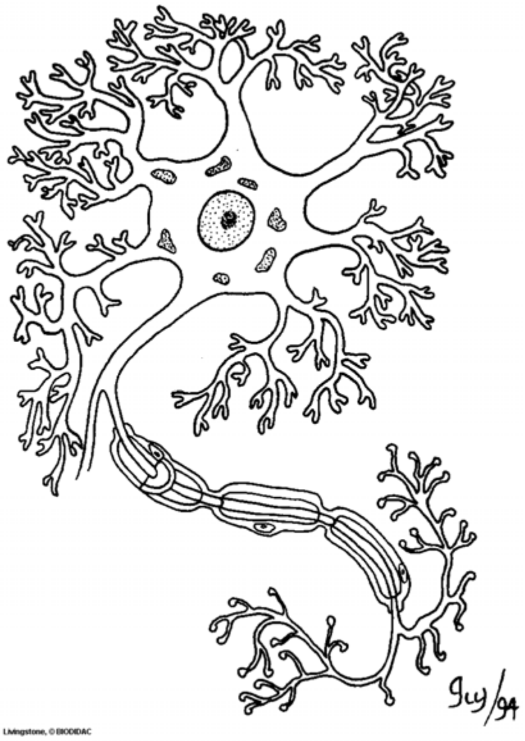
Mechanisms relating to spikes provide explanations for propagation and amplification. Signals are propagated (A) along a structure (axon) that has an increased resistance through the membrane due to its long and thin form, (B) the axon is myelinated by which the resistance through the membrane is increased and (C) the signal is amplified by way of spike generation. Since the residual leaks do still impede signal propagation considerably, it must be amplified during transmission along the myelin sheath, particularly for long distances. Therefore, the myelin sheath is interrupted ('Ranvier nodes'), thereby disclosing active membrane patches on which the signal can be intermediately amplified by a spike generation process. This form of propagation is called 'saltatory'.

This previous paragraphs introduced a basic account of spikes and their function of overcoming the poor performance of the passive membrane. Again, mechanisms on several levels of organization were necessary to form a sound explanation. However, only if a given phenomenon to be explained crucially depends on the exact spatiotemporal coordinates of spikes, it is mandatory to include all the mechanisms introduced in the previous paragraphs. Otherwise irrelevant levels of organization can be neglected. For example, if the precise timing of a spike is decisive for the 'correct' turning direction for the fly's escape maneuver away from the frog's tongue, comprehension of the underlying mechanisms might be indispensable. But, if a binary neuron without spatial extent is sufficient to explain the phenomenon (e.g. if there would be only one escape maneuver that is triggered by a signal representing the frog's tongue), considering spikes on the electric or molecular level of organization might not make sense. Which kind of explanation is necessary, again, depends on the question posed.

1.2.4. The classical neuron

There is no archetypal neuron. Rather, there are countless types of neurons the classification of which would be far beyond the actual scope. The rationale applied throughout this section is to demonstrate how to combine the constituent mechanisms for obtaining sound explanations of a given phenomenon. However, there is one account of a neuron that recurs again and again that shall therefore be called 'classical neuron'. Beyond the equipment of any other biological cell, the classical neuron has responsive passive and active membranes, soma (cell body), a dendritic tree and one axon. Since the functional contact zone between neurons is called synapse, we have postsynaptic terminals in the dendrites (receiving) presynaptic terminals in the axon (transmitting). These elements can be conceived as being strictly hierarchically organized, i.e. being linearly and transitively ordered (see fig. 7).

```
1.  environment
2.  organism
2.1.    system
2.1.1.      network
2.1.1.1.      neuron
2.1.1.1.1.    postsynaptic terminal
2.1.1.1.1.1.     postsynaptic membrane
2.1.1.1.1.1.1.     phospholipid-bi-layer
2.1.1.1.1.1.2.     channels
2.1.1.1.1.1.2.1. passive channel
2.1.1.1.1.1.2.2. active (ionotropic) channel
2.1.1.1.1.1.2.3. active (metabotropic) channel
2.1.1.1.2.    dendrite
2.1.1.1.2.1.     passive membrane
2.1.1.1.2.1.1.     phospholipid-bi-layer
2.1.1.1.2.1.2.     channels
2.1.1.1.2.1.2.1. passive channel
2.1.1.1.2.1.2.2. sodium-potassium-pump
2.1.1.1.3.    soma
2.1.1.1.3.1.     passive membrane...
2.1.1.1.4.    spike initialization zone
2.1.1.1.4.1.     active membrane
2.1.1.1.4.1.1.     phospholipid-bi-layer
2.1.1.1.4.1.2.     channels
2.1.1.1.4.1.2.1. active channel (ionotrop)
2.1.1.1.4.1.2.1.1.      voltage dependent sodi
2.1.1.1.4.1.2.1.2.      voltage dependent pota
2.1.1.1.4.1.2.2. sodium-potassium-pump
2.1.1.1.5.    axon
2.1.1.1.5.1.     myelin sheath
2.1.1.1.5.2.       active membrane
2.1.1.1.5.2.1.     phospholipid-bi-layer
2.1.1.1.5.2.2.       channels
2.1.1.1.5.2.2.1. active (ionotrop) channel
2.1.1.1.5.2.2.1.1.      voltage dependent sodium channel
2.1.1.1.5.2.2.1.2.      voltage dependent potassium channel
2.1.1.1.5.2.2.2. sodium-potassium-pump
2.1.1.1.6.    presynaptic terminal
2.1.1.1.6.1.       membrane vesicles
2.1.1.1.6.1.1.       transmitter
2.1.1.1.6.2.      presynaptic membrane
2.1.1.1.6.2.1.      phospholipid-bi-layer
2.1.1.1.6.2.2.        channels
2.1.1.1.6.2.2.1. passive channel
2.1.1.1.6.2.2.2. active (ionotrop) channel
2.1.1.1.6.2.2.2.1.      voltage dependent calcium channel
2.1.1.2   synaptic cleft
```



Livingstone, © BIODIDAC

***fig. 7: Outline of the 'classical' neuron.** Possible general structure of concepts above and within the 'classical' neuron. The structure has a spatial organization beginning with the largest ('environment') and going to the smallest and it has a temporal organization (following a hypothetical signal) going from presynaptic to postsynaptic. It should be noted that some elements have multiple presence (e.g. active membrane). Moreover, some elements could be positioned on other places (e.g. cytoplasm) constructing other accounts of the neuron, e.g. one that does not go from presynaptic to postsynaptic but from outside to inside: { extracellular space | neuron { membrane | cytoplasm } }. Inset shows a classical neuron as it is similarly presented in many textbooks. Insert © BIODIDAC (license permissive for reprint).*

The classical neuron is particularly appealing because the elements are not only hierarchical, they are also spatially and temporarily organized: an incoming activity has to run through the spatiotemporal organization of the classical neuron, is modified by various elementary mechanisms and thereby transformed to the outgoing activity. In explanations of brain phenomena, the classical neuron is a kind of implicit operating instruction of how to put all the elementary mechanisms together in order to obtain a holistic notion on the cellular level of organization.

For achieving a detailed spatial understanding of a neuron, mechanisms acting on the electric level of organization have to be applied to a given geometrical architecture of a neuron. An example is the construction of a compartmental model of a classical neuron that can explain the generation of spikes. Such a model can be conceived as a large wire frame consisting of numerous RC-circuits. How much RC-circuits are chosen, depends on the problem to be solved; more details call for more RC-circuits. Even though the classical neuron does not have one specific classical function, it has a general functionality: it is usually assumed that the dendrites have ramifications that make contact to different presynaptic neurons. Thus, they integrate activities from several presynaptic neurons. In the situation of the fly trying to escape the frog's tongue, a useful function for a classical neuron would be the pooling of spatial information: since the frog's tongue is represented in the fly's eye as an accumulation of single activities corresponding to single receptive fields of photoreceptors, it might be useful for the fly's escape to integrate these activities into a larger representation of, say, an object (a tongue) or an edge (boundary between tongue and not-tongue). Thus, as an exemplary function of the classical neuron it can be supposed that it pools activities of a specific area (say the lower left corner) of the fly's retina.

The classical neuron is a mental assistance that helps to localize and trace activities entering a neuron and understand their transformations. Multiple activities are received at the postsynaptic terminals and are conducted by the passive membrane. Since there is no strict mechanism that directs the activities (as there is one in the case of spikes), activities spread in all directions ('electrotonically'). Those activities the fate of which it is to dwindle without significant consequences somewhere in the extracellular space do not have to be further considered in the classical neuron. The next relevant situation for conducted activities is encountered when (the residuals of) two activities meet at a ramification point and are merged – a process called ‘spatial summation’. In equivalent circuits, this case can be described as parallel resistors: consider the two dendritic ramifications as two feeding resistors R_1 and R_2 that converge on to the receiving element R. The sum of current must be zero (first Kirchhoff law $I_1 + I_2 - I = 0$), i.e. the two feeding currents are summarized and have the same value as the receiving current. The conductance of the receiving element is limited by the sum of the conductances of the feeding elements ($g = 1 / R$). There cannot be more current (charges per time) than that what is delivered by the feeding elements. The overall potential is determined by the sum of the constituent potentials that, in turn, are limited by the resistances (second Kirchhoff law $U = I_1 \cdot R_1 + I_2 \cdot R_2$). Applying these rules provides possibilities of regulating the current flow by the design of the circuit. If the conductance of the receiving element was too low the activities would be damped and prolonged and activity is lost. Thus, the conductance of the receiving element must be at least as high as the sum of the feeding elements if the sum of the currents shall be transmitted. For a dendritic tree with constant membrane properties (resistivity, constant number of passive channels) this implies that the diameter must be increased for decreasing resistance through the membrane, i.e. the twigs of the dendrite get thicker towards the center. This, in turn, implies an increase of resistance along the membrane and, accordingly, a “loss” of activities ($\lambda = r_{leak} / r_{cyto}$). For loosing least activity, the best design of a receiving element is the one that corresponds exactly to the sum of the conductances of the feeding elements. This example illustrates that dendritic tree design can determine the concrete computation of activities. It should be noted, however, that there might also be cases in which damping a given activity by the design of the dendritic ramifications is functional, for example if ‘undesired’ activity shall be suppressed.

The dynamics of temporal summation also determines the outcome of incoming postsynaptic activities. Since current represents the amount of moved charges per time, temporal summation basically follows the same rules as spatial summation of currents (Kirchhoff laws). Thus, in an equivalent circuit with parallel resistors that represents a ramification of a dendritic tree, two feeding currents are simply added in each time slice in the receiving element. However, the specific structure of the feeding elements can change the time-course of the feeding activity thereby delaying one activity before or after the other. A temporal shift “smears” two initially synchronous feeding activities because the overall duration of activities being present will prolong (accompanied by a reduction of amplitude) because the maximum values of the feeding activities fail to 'meet' in the same time-slice. For instance, thicker elements prolong the time-course of a given activity on passive membranes - the time constant increases with resistance $\tau \sim r_{cyto}$ (see also 1.2.2.5). Considering the molecular level of organization, this can be conceived as currents that take longer to get through a larger volume of medium (cytoplasm).

In the classical neuron, all activities are finally integrated by way of spatial and temporal summation into a single merged activity at the final ramification of the dendritic tree. The merged activity spreads into the regions of the "spike initialization zone". On the electric level of organization, i.e. as an electronic component, this region can be conceived as a varistor with very high responsiveness, i.e. the threshold for initially increasing conductance is very low. On the molecular level of organization this property can be conceived as a high density of voltage dependent sodium channels. Thus, in the course of activity traveling through the classical neuron, now an active membrane is present. Here, a spike is generated if the merged activity arriving from the dendritic tree succeeds in activating enough voltage dependent sodium channels for initiating the self-amplifying process of channels activating channels... This process also depends on the time-course of the merged activity. The longer charges are present the more channels will be activated by these charges. But the longer the charges are present the more channels will get inactivated, too. Thus, a brief, high-amplitude activity is thought to be optimal for spike generation. Therefore synchronous activities in the presynaptic region are thought to be more likely to elicit a spike (given an equal propagation over the dendritic compartments).

Eventually, an incoming activity ends either as contributor to a spike that travels saltatoriously down the axon in order to be transmitted to the subsequent cell on the presynaptic terminal - or it dwindles as leakage current in the extracellular space. For this strict division between 'yes' and 'no', spike generation is frequently called an all-or-none process or a threshold procedure. The form of the threshold, i.e. the amplitude and duration of the pulse necessary to elicit a spike, depends crucially on the density of the voltage dependent sodium channels but also on their kinetics, i.e. time constants for the phase transitions resting-active-inactivated (see also 1.2.3.3).

The classical neuron is a simplification, a model in exactly the same sense as the binary neuron – it is only more complicated. Expressed a bit sloppily, it assumes that all dendrites pick up little packages of activity, bring them to the spike initialization zone where they all are collected and evaluated so that the spike supervisor says: "Ok, that's enough for a spike. Fire!" or "No, throw the old activity stuff away. We begin anew!" If the spike is fired it travels all the way down to the axon to the next relay station. This account is very similar to the binary neuron. The concrete forms of spatiotemporal integration of activity in the dendrites (currents flowing 'backwards' in neighboring dendrites, for example) are often neglected. The possibility of neurons without axons, or axons having ramifications in their terminal region are blended out and the question whether activities other than those resulting from the stimulus that one has in mind (e.g. the frog's tongue) might determine whether or not a spike is generated are usually not considered. Thus, thinking in 'biological terms' (classical neuron) can make possible thinking in terms of mathematical functions (binary neuron) by assuming a very sound case that does not afford much mental capacity by invoking problematic explanatory sidetracks. Even experts for the lower levels of organization (i.e. electric or molecular) can benefit from thinking in terms of the classical neuron, as it makes possible to ignore complexity meanwhile. However, the type and number cognitive processes involved (mental models activated or schemas instantiated) depends crucially on the prior knowledge of the auditor. Someone can think of a neuron as a binary element just differentiating between yes and no, while another one thinks of a color coded dynamic 3D full featured membrane model ad yet another one of a system of differential equations. There is no wrong or right – there is just (in)adequacy for explaining a given phenomenon.

1.2.5. Synapse

The synapse is both a crucial *spatial* connector since it explains mechanistically the transition between the level of neurons and the level of networks and the synapse a crucial *temporal* connector since it explains mechanistically how real-time phenomena (e.g. sensation) relate to 'persisting' phenomena (e.g. memory). The real-time functionality of the synapse comprises the transmission mechanism between two neurons in its entirety. The synapse is not a straightforward spatial unit: Does the synapse belong to the presynaptic or the postsynaptic neuron? The synapse is neither a temporally discrete event: When does 'before' begin and 'after' end? It is literally the 'in-between' two neurons, two instances of time etc.

As all brain phenomena, synapses can be explained on different levels of organization. Particularly, two types of synapses call for explanations on different levels of organization: electric synapses vs. chemical synapses (see e.g. Bennett 1997). The 'story' about electric synapses is most often briefly told (even though they might actually deserve more detailed treatment): two neurons are connected by specific functional structures called 'gap junctions' that allow a nearly instantaneous, largely unchanged transmission from one neuron to the next.

The chemical synapse is a functional unit in which neural activities can be significantly modified so that they have a completely different form after transmission. Thus, beside the function of merely transmitting a given activity, chemical synapses are thought to transform activities. Both neurons, the transmitting neuron with the presynaptic side and the receiving neuron with the postsynaptic side contribute to a synaptic transformation. In between lies a specific part of the extracellular space, called synaptic cleft. Anatomically, the whole unit is typically characterized by a membranous bulb-like protuberance on the presynaptic side, sometimes also on the postsynaptic side.

Chemical synapses transform a presynaptic electric activity into largely non-electric, molecular elements that in turn is transformed to a postsynaptic activity. At synapses, the electric level of organization (i.e. electric changes measurable with microelectrodes) is only tangible by comparing the presynaptic activity with the postsynaptic activity. Such an input-output analysis tells nothing about the molecular mechanism that characterizes synapses, but it can characterize synaptic phenomena that are then to be explained in terms of molecular mechanisms. A typical phenomenon found at chemical synapses is the 'delay' that synaptic transmission introduces: The time it takes for an activity to get from one neuron to the next is considerably longer than the time it takes for the activity to travel down the same distance an axon. Furthermore, the activity can be changed from a given type of presynaptic signal to almost any other type of postsynaptic signal. What are these transformations? Consider a current step as presynaptic activity. Depending on the synapse, this step on the presynaptic side can look like an impulse on the postsynaptic side, but it can be also prolonged, it can even made negative or have a positive peak at the beginning and a negative peak at the end. It can stay largely unchanged, though.

Explaining synaptic phenomena in terms of mechanisms is best possible by considering both the electric and the molecular level of organization. It is thought that the activities are transformed from the electrical-chemical into the chemical-molecular domain and back again. Thus, if only one level was considered, an incomplete account of synaptic phenomena would be provided. An explanation that reduces to the electric level of organization must fail at this instance because it could simply not detect the chemical processes involved. An explanation that reduces to the molecular level of organization must fail because the computational context would be missing when electric activity is neglected. This illustrates very concretely that a mix of levels of organization is the appropriate approach for explaining brain phenomena.

The key process at the synapse is the involvement of so called transmitters. The rationale is that an activity in the form of an ionic current of a given amplitude and duration is translated into a given amount of transmitter that transports the activity over the synaptic cleft, there causing a retranslation into an ionic current of a given amplitude and duration. (In this period of time, the neuronal activity relating to the frog's tongue is not an electric potential but an amount of a chemical substance.) The process in between has the form of a complex cascade of events that shall not be taken into account in every detail. The principle, however, is quite simple: transmitter is stored on the presynaptic side in vesicles, i.e. membrane inclusions. Activity arriving at the presynaptic side activates voltage dependent calcium channels, thereby causing an inward calcium current. A series of events is triggered by the incoming calcium ions. Effectively, this series of events causes the vesicles to fuse with the presynaptic membrane and transmitters are released into the synaptic cleft. Apart from some recycling processes on the presynaptic terminal, the activity has now left the presynaptic neuron. (No frog's tongue represented anymore in that neuron.) Transmitter spreads through the synaptic cleft and binds to specific receptors on the postsynaptic membrane thereby initiating another series of events, this time on the postsynaptic membrane. (Ionotropic transmitters directly bind to channels, metabotropic transmitters act indirectly by binding to other proteins that in turn trigger further processes.) Effectively, transmitter binding causes ion channels to open or close by a change in their conformational state. This opening or closure on a fraction of channels on the postsynaptic side changes the current of that ion species that moves through the channel thereby influences the postsynaptic potential. Thus, the selectivity of the channel for the transmitter on the one side and an ion species on the other side realizes the retranslation from the molecular into the electric domain. This newly formed activity can spread over the postsynaptic area (usually a dendrite). Apart from some clearing works in the presynaptic cleft (reuptake of transmitter) the synaptic transmission is completed. The postsynaptic activity is ended by inactivation of the postsynaptic channels.

The account of synaptic transmission introduced so far can explain some of the synaptic phenomena introduced above, for instance, the delay that is typically observable when the presynaptic and postsynaptic signal are compared. Obviously, the reason for the delay lies in the transformations from ionic currents through channels on the presynaptic side to other molecular mechanisms at the synapse and back to ionic currents. Compared to the electrochemical processes on a passive membrane that determine the dynamics of a given activity, the molecular processes involved in synaptic transmission take considerably more time. In the case of the passive membrane, a given activity causes capacitive current that spreads almost immediately. The subsequent ionic current has a considerable delay, but solely depends on diffusion. In contrast, consider the multiple processes involved in synaptic transmission: opening of calcium channel --- activation of the vesicle --- fusion with membrane --- spread over cleft --- activation of postsynaptic channels --- (lines indicate the presence of intermediate processes). These processes have to occur one after the other for building a causal chain and, therefore, the durations of the single events have to be summed up. Thus, even if only a minimal duration of each process is assumed, it should become clear that these processes take longer than a diffusion process along a gradient through open channels. It should be noted that the overall time for synaptic transmission is not that much longer than a spike traveling down an axon (both are in the transition between μs- and ms-range), which, however, has to overcome a much longer distance (mm-range for axon vs. nm-range for synapse).

Activity cannot only be delayed at a synapse, but can be also transformed. How can these transformations of a presynaptic signal into a sometimes completely different postsynaptic signal be explained? Again, the participating elements and activities are manifold and constitute a complex series of events. However, the principle is simple again: the mechanism is grounded in the selectivity of the channels for a given ion species. If sodium channels are opened, a sodium inward current and a depolarization is caused, i.e. the postsynaptic potential is positive or excitatory (EPSP). If chloride (or potassium) channels are opened an inward chloride (or outward potassium) current and a hyperpolarization is caused, i.e. the postsynaptic potential is negative or inhibitory (IPSP). The activation of a given channel type is determined by the transmitter that influences the states of the channel. Thus, combining appropriately transmitters and receptors yields the different possibilities of altering activities.This explains the direction of the postsynaptic effect, i.e. if it is positive or negative, but what about amplitude and duration of the postsynaptic signal? Quantitative transformation can partly be attributed to the characteristics of the synapse; the other part is determined by the incoming activity. The characteristics of the synapse that influence the amplitude of the postsynaptic signal are determined by the density of postsynaptic channels through which postsynaptic currents flows, by the membrane equipment, e.g. the different ion specific channels involved, the reversal potentials of the ions involved etc. Furthermore, the effectiveness of the whole transmission process is important: How much transmitter is activated per unit presynaptic activity and how much of the transmitter binds effectively to postsynaptic channels? This is the point where the stimulus-induced component takes effect - coarsely said: the more activity, the more transmitter.

Dynamics (temporal prolongation or truncation) of the activity can, for instance, be determined by the kinetics of the postsynaptic channels, i.e. how long they are activated per transmitter action. (These dynamics can be conceived in the same way as the parameters *n, m* and *h* affect the membrane current in the Hodgkin-Huxley equation, see also 1.2.2.3). Short opening and long inactivation phases of the channel lead to truncation. For instance, a postsynaptic event has stopped, although a presynaptic event is still present. Long opening times can lead to prolongations and possibly to amplification of the presynaptic signal. The mechanisms involved in synaptic transmission can also explain phenomena that exceed 'real-time', i.e. the timescale of an actual behavior (e.g. the fly escaping the frog's tongue). Consider the marine snail *Aplysia* that withdraws its gill in order to protect itself against environmental dangers when specific sensitive areas (the 'siphon' or the head) are touched (see fig. 8). Now, certain events can change the behavior over a larger timescale. For instance, touching the siphon repeatedly causes the snail not to withdraw its gill. (This might be functional or economic since the stimulus is obviously not dangerous because the snail is still living and able to withdraw its gill.) An explanation in synaptic terms can be given in terms of an inactivation of calcium channels (presynaptic to the decisive motor neuron) that leads to a decrease in calcium influx, a decrease in vesicle activation etc. In this way, the effectiveness of synaptic transmission is changed over seconds, minutes or even hours. This process is called 'habituation' and the synaptic mechanism 'presynaptic depression'.

Sensitization (sometimes also called dishabituation) is observable as an increase in the strength of gill withdrawal. This can be brought about by applying a new stimulus on the head. (This could be dangerous!) An explanation in synaptic terms assumes that the additional stimulus results in an activity of another neuron releasing a transmitter that can bind on the site that also is responsible for habituation, i.e. the site presynaptic to the decisive motor neuron. Effectively (after a series of events), it is thought to inactivate potassium channels, thereby prolonging the presynaptic activity. Thus, effectiveness of synaptic transmission to the decisive motor neuron is increased by way of 'presynaptic facilitation'.

Even longer timescales than the minutes of habituation can be approached by applying very strong stimuli on head and siphon. Then, the concentration of intermediate products of a first "series of events" (cAMP and calmodulin) is increased considerably and activate another "series of events", namely the activation of processes in the genetic material. Generally, gene expression allows almost any conceivable change in a neural system: changing of the synaptic equipment (e.g. variation of the number of channels or addition of a new channel type), induction of growth processes of new synapses etc. (It should be noted that these processes are not exclusively synaptic because the whole cell comes into focus.) Since the marine snail can be caused to withdraw the gill with increased strength for days, the phenomenon is termed ‘learning’, more precise ‘associative learning’.

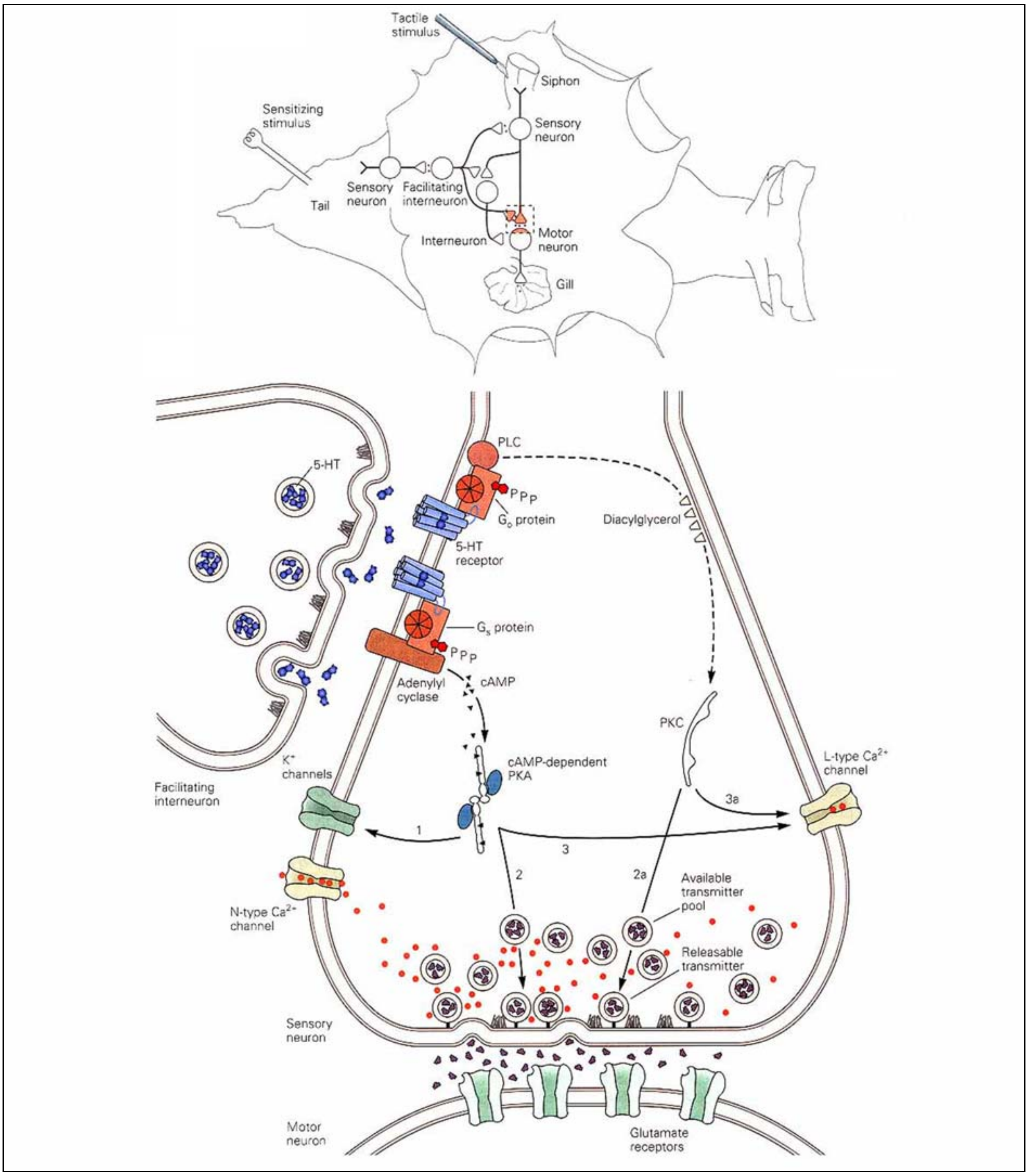


fig. 8: ***Plasticity in Aplysia.*** *Upper: Schematic neural circuit. Tactile stimulation of the siphon causes withdrawal of the gill. Response decreases, however, when stimulation is repeated (habituation). The initial response strength can be brought back by applying another stimulus on the tail (dishabituation, sensitization). Lower: Molecular mechanisms habituation and dishabituation on a synapse (zoomed from dotted line in the left figure). Habituation is explained as resulting from decreased activity transmitted to the motor neuron (synaptic depression). Sensitization is explained by activation of additional molecular mechanisms by a third synapse (presynaptic facilitation). The figures illustrate the principle of explanations on different levels of organization. from uscd.edu.*

A lot more could be said about synaptic transmission, particularly about more sophisticated computations at synapses or about learning. But the section so far should last to illustrate what synapses are about. Synaptic transmission involves multiple mechanisms on different levels of organization. Moreover, it prepares the ground for a general change of the level of organization on which the functionally significant processes are located: whereas the representation of the frog's tongue is in the classical neuron thought to be predominantly observable as an electric phenomenon, it is the concentration of transmitters that contains the representation during chemical synaptic transmission. But seen from the design stance, it seems circuitous to introduce such a complicated process. Why not just take electric synapses? An obvious benefit of such a transformation stage is their disposition for alteration of a given activity. Most of the conceivable transformations that can be performed on a given activity can be realized at the synapse. Thus, the transmission characteristics of the synapse determine the 'fate' of a given activity. Moreover, plasticity allows that the transmission characteristics themselves can also be altered. Plasticity introduces a strong dynamic quality since the activity itself changes the transmission characteristic. For example, if a synapse showing plasticity transmits the same presynaptic activity several times, each postsynaptic signal may look different. Therefore, a clear separation of stimulus (activity) induced component in a postsynaptic signal and the contribution of the synaptic transmission characteristic to the postsynaptic signal cannot easily be maintained. Of course, dynamics are introduced at the cost of making the synaptic activity more difficult to decode – but the benefit is an extension of representational power: an external event can not only be represented as a short, transient effect of an external change, but can impose longer lasting changes as in the case of habituation or can become (quasi) persistent such as an association. The transitions between these forms of representation, however, are only gradual.

For assessing the overall effect of synaptic transmission on the postsynaptic neuron, it has to be taken into account that usually multiple (thousands) of synapses are activated in parallel. In a 'classical neuron', these are the activities that enter the spatial and temporal integration processes ('summations') that extend from the dendritic ramifications to the spike initialization zone. Reconsider equilibrium potentials, conductance etc. from the explanations of a simple current pulse on a patch of membrane (see also

1.2.2). Now, think of thousands of those activities in a dendritic tree. If it is then taken into account that all these activities come from synapses and that each synaptic transformation can be different, it becomes clear that synapses extend the scope of brain phenomena significantly: they do not only represent the crucial spatial and temporal connector for explanations of brain phenomena, they make conceivable almost any alteration or – seen in a functional context – computation of activity.

Since the synapse introduces many new phenomena, it shall briefly be discussed what state of explanation is reached so far. Whatever complex the mechanisms necessary to provide a sound explanation for a given phenomenon might appear, the building blocks still are few. There are three membrane types and a gap, i.e. a passive, active, synaptic membrane and the synaptic cleft. These membranes are functionally determined by the two channel types 'passive' and 'active', the latter of which being ionotropic (controlled by ions/voltage) or metabotropic (controlled by ligands) and direct (activated by transmitters) or indirect (activated by mediators). Ions are present in differential concentrations that can represent environmental changes by moving through the channels. This ionic current representing environmental changes (e.g. a dangerous stimulus) can subsequently trigger processes that control the concentrations of transmitters that control the concentration of other substances triggering gene expression that controls the building of membranes and channels …

It should be noted that neither the elements themselves nor the functions in which the elements are embedded are too difficult to imagine. Also, it is not difficult to integrate these into a mechanism – as long as a mechanism considered isolated. But when multiple mechanisms are considered, multiple causes act on the elements (e.g. ions) and the possibilities of thinking causal relations increase. (Think of different receptor types on a patch of synaptic membrane determining together postsynaptic currents.) If, additionally, dynamics are considered (e.g. adaptation), yet another quality is introduced. Finally, the possibility of temporal orders determining temporal orders (plasticity) and the generative mechanism of creating new elements (genetic expression) makes the possibilities endless. Thus, the principles of synaptic transmission supply explanations of brain phenomena with the power to equip neural systems with much more computational functionality than with neurons alone. Moreover they allow to connect neurons spatially and temporally and to design networks. When networks of several neurons with thousands of synapses per neuron and thousands of dendrites (or RC-circuits) per neuron are considered, complexity can easily get out of hand. Imagine how a given activity spreads in such a network! It should become clear at this instance how simplifying the classical neuron is. However, these remarks should not be deterring. Rather, they shall remind to keep the focus controlled: explanations on brains are easy as long as the explanatory aperture is narrow, but bear the power of unfolding massive complexity as the aperture widens.

1.2.6. Networks

Is any brain phenomenon explainable with the building blocks introduced so far? Do the explanations of brain phenomena considered so far imply that nothing new can be expected? Yes and no! Yes because no new building blocks are necessary and no because their combinatorial powers allow the construction of novel phenomena. This means: new are the phenomena, the observations to be made. What about networks, i.e. multiple neurons connected by synapses? Is it necessary to introduce a new level of organization, called networks? Isn't it possible to explain any function of a network with the explanatory means gathered so far? It might be possible to explain network phenomena by low level phenomena – the functional result might be the same, no matter if a network is made up of electric circuits, dendritic trees or neurons. As long as the high levels are reducible to low levels, explanations can be based on any level of organization. But, as has been pointed out before (see also 1.1.12), it is not necessary in any case to explain any phenomenon down to the lowest level because possibly no gain is achieved. Moreover, sometimes dynamics and complexity do not allow for consideration of low-level phenomena since mental capacity is exceeded.

Reconsider the classical neuron that is applied in explanations of brain phenomena as a kind of 'concept container': it reflects a fairly simple account of neurons that allows to apply it as an entity that is not further scrutinized, i.e. reduced to lower levels. But it is implicitly assumed that the entity can be reduced to low levels anytime when necessary. Thus, the conception as an entity allows for the reduction of complexity. The classical neuron serves as mental packaging mechanism. For example, the classical neuron makes possible to think of neurons as binary units: spikes help to justify why complex activities of neurons can be simplified to a being active or being inactive. Now, the network level of organization is often characterized by a neglect of the cellular level: neurons are treated as abstract units that can carry out various computational operations on activities of precursors and pass the results of these computations to successors. It is not considered what spatial organization the computational units might have. In this sense, packaging low levels with the classical neuron makes possible to adopt the network stance.

The transition from neurons to networks implies several changes in the levels-of-organization framework. When the spatiotemporal aperture is widened, phenomena resting on more than one interconnected neuron appear. When the functional aperture is widened, an increased computational complexity is revealed. These perspectives correspond to two prevalent notions of networks: first, small circuits made of neurons, chemical synapses, molecules etc. and, second, artificial neural networks. Even though these two notions are considerably different, both typically focus on computational phenomena – one with respect to biological realization, the other one with respect to algorithmic realization. Both assume that networks as opposed to single neurons realize that activity is processed parallel and distributed. In sum, networks are a distinct level of organization. The neglect of details makes possible compact and comprehensible explanations of parallel distributed processing.

1.2.6.1. *An example: Motion detection*

A phenomenon that offers explanations for both biological realization and algorithmic realization is motion detection (Borst & Egelhaaf 1989). However, biological realization was already treated in depth for the cellular level of organization in the previous sections. Therefore, a simple algorithmic account (that is characterized by a neglect of details) is provided here (see fig. 9). Motion is a change of an element in space and time, i.e. a spatiotemporal displacement. If an element on position x_1 (left) changes to position x_2 (right) from one instant of time t_1 to the next t_2, it has moved (from left to right). Now, how can a device be designed that can detect such a change? For the element to be detected, a receptor r_1 for visual stimuli coming from x_1 has to be present. This receptor can detect the change that occurs when the element disappears at t_2. Thus, a single receptor can detect if something has disappeared, but not unambiguously detect if it has moved: therefore, a second receptor r_2 for visual stimuli coming from x_2 has to be present. If (and only if) the other receptor r_1 detects the stimulus at the same instant of time the first receptor detects nothing anymore, the movement can be detected.

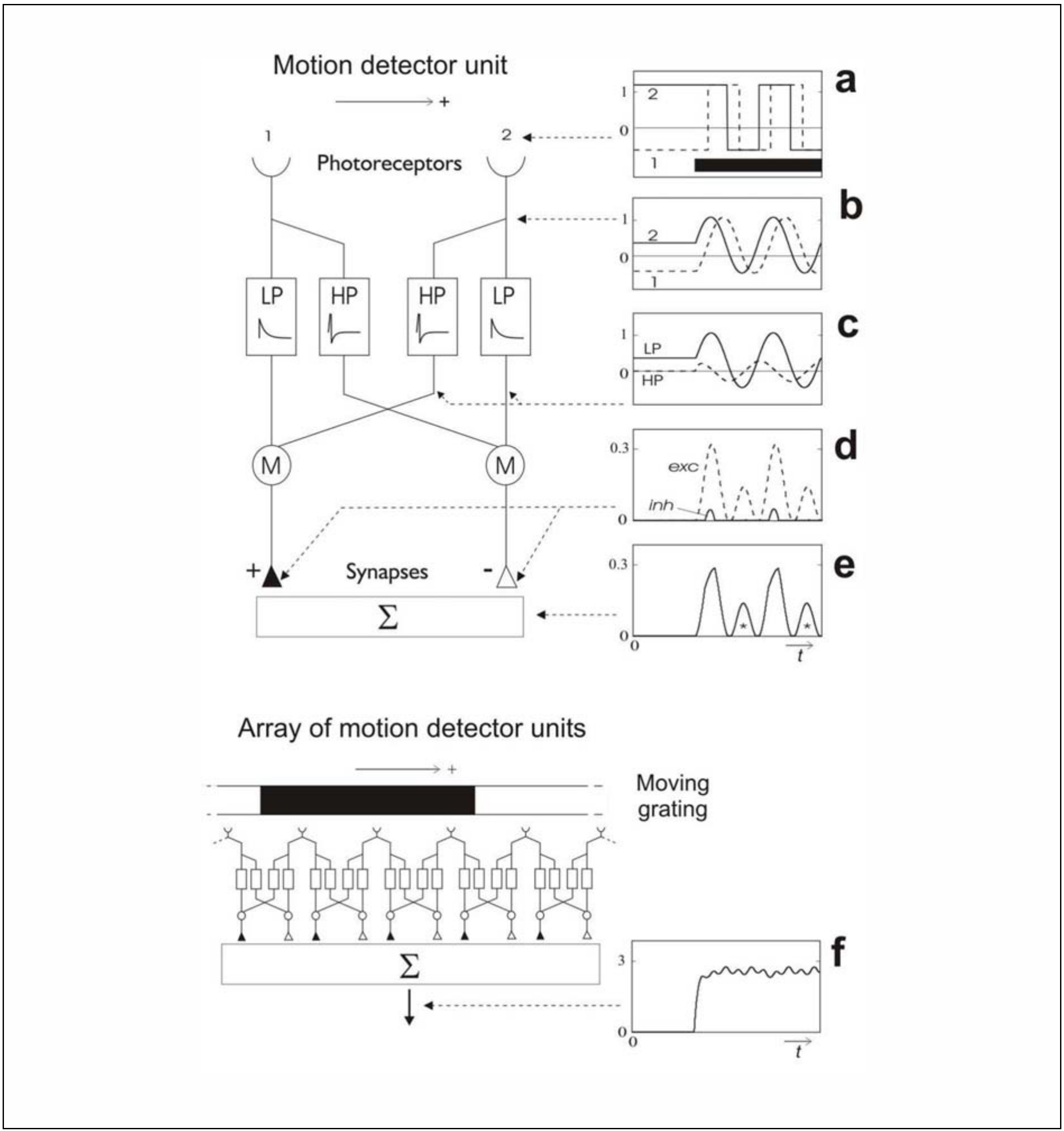


*fig. 9: **Motion Detector.** Upper left: Elementary motion detector with different stages of processing (a-e) responses to a square-wave grating moved in the preferred direction. Two symmetric half-detectors are shown that 'use' the same receptors and the responses of which are subtracted from another in order to obtain an unambiguous response. Lower left: Array of motion detectors that illustrates the principle of combining EMDs for sampling larger visual fields (from Haag et. al. 1999).*

Initially, the representation of the movement is distributed over the two receptors. For an integrated representation of movement the two receptor activities have to be computed. A simple computing element could add up both signals. But such a computation would be not specific enough, because its activities would be the same when movement is present or a single receptor was active over both instants of time, i.e. both conditions $x_1(t_1) = 1$ and $x_1(t_2) = 1$ *or* $x_1(t_1) = 1$ and $x_2(t_2) = 1$ amount to "2". Thus, unambiguous results can be obtained only by way of evaluating the activities of both receptors at both instants of time. If these activities were summed up, the evaluating element would receive activity for both instants of time $x_1 + x_2(t_1) = 1 + 0 = 1$ and $x_1 + x_2\ (t_2) = 0 + 1 = 1$. An observer could not see a difference in the resulting activity of the evaluating element between the two instants of time: no movement is detected. Therefore, the second instant of time t_2 has to be compared to the first instant of time t_1. This can be managed by making the activity of the first instant of time 'wait' for the activity on the second instant of time, i.e. delaying the first activity. If the duration of the delay matches exactly the time the stimulus takes to move from one location to the next, the resulting activity of the evaluating element is "0" on the first instant of time and "2" at the second instant of time. t_1: $r_1(t_0) + r_2(t) = 0 + 0 = 0$, t_2: $r_1(t_0) + r_2\ (t) = 1 + 1 = 2$. Thus, the delay provides a means for obtaining detection of movement.

Up to this point, it was neglected that a stimulus can also be longer than only one instant of time. Such a continuous stimulus makes considerable problems for the motion detector introduced so far: a continuous stimulus that activates both receptors would result in the same activation of the evaluating element as a moving stimulus t_1: $r_1(t_0) + r_2\ (t) = 1 + 1 = 2$, t_2: $r_1(t_0) + r_2\ (t) = 1 + 1 = 2$. Undesirably, there would be not only detection of movement, but also detection of continuous stimuli. A solution to this problem would be to allow only 'changes' to enter computations in the evaluating element. The desired response is called 'phasic' (biological term), differentiated (mathematical term) or high-pass filtered (cybernetic term) response, which means that only the beginning (and possibly the end) of a stimulus is transmitted, while the continuous part is neglected t_1: $r_1(t_0) + r_2(t) = 0 + 0 = 0$, t_2: $r_1(t_0) + r_2(t) = 0 + 0 = 0$. Thus, the most effective method to handle the problem of continuous stimulation of receptors in motion detection is to simply prevent their appearance in the computation. After all, it is change that matters to motion.

Of course, some problems are left aside. One problem is logical: consider two motion stimuli that are applied one after the other. If the first stimulus is delayed as long as the second one needs to arrive at the second receptor, these two are paired and motion is detected. However, two different objects caused the motion. The first stimulus has already passed the receptor before the second arrived. This phenomenon is called 'aliasing' and can be observed for durations of inter-stimulus-intervals that are multiple integers of the delay.

Moreover, two synchronously appearing stimuli on both locations (no motion) would still cause the device to respond $t_1\!: r_1\,(t_0) + r_2\,(t) = 0 + 1 = 1$, $t_2\!: r_1\,(t_0) + r_2\,(t) = 1 + 0 = 1$. Even though this is not the maximum response (which is "2") the process is still not exclusively specific for motion. Exclusive motion information could be obtained by applying a threshold function, though. Another simple change also helps: Instead of summing the incoming activities, they can be multiplied. Then, the criterion from the beginning is sufficed: if (and only if) r_1 and r_2 are activated on two subsequent instants of time t_1 and t_2, motion is detected: $t_1\!: r_1(t_0) \cdot r_2(t) = 0 \cdot 1 = 0$ and $t_2\!: r_1\,(t_0) \cdot r_2(t) = 1 \cdot 0 = 0$ in case of synchronous appearance, but $t_1\!: r_1(t_0) \cdot r_2(t) = 0 \cdot 0 = 0$, $t_2\!: r_1(t_0) \cdot r_2(t) = 1 \cdot 1 = 1$ in case of movement. The computational principle exploited here is "coincidence detection“, i.e. the detection of the circumstance that two events happen at the same time. However, these events are originally not synchronous, but coincide by computation.

Another obvious problem is introduced by the delay: if the delay does not exactly match the time it takes the stimulus to move from one receptor to the next, no motion at all will be detected. This problem is actually not easily overcome, but can be extenuated, e.g. by softening the hard criteria introduced above. For instance, the hard criterion of deleting all continuous parts of a stimulus can be softened by designing a filter that maximally transmits the beginning of the stimulus, but then gradually decreases while the stimulus stays the same. The portion of decrease of a step function can be manipulated by the time constant of the high pass filter: the larger the time constant, the longer the response, the shorter the time constant the sharper the response. Thus, the result of a softer filter is that a moving stimulus leaves behind a trace a bit longer so that there is still something other than zero to be multiplied when the other signal arrives. The behavior of a hard high pass filter is not likely to occur in a biological system. For example, inactivation of channels will never occur perfectly simultaneously for all channels. The occurrence of a pure delay is likewise biologically implausible. Delays are typically introduced by a gradually increasing response as it can be modeled with a temporal low pass filter. The maximum response to a step function is not reached instantaneously but at a later instant of time. Thus, this is no pure 'down time' of the system just a delayed maximum. (Think of a gradually increasing number of activated channels…)

A consequence of introducing the temporal low pass filter is that the delay is readily implied in the filters and does not have to be represented explicitly by a delay function. Depending on the time constant of the low pass filter, the activity of the system has a maximum at a delayed instant of time and decreases gradually afterwards. Thus, the low pass filter contributes to a prolongation of the signal. The time constant of the low pass filter therefore determines the tuning of the system for certain velocities, i.e. where the optimum is and how "tolerant" (broad) it is.

Even though the motion detector will always have an optimum for a given stimulus velocity, the range of velocities of moving stimuli can be extended. Detecting stimuli of other velocity ranges can not only be realized by varying time-constants, but also by varying the distance between receptors: a wider distance in an otherwise unchanged configuration would detect faster stimuli. Of course, it seems unlikely that a biological system changes its receptor spacing. But consider an array of more than two receptors: given that wider spaced receptors are connected as a motion detector, faster moving stimuli can be detected. In this manner, different velocities can be detected without changing time constants – just by correlating elements with different spacing!

The last problem considered here refers to direction. Until now, it was not considered what happens if a stimulus moved from right to left. It would be easily possible to extend the functionality of the motion detector if the same design that led to the detection of motion in one direction was applied for the opposite direction, i.e. if the design was symmetrically mirrored. For this to be realized, the signal of the right receptor has to be delayed and fed into a second correlating element. There, it has to be compared to the signal of the left receptor. However, the two correlating elements provide outputs with the same sign, so they have to be differentiated by subsequent elements or have to be sign-inversed (biologically realizable during synaptic transmission). An integrating element onto which activity about motion in one direction (say from left to right) enters with positive sign while the other direction is inversed, unambiguously responds to motion in both directions. Such a device is called elementary motion detector, EMD (Reichardt 1954).

Finally, consider a row of receptors with elementary motion detectors whose activities are all summed up by a large integrative element. This device would detect global or 'large-field' motion. For detecting motion along this row of receptors, the same device has to be built in the orthogonal direction. Usually, biological systems have a matrix of receptors (a retina) that can be thought as being organized in horizontal and vertical directions. By computing the activities of two large-field elements that integrate orthogonal directions, an overall estimate of the motion direction is generated. Consider a fly over a pond: in a patterned environment, such a motion detection system can 'tell' the fly when it rotates. It will also respond to a frog's tongue. (Of course, for detecting the frog's tongue, a mechanism of figure detection is still missing, but see Egelhaaf & Borst 1993 for an introduction).

The mechanism of motion detection was introduced here without questioning how it can be biologically realized. As it is typical for the network level of organization, it was primarily based on computational arguments. However, there is vast amount of experimental evidences that a motion detection mechanism can be realized biologically in a similar form (see again Egelhaaf & Borst 1993). The mechanism predicts very well the behavior of flies (or flies' neurons), for example, and even of single cells responding to motion stimuli. Thus, it is easily conceivable to think of the elements of the movement detector in cellular or biophysical terms: a high pass filter as a membrane with inactivating channels, a low pass filter as a membrane with resistive properties, an inversion of signs as a synaptic hyperpolarization etc. But the consideration of details complicates the understanding of the network, while the computational account makes room for understanding the principles of motion detection. It is very well possible to realize the motion detection mechanism formally and integrate it in a technical system. Computational explanations are typical for both notions of networks cellular circuit and artificial neural network (see also 1.1.3). The difference is to be found in the smaller size of circuits that allows consideration of more details.

But it is frequently assumed that also the computational explanations of artificial neural networks are consilent with neural accounts. Computational explanations mediate between functional complexity and mechanistic explanation because they allow to neglect (mechanistic) details that release cognitive capacity for understanding functional principles. Imagine, the explanation of motion detection provided above in completely neural terms – it would be multiple times longer. However, this benefit is gained at the cost of the assumption that a given function is largely independent of the concrete realization (see also 1.1.3.1).

1.2.7. Behavior

The explanations of brain phenomena considered so far covers a wide range of phenomena that can be encountered in relation to brains. Still, one important level of organization was touched only implicitly: behavior. Behavior is the one phenomenon enclosing all the constituent brain phenomena. Behavior is the first significant level of organization where the brain does not contain other elements (as the brain contains membranes, neurons, networks etc.), but is itself contained, namely by the organism that shows behavior. Thus, behavior as it is meant here does not refer to an arbitrary behavior such as the 'behavior' of a network – it refers to the behavior of a biological organism in its environment (i.e. an animal). This level of organization is relevant for an understanding of the brain because it is thought to be a phenomenon predominantly *caused* by a brain. It should be noted that a significant step is made in the succession of levels of organization. The brain is put in a functional shell of the organism. Thus, by taking into account the behavioral level of organization, the brain is solidly embedded in nature's established categorical system of animals, flowers, earth, life etc. that does not necessitate dissection, but is overt and evident for everyone. The behavioral level of organization entrenches the brain by mediating an immediate and intuitive understanding of what the brain is all about, what function the brain has: the brain enables an animal to align with its environment.

It might seem surprising that it is suggested to consider behavior as the relevant characteristic of possible levels of organization upward from the brain – the organism might seem more appropriate at first sight to constitute the next level. But some aspects make the organism a sub-optimal choice as an exclusive level of organization. First, the organism as the sum of all bodily function (e.g. digestion, extremities etc.) is not the appropriate functional successor of the brain. The organism might be a container for the brain with respect to spatiotemporal organization, but not with respect to causation. It is not the organism that determines how the brain relates to the environment – if there is causation of adaptive behavior in the sensorimotor loop, then it is attributed to the brain. In other words, brain and behavior are causally more directly coupled than brain and organism or organism and environment. Of course, the brain needs to be embedded in motor and sensory systems, but the control of changes imposed to the environment is thought to reside in the brain (see Chiel & Beer 1997 for a detailed discussion). It should be noted, however, that this functional account of the levels of organization elicits the metaphoric aspect of the term brain (see also 1.1.1): organisms without brains can very well behave adaptively and nervous, motor or sensory structures can equally well contribute to adaptive behavior. The brain as it is understood here is, by definition, the cause underlying the generation of adaptive behavior. (Of course, evolution is a cause of adaptation, but the brain is the immediate cause of behavior.)

The consideration of causal relations between brain, motor and sensory systems, behavior and environment already shows that the definition of borders between concepts is not trivial. But there is more: for example, the problem of defining internal and external factors. Generally, an environment can be described as a situation the animal resides in. Particularly illustrative, intuitive cases are provided when external, environmental factors directly determine the situation of the animal such as the environment of the fly being determined by the frog's tongue. For their intuitive appeal, external factors are frequently overestimated and internal factors determining the situation are commonly underestimated. (With respect to humans this relation might be inversed.) But internal factors definitely determine the situation, too. Consider the energetic state of the fly: if the energy stores are empty and the muscles do not work optimally they might determine success or failure of behavior. Similarly, if the fly had no appropriate sensory system to extract motion information (e.g. because temporal resolution is too low to detect the fast tongue), adaptive behavior might not be realizable. So, where to draw the line between internal and external? Is it really the frog's tongue itself that immediately determines the behavior or is it the fly's internal representation of the frog's tongue? Some internal representations of external elements or activities are directly caused by external activities, such as the frog's tongue. Other internal representations are not immediately caused by external factors. The energetic state of the fly, for instance, is determined by nutrition uptake that is determined by external factors, but it has undergone considerable changes while it was transformed by digestion processes that turned it into an internally caused factor. Sensory representations of the frog's tongue could be compared to the stage of chewed nutrition (an externally caused factor) but not to glycogen saturation in the muscles (an internally caused factor). This example illustrates that the type of transformation performed on externally caused factors somehow determines whether or not it becomes internal.

With respect to the control of behavior, it is the brain that performs these transformations on externally caused factors, sometimes adds internally caused factors and eventually generates a new internally caused factor that is applied in behavior. Therefore, it is the brain that draws the line between organism and environment for behavioral issues. Since the organismic level of organization is soaked with environmental factors on the sensory side and molds the environment on the motor side it is a rather indirect approach to the explanation of behavior. Nevertheless, it is important to find the interfaces between environment, sensory and motor processes in behavior. The brain is the most potent approach that can be taken to define these interfaces because it causes and controls the generation of adaptive behavior.

1.2.7.1. *An example: Phototaxis*

In order to demonstrate explanations of brain phenomena for behavioral issues, consider the very simple behavior of phototaxis, i.e. directed motion of an organism that relates to a light source. The simplest scenario necessary for explaining phototaxis is an environment that contains an organism. Seen from the perspective of an observer, the environment contains external activities, i.e. visual stimuli. For realizing phototaxis, visual stimuli must somehow lead to motor activities that propel the organism towards the light source – they must be processed in between uptake and delivery. Principally, the light source can be conceived as a brightness contrast, i.e. there must be a bright zone and a dark zone. Thus, the organism must be able to compare an activity relating to a bright zone and another an activity relating to a dark zone. Most easily, this can be achieved by an organism with at least two equally sensitive receptors that look at different locations. The strength of a receptor's activity differs depending on the brightness of the location it represents. The difference between the two externally caused activities indicates the direction of the brighter location. Given that brighter stimuli cause stronger, positive activities, for instance, subtraction of the activity of the right receptor from the activity of the left receptor provides positive values for a brighter left location and negative values for a brighter right location. Thus, the activity of a subsequent unit that integrates both values could represent the possible turning direction for approaching the light. However, an explicit representation of the difference is not even necessary if it is assumed that the receptor signal directly effect motor elements – one standing for movement in one direction, the other on standing for movement in the other direction. Given that the stronger signal causes stronger motor activity on the same (ipsilateral) side (i.e. does not cross to the other, contralateral side of the body), the organism moves to this side (consider a fly increasing its flap frequency on one side, for example).

Is it also possible to design an organism with just one receptor that can do phototaxis? For a creature with only one receptor such an operation is not trivial. The only way to achieve the two contrasts necessary is to transform the spatial contrast into a temporal contrast by evaluating the time course that arises from self-motion, for example when the organism scans the environment for brightness contrasts by moving the receptor. A possible computation is the differentiation of the receptor signal and an evaluation of negative values: as long as the differentiation of the receptor activity stays constant or increases, the brightest point is not reached. But when the differentiation becomes negative the brightness contrast decreases and the scan must be stopped. Since the position of the brightest point encountered in the scan was reached an instant before the scan, motion in the direction of the actual receptor position is a good guess. Alternatively, the motion can be directed to the point the receptor was before, i.e. be corrected. However, when it moves and the brightness (and the differentiation) decreases, it will change its direction anyway. This question for a realization of phototaxis with only one receptor illustrates that a sequential system with just one pathway that is parsimonious in the number of elements imposes constraints for complicated computations. The benefits of investing a few more elements to allow for parallel distributed processing pays off in offering much simpler computational solutions.

Phototaxis does not have to be positive, i.e. yield motion towards the light – there is also negative phototaxis when organisms turn away from the light source. Altering positive phototaxis to negative phototaxis can be easily achieved by crossing the connection from receptor to motor. A stronger activity from the other (contralateral) side results in a turn away from the light source.

A creature with this configuration is widely known as a simple Braitenberg vehicle named after the author (Braitenberg 1984). It shows behavior in the sense that it suffices the minimal needs for explanations of organismic, individual behavior, i.e. it applies the same scenario containing environment, individual, stimulus, mechanisms etc. Moreover, the Braitenberg vehicle offers a mechanism of phototaxis that can be well imagined to be realized in a biological organism. However, a Braitenberg vehicle definitely is no organism because the realization of the behavior differs substantially from the organismic realization: apart from all biologic functionality (autonomy, reproduction etc.), the molecular realization of the control mechanism of behavior is missing. So, does it really help to compare an organism with a vehicle? Yes and No! Taking into account the vehicle helps as long as the differences in realization do not determine functionality, i.e. as long as long as organism and vehicle behave the same. Of course, this functional equivalence is extremely dependent on the situations taken into account. If only simple cases are considered, the behavior will perhaps be the same in organism and vehicle. But, if a behavior is extensively tested, differences in the behavior will presumably appear. If these differences are to be explained in terms of the differential mechanisms on differential levels of organization, the use of the vehicle for explaining the behavior ends. The vehicle can be refined with new models to achieve functional equivalence again and the procedure can begin anew. Since in praxis always differences will appear, a consequent application of this comparative procedure would inevitably lead to a vehicle that accounts for phenomena on all levels of organization, i.e. systems, networks, neurons, molecules and atoms. However, such a simulation is only advisable if the vehicle still shows benefits in the form of clear or illustrative explanations. Thus, the explanatory target (the concrete question concerning the phenomenon to be explained) determines if the vehicle is helpful or not (see also 1.1.12). The scope of the simulation (e.g. Braitenberg vehicle) for explaining brains ends where mechanisms that realize organismic behavior (primarily molecular mechanisms) are as comprehensible as the simulation.

1.2.7.2. *Further behavioral issues*

The consideration of the Braitenberg vehicle demonstrates that behavior, as all other brain phenomena, can be explained on different levels of organization. Narrowing the explanatory aperture reveals brain mechanisms on different levels of organization that realize the behavior. Widening opens the view for a purely functional approach. Considering the behavioral level of organization establishes an explanatory interface between brain phenomena and function, which is essential for brain sciences. Without this interface, brain studies are endangered of being *l'art pour l'art*. This becomes particularly clear when evolutionary questions are taken into account. Evolution encircles brain sciences from two sides: on the behavioral and the molecular level of organization. Behavioral, because evolution theory provides means for determining the functional value of a given behavior and therewith for brain mechanisms. Often it seems difficult to assess the value of a given brain mechanism: whether a certain synapse is inhibitory or excitatory appears to be an arbitrary question if considered isolated. But, if this synapse is decisive for showing positive or negative phototaxis, a functional value is added. If, additionally, it is taken into account that the kind of orientation leads the organism to nutrition or keeps it from being caught by a frog's tongue (i.e. influences the fitness of the organism) the functional value is immediately clear. Evolution theory provides means to quantify these values and translate it into molecular mechanisms (genetics) that are placed on the other (low level) side of the scope of brain science. In evolution theory, phenomena on the molecular level of organization are explained on a larger timescale than the 'real-time' typically taken into account in brain science, i.e. ontogenetic and phylogenetic timescales. Thus, evolution theory extends the focus of brain science primarily temporally.

Behavior is not only constrained by evolution and low-level neural mechanisms, but also can depend on cognition. Cognition can be conceived as a specific group of internal factors that determine behavior. In this conception, cognition is directly traceable experimentally by analyzing the effects cognition has on behavior. The most common paradigms in cognitive science are performance tests on humans that measure reaction times or the quality of verbal reports in relation to certain treatments (environmental situations). Explanations resulting from these paradigms are largely independent from brain mechanisms. However, since cognitive factors are internal factors caused by the brain (see above) they are explainable in terms of brain phenomena. Whether the explanation of cognitive phenomena in terms of brain phenomena is adequate, depends, again, on the question posed: as long as an explanation in pure cognitive terms suffices, there is no need to mobilize brain mechanisms.

Considering brain mechanisms for explaining cognitive phenomena can be helpful to describe the transition between cognitive and non-cognitive phenomena: for something to be cognitive, internal representations (no matter if externally caused such as the frog's tongue or internally caused such as a dream) are transformed by the brain in a way nutrition is digested by the digestive system. The result of this transformation is an internal representation of the situation the animal resides in (a 'model'). If (and only if) another transformation is performed on this model (of the situation the animal resides in), the situation the animal resides in is no longer caused by the environment – then it is caused by the brain. A behavior rests on cognition, thus, if it can be traced back to a model of the situation that was transformed in order to generate the behavior. The fly would have cognition in case it can be shown that its escape behavior rests on a model of the situation that was transformed (e.g. evaluated) in order to generate the behavior. Usually, it is not assumed that the fly has cognition because it is sufficient to evaluate internal representations of an externally caused factor (i.e. direction and velocity of a moving object) to explain the fly's behavior. (And, applying 'Occham's razor' according to which parsimonious explanations are superior, no superfluous assumption should be made in scientific explanations.)

In an overall view, the behavioral level of organization integrates the levels-of-organization framework underlying explanations of brain phenomena by determining its function: brain phenomena cause adaptive behavior! At the same time, behavior provides explanatory interfaces between the brain and further levels of organization outside the brain (e.g. evolution) and inside the brain (e.g. cognition).

1.2.8. Learning

Explaining learning as a brain phenomenon demands consideration of all levels of organization that are usually taken into account in brain studies, i.e. molecular, cellular, networks etc. Plasticity, the neural basis of learning, was already introduced in the context of synapses. Plasticity denotes a general characteristic of systems, while learning as a brain phenomenon is typically assessed with reference to a changed behavior of an organism. But learning is considered as being on a further level of organization than behavior because it exceeds real-time (the immediate time span it takes to generate adaptive behavior) and shall conclude the explanation of specific brain phenomena. Learning is a phenomenon in the transitory area between cognitive and non-cognitive phenomena. A case of non-cognitive learning is given when externally caused elements or activities (e.g. the tone in the Pavlov paradigm) become internal representations that are applied in the generation of a behavior, even though the externally caused activity is not present in the situation the behavior is generated. This learning is non-cognitive because it is not necessary to transform the internal representation, but rather to apply it. A case of cognitive learning would be given if a behavior is only explainable by assuming that a learnt representation is transformed before or during the generation of behavior, e.g. if two concurrent stimuli are evaluated in a situation model. The criterion for cognition of transforming the representation of a situation critically depends on the concept of transformation. For achieving a more precise notion of transformations they can be described as a set of procedures or rules operating on representations, short ‘computations’. In this sense, the concept of computations can also help to clarify cognition: the kind of computations made can determine if a behavior rests on cognition or not. However, developing such criteria in depth is not a primary objective in the present context. The comments shall rather illustrate that the concepts brains, learning, computation, cognition etc. can be arranged meaningfully by relating them to behavior.

1.2.8.1. *Conditioning*

Multiple versions of explanations of learning can be construed, ranging from purely formal over behavioral versions to those incorporating brain mechanisms. Here only a very basic example shall be introduced, again.

Since, unfortunately, a fly escaping a frog's tongue does not learn that much, another case has to be chosen. Consider the simple example of conditioning (Pavlov 1904) and the well-known case of conditioning of the eye blink response: a human subject or an animal (frequently a rabbit) receives an air puff directed to the cornea that elicits a blink reflex. This response is thought to be a protection against noxious stimuli. If the air puff is repeatedly preceded (e.g. after 20 trials) by another stimulus, e.g. a tone, this stimulus will eventually by itself elicit a blink even before the air puff is applied. In this example, the eye blink is the response (R), the air puff is the unconditioned stimulus (US) and the tone is the conditioned stimulus (CS). Thus, the principle of conditioning is to modify an existing behavior 'air puff causes eye blink' (US > R) to 'tone causes eye blink' (CS > R) by associating air puff and tone (CS<>US). Here, learning takes place only at one specific instance, namely the generation of the association between air puff and tone (CS<>US).

On the behavioral level of organization, the quality of learning can be assessed by defining the cases 'tone causes eye blink' (US > R) as success and then evaluating the frequency of successful trials. At the beginning of the learning procedure, the tone alone will not elicit an eye blink. After some trials the tone alone might be already successful in single trials. After more trials the tone alone might lead to success in almost each trial. This learning process yields an individual learning curve that shows rising performance with time or, better, trials. Making a lot of these experiments and pooling the gathered data yields an expressive learning curve that describes the learning process as such. Of course, various factors influence this learning curve and for obtaining a valid general learning curve, the conditions in the experiments determining the individual learning curves must have been approximately identical (i.e. strength of the stimuli, timing of the stimuli etc.). Assume that the 'decision' whether an eye blink is elicited by the tone alone is determined by a single connection (or switch) being activated. Then, learning is to be understood as an increase in the probability of this connection being activated. Applied to the stimuli (air puff<>tone), this means: the connection between CS and US sometimes functions and sometimes fails with a given probability. The actual association strengths between CS and US represent this probability. Under this premises, a value of 0.5 in the learning curve means that in 50% of the cases an eye blink is elicited by the tone alone and the US-CS-connection is activated.

1.2.8.2. *Formal explanation of learning*

The overall learning process depends on several factors. An important factor is the maximum value that can be approached, i.e. the highest probability of successful trials. Even after extensive training some trials still might fail so that the maximum value is under 100%, e.g. $V_{max} = 0{,}9$. The amount of learning occurring between two trials Δ can be described as the difference between the maximum value and the learning that has already taken place in the previous trials $\Delta\ V(n) = V_{max} - V(n-1)$. This Δ represents the missing amount of learning or the 'error' that is still in the learning process.

Furthermore, the specific character of the stimuli is important for the learning process. As already mentioned above, the 'salience' of the stimuli (e.g. strengths of the air puff, frequency of the tone) can influence the probability of an eye blink response by a factor K that is also between 0 and 1. Alternatively, the constant K can be separated in α for the unconditioned stimulus and β for the conditioning stimulus: $K = \alpha\ (US) \cdot \beta\ (CS)$. The learning curve can be described in terms of the different factors by $\Delta V(n) = K \cdot (V_{max} - V(n-1))$. Since every trial yields a certain increase in learning $V(n)$, the curve approaches V_{max} - but only the amount that the stimulus saliency K allows. This description is a simple version of the Rescorla-Wagner model (Rescorla & Wagner 1972). It allows not only for the modeling of classical conditioning but also for various other conditioning phenomena such as 'blocking' (primary US already learnt prevents secondary US from being learnt) or inhibition (the absence of a US is learnt).

The principles of the Rescorla-Wagner model are also applied in the field artificial neural networks: the so called 'delta-rule' (see e.g. Rumelhart *et al.* 1986) is an equivalent of the Rescorla-Wagner model that is applied to the weights of artificial neurons. ('Weight' is just the technical term for connection strengths between units.) They determine the magnitude of the effect that an element x (a given input signal) has on an element y: $y = w \cdot x$. The artificial neurons are organized in layered networks. For illustration of the function of such a network, consider an industrial robot that controls if a panel with six eggs is completely filled and that has to supplement missing eggs. The computations necessary for accomplishing this task can be achieved by teaching the network to associate all possible input patterns to a correct output pattern. For instance, the network has to learn to respond to an input pattern 1-0-1-1-1-0 with an output pattern 0-1-0-0-0-1. Several combinations have to be stored in the network. The delta-rule is a means of teaching the network: the connection ('weight') between an input element *i* and an output element *j* (w_{ij}) is changed by adding the actual Δ to the weight of the previous trial $w_{ij}(n) = w_{ij}\,(n - 1) + \Delta(n)$. The constant *K* from the Rescorla-Wagner model is usually conceived in a slightly different way, namely as determined by a learning rate ϵ and the input value x_j: $K = \epsilon \cdot x_j$.

The differences between Rescorla-Wagner model and delta-rule are primarily to be found in the intention of modeling and in the organization ('architecture') of the system. The Rescorla-Wagner model was developed along experimental, behavioral data and usually refers to one or a few connections (stimuli), while the delta-rule is applied to train artificial neural networks with multiple units or massive connectivity.

As can be concluded from the short account provided above, learning can very well be explained and formally described solely on the behavioral level of organization without reference to the brain mechanisms. The iterative observations of the behavior and the change in performance ('error') yield self-evident explanations and do not necessarily call for mechanistic-causal explanations. However, explaining learning as a brain phenomenon calls for other levels of organization. For the eye blink case, the general claim has to be that somewhere in the brain a connection between the tone and the eye blink has to be established.

1.2.8.3. *Neural Explanation of learning*

An explanation of learning on the neural level of organization demands an application of most of the brain phenomena considered so far. The following presentation is intentionally highly condensed in order to provide an example of what applied cases of explaining brains might look like. (Most explanations given so far presumably were quite lengthy for a brain expert). Consider the following premises as given (see fig. 10): (1a) the air puff, detected by sensory cells in the eye, is processed on a reflex path (via the trigeminal nucleus) over the brain stem (the reticular formation) to motor centers where the eye blink is generated. (2a) The tone is processed in auditory nuclei all over the brain. (3a) Motor learning can take place in the cerebellum. For learning to take place it must be given that participating regions are interconnected: (1b) The trigeminal nucleus is connected to the cerebellum via ‘climbing fibers’. (2b) The auditory nuclei are connected to the cerebellum via ‘mossy fibers’. It can be shown that electric stimulation (CS) of the mossy fibers paired with air puffs (US) leads successfully to conditioning. (3b) The cerebellum is connected to the cranial motor nuclei for triggering eye blink generation. These exemplary evidences indicate that the circuitry is apt for realizing the connection between tone and air puff. The place of the connection, however, can be anywhere between the location where the pathways of the stimuli ‘meet’, i.e. in the cerebellar region and during generation of the eye blink (i.e. the cranial motor nuclei). Since (reversible) chemical inactivation of the cranial motor nuclei prevents conditioning, the location of learning is concluded to be the cerebellum. How is the connection realized on the cellular level of organization? The Purkinje cells project in the direction of the motor nuclei, i.e. are the output neurons of the cerebellum. The dendrites of the Purkinje cells are connected to both climbing fibers (involved in the air puff processing) and mossy fibers via parallel fibers (involved in tone processing). Synchronous activation of climbing and parallel fibers is thought to lead to a depression of the Purkinje cell that unblocks (releases) the connection between the two stimuli.

On the sub-cellular level, the connection is thought to rest on a molecular mechanism called long-term depression. Principally, temporally correlated presence of the two stimuli (tone and air puff) leads to synchronous activation of the transmitters at two sites presynaptic to the Purkinje cell dendrite, namely at the climbing fiber and at the parallel fiber. The resulting synchronous activations of metabotropic receptors on the postsynaptic membrane of the Purkinje cell result effectively in a persistent increase in the concentration of the enzyme Protein Kinase C (PKC). The connection of the stimuli is brought about by the condition that the effective activation of PKC in the Purkinje cell dendrite depends on synchronous activities on both synaptic sites: cascades of intermediate activities at the parallel fiber synapse and calcium influx at the climbing fiber synapse. Since, it is known that PKC is a major regulatory enzyme and plays an important role in signal transduction as well as cell growth, differentiation and gene expression, the proposed mechanism for the establishment of the connection between the two stimuli focuses on PKC.

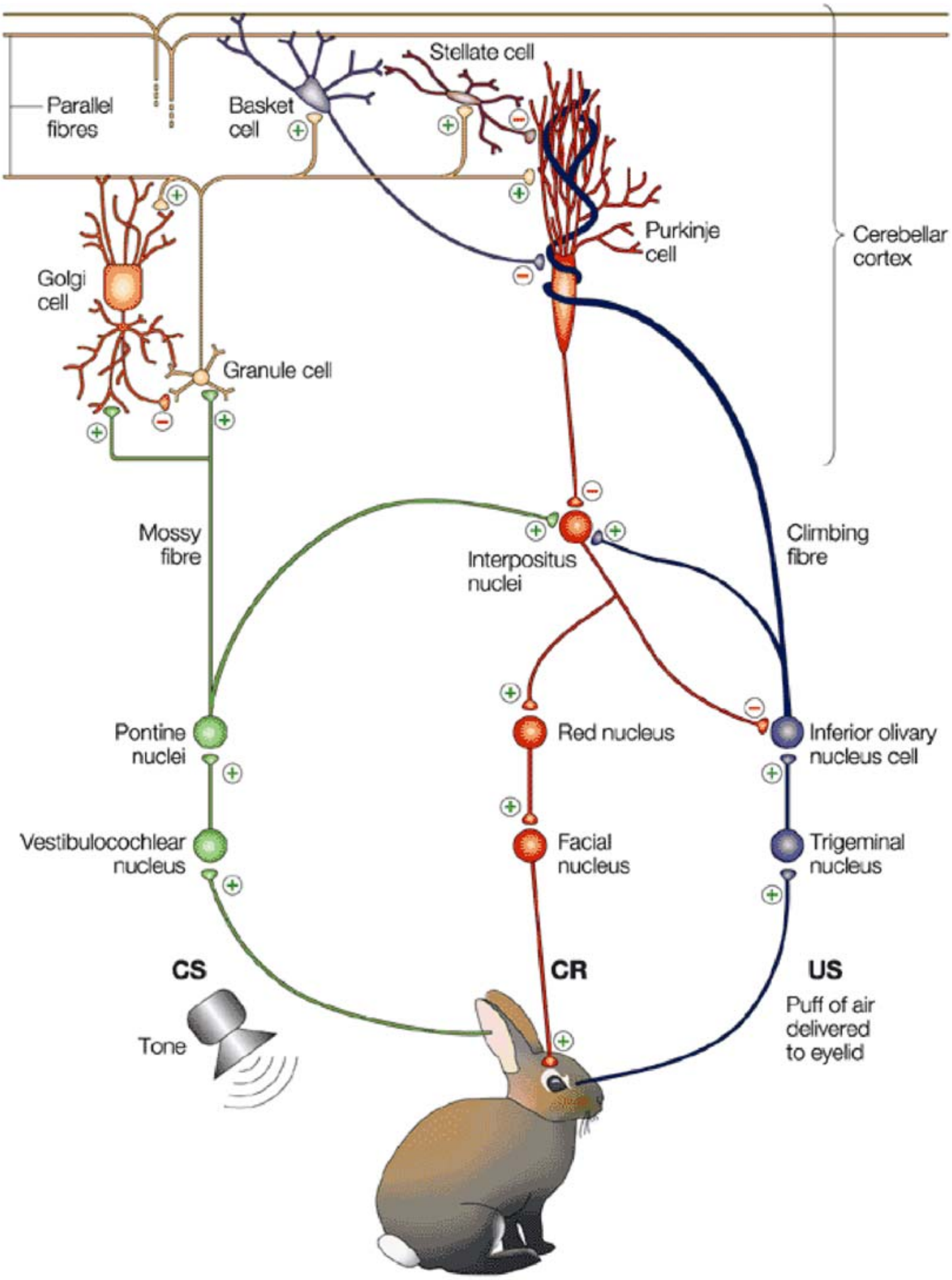


fig. 10: ***Eye-Blink Circuit.*** *Eye blink conditioning. The rabbit responds to an air puff (unconditioned stimulus, US) with eye blinks. When the US is repeatedly preceded by a conditioning stimulus (CS) such as a tone, the eye blink is eventually elicited by the CS alone. Shown is the neural circuit hosting the mechanisms underlying the association. For details see text. © The Mauk Lab (permissive license).*

Learning, as exemplified by conditioning the eye blink response, is a brain phenomenon that spans from the behavioral to the molecular and electric level of organization. Furthermore it extends from microseconds over real-time to ontogenetic time (or even phylogenetic if the causation of gene expression and fitness value is considered). Therefore, a conclusive explanation must consider far more evidence than only the few of the vast number of evidence existing for the realization of the eye blink conditioning that were taken into account above. But for serving the present illustrative purpose - introducing the principles of learning as a behavioral brain phenomenon - this simplification should be allowed. Applying the explanatory construction for brains on any level of organization will unfold an almost arbitrary large complexity underlying the phenomena of learning. Experimental evidence, for instance was hardly taken into account. The experimental machinery behind evidences for mechanistic explanations of brain phenomena sometimes appears to be very remote from a behavior of an individual organism. For example, if someone investigates the regulatory capabilities of PKC, it is not immediately evident that this can be relevant for understanding eye blink conditioning. This difficulty of conceiving simultaneously the explanation of the phenomenon and the experimental procedures demands simplification strategies. Complex phenomena that incorporate explanations on many levels of organization and extend over larger timescales – such as learning – are primarily explained with strong simplifications, only taking into account such activity on such levels of organization that is indispensable for explaining the mechanism. More detailed aspects are omitted, not because of their irrelevance, but because they overload the explanations. As a consequence the explanatory focus ‘jumps’ from behavior to molecules to systems and from microsecond to days to minutes. In this way it can be understood how the “strange brew” of levels of organization often found in explanations on brains emerges.

1.2.9. Summary and conclusion

Brain phenomena play a crucial role in explaining adaptive behavior mechanistically. The explanations apply a specific explanatory framework that is dominated by the implicit notion of levels of organization. The elements that play the leading part in the explanation of brain phenomena are neurons. Therefore, the cellular level of organization is a common checkpoint in explanations of brains. One essential property of neurons is their responsiveness: external (environmental) elements and activities that demand an adequate (adaptive) behavior cause correspondent activities in the brain. These responses can be passed to elements downstream in processing that also respond. In this way, responsiveness makes possible that representational activities cascade from reception over processing (evaluation) to a possible reaction through the organism. Responsiveness can basically be explained as an electric phenomenon that takes place at the neuronal membrane. Electric accounts elicit the weakness of membranes to conduct activities efficiently so that 'workarounds' such as spikes as amplification mechanisms play an important role in processing responses. Even though the principles of processing activities without spikes (on a 'passive membrane') or with spikes (on an 'active membrane') are not too complicated, the possibilities of processing explode already when the cellular level of organization is reached. Again, dynamics and complexity resulting from the spatiotemporal and causal variability prevent the formulations of a general account. Therefore, processing is often conceived in a simplified framework: the 'classical neuron'. It performs spatiotemporal integration of activities collected from other neurons in dendrites that have a tree-like shape and evaluates the incoming activity by a threshold procedure at the spike initialization zone. Whether or not the activities are processed further depends on their ability to elicit a spike (when the activity is strong enough). If a spike is elicited an activity is processed further and can be passed to the next neuron. The mechanism by which the activity is passed to the next neuron is called synaptic transmission. Synapses are functional units that extend the range of processing spatially, temporally and causally (e.g. by inverting the sign or the form of the signal). The resultant level of networks of neurons is characterized by the distributed processing of several activities in parallel and is therefore thought to be the level where multiple representations (e.g. resulting from different elements and activities in the environment or different instances in time) are combined, compared and evaluated in order to generate adaptive behavior. Behavior is at the interface between brain and external world and allows for drawing further conclusions on the function of brain phenomena, e.g. if the brain phenomena are cognitive or have evolutionary significance. Learning extends the significance of brain phenomena beyond the actual behavioral context in a domain of possible behaviors that are themselves represented as brain states.

In this section, the explanatory framework provided in the previous section was applied to some concrete brain phenomena. At the beginning, it was stated: "Brains are no easy things to explain". But the systematic application of an explanatory framework ('levels of organization') to specific brain phenomena illustrates that the constituent brain mechanisms are fairly easy comprehensible. Complex phenomena become intelligible by applying simplifications such as the 'classical' or the 'binary' neuron. These simplifications can be compared to a 'packaging mechanism' that functions by considering a certain phenomenon as being self-contained. The underlying complexity is simply neglected for the moment and is only considered if the situation affords it. In this way, a phenomenon can be checked off the list of comprehension problems and, thus, releases cognitive capacity for thinking in a greater context. The problems of explanation arise primarily when the adoption of simplifications such as the 'classical' or the 'binary' neuron are refused. Then, too many levels of organization are taken into account and the unfolding complexity (or dynamics) might render the explanation completely unintelligible. Of course, in specific cases the consideration of detail is mandatory, e.g. when inconsistencies are detected or new simplification mechanisms shall be developed. And what level of detail is appropriate? The significance of a given detail is only determined by the question posed, i.e. if it is needed to produce a sound explanation. In this sense, there is no general rule for explaining brains – there is just an explanatory construction kit and some operating instructions.

Reading Advice

In the previous section, the first Brain-Special provided an extended insight into explanations of brain phenomena that illustrated the influence of dynamics and complexity. This will be important for understanding the role of simulation in explaining brains that is further analyzed in the next chapter. But first, the second Brain-Special in the next section will have a good look at the consensus on explanations of brains (as it was appealed in the first Brain-Special). The various disciplines contributing to explanations of brain phenomena are analyzed with respect to the commonality in their underlying 'ontology' that becomes obvious when their information resources such as databases, thesauri and textbooks are scrutinized.

1.3. "Where to put information on brains!?"

"Brain Theory" has become an integral part of the scientific and even the public understanding of nature. But the notion of a uniform framework behind explanations of brain phenomena might be deceptive. In everyday practice, the various disciplinary approaches to the brain sometimes rather appear as an ever-growing amount of scattered explanatory fragments than as a homogenous 'theory'. Consider a student of brain sciences sitting in front of a library computer: a query on the term 'adaptation' in the literature database might bring to light thousands of hits from hundreds of specialized disciplines - each one with special, if not contradictory, assumptions about adaptation. The following analysis mediates an impression of this variety of approaches to the brain. It will be accomplished by analyzing existing information resources on brains such as databases and textbooks. The purpose of this study is to test the integrity of the brain sciences. Differences and commonalities of conceptual frameworks in specific approaches to the brain shall be elaborated. The differences call for strategies to handle them and the commonalities should help to filter out a common conceptual framework for the brain sciences. For not becoming top theoretical also a practical application shall be envisaged. In the student's database query mentioned above, a 'brain navigator' would be helpful to guides the student through the search results, suggests ranking of relevance, gives short tutorials, shows where to put gathered information in the records etc. But is there a common conceptual framework a brain navigator can be based on? This depends on the state of theoretical integration in the brain sciences. Thus, the brain navigator serves as a hypothetical benchmark test for the state of theoretical integration in the brain sciences?

1.3.1. Introduction

'Neurobiology', 'artificial neural networks', 'cognitive psychology' and related research areas obviously have something in common. All these disciplines seek to contribute to explanations of brains. However, even though there is a common explanatory target and despite political proclamations such as "the decade of the brain" (Bush 1990), or large scale programs such as "the human brain project" (Shepherd *et al.* 1998; Arbib & et al 2001; Koslow 2001), a unified brain theory is not within reach. Rather, numerous highly specialized disciplines raise their voice in a chorus of explanatory fragments.

This is not necessarily a fundamental problem: it seems reasonable to assume that, when research has generated enough knowledge, all fragments will automatically assemble to a well-formed composition. At the moment, however, brain sciences appear to be much more differentiated than integrated. This is not so much a difficulty for researchers working in one of the specialized disciplines. These specialists have their own explanatory frameworks and manage it very well. But the lack of a common conceptual framework is a problem for other groups: first, for people who commute between the approaches i.e. interdisciplinary workers and, second, for novices of brain sciences. This lack may lead to a weak public understanding of the brain. If these groups ask cross-border questions such as "In which sense does an artificial neural network have a memory?" they might receive as many different responses as there are sub-disciplines: a neuropharmacologist could try to explain memory with the analogy of influences of pharmaceutical treatment, the computer scientist by forwarding storage routines of programming languages and the neuropsychologist by talking about diagnostic criteria of traumatic defects. This condition is certainly not an indicator for solid explanatory quality of brain phenomena!

In general, an ambiguity is evident: on the one hand, there is a ubiquitous supposition of a common explanatory target, i.e. the brain. On the other hand, there is no commitment to a common conceptual framework. The general objective in the background of this study is therefore to find a common conceptual framework for explaining brains. However, in absence of a general brain theory that would allow us to deduce a conceptual framework in a top-down manner, an 'empirical' method that analyzes existing information resources on brains (textbooks, databases etc.) in a bottom-up manner is applied. First, existing resources are analyzed with respect to their classification systems (e.g. keywords in database structures). The grade of variance and redundancy between the specific classification systems can serve as an indicator of integrity of brain sciences. The current state will be assessed by attempting to integrate the different resources into a single account. A single, consistent and comprehensive classification system for the brain would bring us several steps closer towards a common conceptual framework. In the example of the student of brain sciences sitting in front of a library computer such a unified classification system would mean that the student does not have to switch between different search engines, glossaries, data records etc., but would have an integrated access (i.e. the brain navigator).

In the second part of this section, existing resources on brains are analyzed with respect to their underlying specific conceptual frameworks (e.g. in textbooks). Consistencies between the accounts can serve as clues for constructing a common conceptual framework, inconsistencies as warning signs for difficulties of theoretical integration. A common scheme for the brain sciences as an early stage of a common conceptual framework shall be proposed. For the student's brain navigator, this would imply some kind of standardized tutorial of the brain. But first, for not getting lost in the 'jungle' of disciplinary variety with all their specialist accounts, a roadmap will be sketched.

1.3.2. A sketch of the brain sciences

Is there a term that subsumes all the disciplines that are concerned with the brain? It seems as if there were arguments to be found easily against any term. The term 'brain science' is somewhat void in that it does not present a certain phenomenon under scrutiny but rather an anatomical structure. The term 'neuroscience' comprises a wealth of the brain sciences, but it might come across as being insolent to silently include all cognitive aspects in the neural domain that might deserve their own credits. The term 'cognitive neuroscience', on the other hand, narrows the focus too much on cognition. The construct 'neural and cognitive sciences' misses the computational aspect and, hence, underrates certain theoretical approaches (neural networks theory) as well as practical applications (i.e. neural engineering and software engineering). The term 'Neuroinformatics' encompasses both these technical aspects and the theoretical aspects exemplified by artificial neural networks[3]. Taken together and facing the facts there is no single term properly representing all disciplines around the brain. Therefore, the term brain sciences is used here in a metaphorical sense – not simply referring to sciences that study the brain as an anatomical structure, but including all disciplines seeking to explain adaptive behavior of animals (including humans) with respect to brain phenomena.

[3] Interestingly, the relatively new term 'Neuroinformatics' was adopted by the Human Brain Project at the NIMH (Koslow 2001; Shepherd *et al.* 1998) and a thematic network from the European Union 'computational neuroscience' (Schutter 2002).

table 1: A collection of disciplines around the brain. *Only disciplines that incorporate one of the three aspects 'neural', 'cognitive' or 'computational' were taken into account. A total sum of 1,0 was split up between the three domains indicating the degree to which the domains are reflected in the respective discipline.*

	Neural	cognitive	computational
Cognitive Neuroscience	0,4	0,4	0,2
Computational Neuroscience	0,4	0,2	0,4
Neurobiology (Biophysics, Neurophysiology)	0,8	0,1	0,1
Neuroethology	0,7	0,2	0,1
Cognitive Psychology (Theoretical Psychology)	0,2	0,6	0,2
Connectionism	0,2	0,3	0,5
Neurophilosophy	0,4	0,5	0,1
Philosophy of Mind	0,2	0,7	0,1
Artificial Intelligence	0,1	0,3	0,6
Cybernetics (System Theory)	0,1	0,1	0,8
Neuroinformatics	0,3	0,2	0,5
Neural Engineering	0,3	0,1	0,6
Robotics	0,1	0,1	0,8

The search for a name already indicated that there are three major domains to which brain phenomena can be attributed: the 'neural', the 'cognitive' and the 'computational'. A collection of disciplines from these domains is shown in table 1. Since a comprehensive collection could begin with physics and end with religion, it appears to be sensible to restrict it. The criterion for 'relevant' disciplines is that the properties 'neural', 'cognitive' and 'computational' all have to be a valid classifier for the subject of the respective discipline. If one classifier does not apply (i.e. if it is outside the circle in the inset of fig. 11) the discipline is not considered. The degrees to which the classifiers apply to a given discipline vary and, thereby, determine the position in the 'landscape' of the brain sciences. Each one is assigned a value that indicates the degree to which a given classifier applies. (Since this is done only for illustrative purposes, the affiliation was simply subjectively rated than determined by laboriously defined objective criteria.) The result is normalized by the constraint that the total must amount to 1,0. The splits are summarized in table 1. For example, the case for computational neurosciences can be read as: "Computational Neurosciences is concerned with 40 percent neural, 20 percent cognitive and 40 percent computational issues." The point resulting from these proportions is plotted in a coordinate system that is constituted by the three classifiers. The results provide a sketch of the brain sciences (see fig. 11). It illustrates the disciplinary variety and shows that the affiliation of a discipline can be applied meaningfully to arrange a disciplinary landscape. The three domains ‘neural’, ‘cognitive’ and ‘computational’ are advocated by three exemplars of scientific disciplines: Neurobiology, Cognitive Psychology and Neuroinformatics, respectively, because they all are characterized by showing a clear commitment to one of the domains, but are simultaneously affiliated to the remaining two domains.

1.3.3. Classification methods

The general sketch of the brain sciences provides an orientation in terms of disciplines and helps to develop a coarse orientation in brain sciences. However, disciplines do not contain explicit statements about the conceptual frameworks of the brain sciences. For example, disciplines are not very telling for a novice student on the library computer to file the information, e.g. search results. Generally, other classification systems are applied. But what do classification systems for information on brains look like?

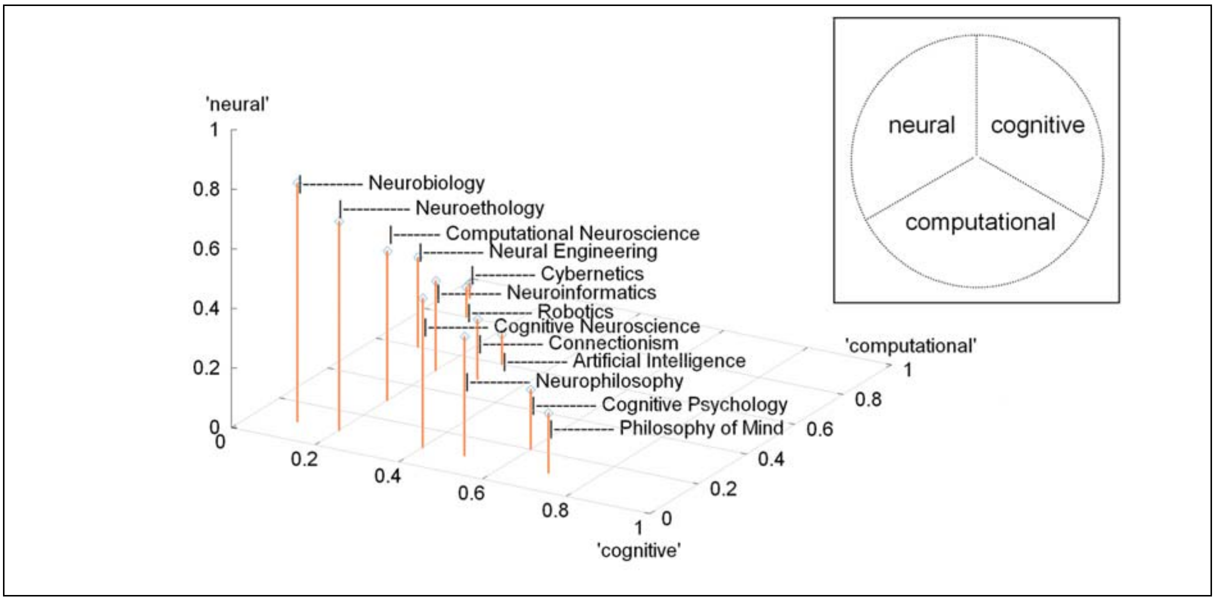


*fig. 11: **A sketch of the brain sciences.** The locations of disciplines as determined by the classification used and explained in table 1. For clearer presentation, bars (rather than ‘vectors’ originating from coordinate 0,0,0) were chosen to indicate from where and how far disciplines protrude into the neural domain. Only disciplines for which each of the classifiers applies were taken into account as indicated by the circle in the inset in the upper right corner*

For a better understanding, some general remarks on classification systems are provided. First, two methods of classification are well known: a ‘loose’ classification by keyword and a ‘fixed’ classification in a hierarchical directory (e.g. a table of contents). In between these poles are several intermediate classification methods. How can they be characterized? The simplest classification is given if a single keyword serves as a classifier, for example in an alphabetical *list*. Here, one keyword does not contain any relation to other keywords. A given keyword might imply a totally different meaning in different contexts (e.g. ‘agonist’ as functional descriptor for a muscle and as a receptor activating substance). Therefore, keywords often have a short description that allows distinguishing between different meanings. A keyword with a short description constitutes a simple *glossary*. If the keywords have relations to other keywords (“relates to“, “related terms“), a *‘referenced glossary’* is given. If, additionally, the parent-child relations “is part of” (narrower) and “contains” (broader) are allowed, it is a *thesaurus*. The forms of the different methods are shown in table 2. For general information on thesauri and indexing, see the specification of ISO standards 5963 and 2788 (ISO International Standards Organisation 2002; 2002).

table 2. Structure of different approaches to classification. *Keywords as the simplest form can be extended to glossaries by provision of short descriptions and, further extended to referenced glossaries by provision of related terms. Thesauri, additionally, provide hierarchical relations and present the most comprehensive approach to classification.*

keyword	**glossary**	**referenced glossary**	**thesaurus**
‘neuron’	‘neuron’	‘neuron’	‘neuron’
	description	*description*	*description*
	excitable cell…	excitable cell …	excitable cell…
		related	*narrower*
		axon	axon
		dendrite	dendrite
		soma	soma
		…	…
			broader
			network
			brain
			…
			related glia cell
			…

Obviously, a thesaurus is the most potent approach to describe a knowledge domain comprehensively. Since the relations in a thesaurus can be multiple (e.g. ‘synapse’ relates to ‘neuron’ but also to terms of a different class such as ‘calcium’), there can be no unique directory that represents the knowledge domain but rather a (semantic) network[4].

In print media, this structure of thesauri can be represented by ordering the concepts alphabetically in the form shown in table 2. The reader has to leaf through the print to find the desired trajectory. In digital media, hypertext can be used and users can jump from concept to concept. Additionally, multiple directories can be generated using the selected concept as root element. The concrete configuration of multiple directories is determined by the choice of the number of nodes around a given concept (higher, lower, related) shown, i.e. the aperture of concept ramifications. Multiple directories can be applied for the presentation of concept positions (see table 3). If a general concept such as ‘learning’ or ‘sensory systems’ is selected, multiple directories can serve as some kind of table of content.

With thesauri it is possible to build arbitrary trajectories through information on brains that resemble free associations, e.g. going from neuron to synapse to calcium, second messenger, adaptation, learning, Hebb’s rule, etc. However, the resulting system is not easy to comprehend. A novice student of brain sciences, for instance, would get lost almost immediately when confronted with such a representation of the brain sciences. An easier approach would be a static directory (e.g. a table of contents). Reducing a thesaurus just on the root element and defining specific classification parameters can achieve this static directory. For example, if ‘nervous system’ was the root element it would be possible to build the children ‘anatomy’ and ‘physiology’ that are based on a distinction between structure and function (the classification parameter). Structure could further be divided into subsystems (e.g. vegetative and somatic) and function into sub-functions (e.g. supply and information processing). *A directory is the result of mapping an ontology (e.g. a thesaurus) onto a specific conceptual framework*[5].

[4] Imagining the network character of thesauri is not trivial. It might prove helpful to click through a hypertext representation such as MeSH Browser: http://www.nlm.nih.gov/mesh/ (National Institute of Health 2002b).

[5] For the sake of clarity, a terminological note: ‘ontology’ corresponds to ‘conceptual framework’ or ‘theory’. ‘Theory’ is common in everyday language, while ‘ontology’ – in its non-philosophical but rather technical sense – is frequently used by knowledge and database engineers and in the field of artificial intelligence.

table 3: Concept positions in multiple directories. *Shown is where the concept 'neural networks' appears in the MeSH. (The choice of this example does not imply a recommendation; see text)*

Natural Sciences	**Information Science**	**Information Science**	**Information Science**
Mathematics	Medical Informatics	Medical Informatics	Pattern Recognition
Math. Computing	**Computing**	**Medical Inf. Comp.**	*Neural Networks*
Decision Support	Computing Method.	. Computing Method.	
Data Interpretation	Artificial Intelligence	Mathematic Comp.	
Decision Theory	Expert Systems	Decision Support	
Neural Networks	Fuzzy Logic	Data Interpretation	
	Natural Language	Decision Theory	
	Neural Networks	Decision Trees	
		Neural Networks	

Two disadvantages of a static directory are obvious: first, as any selection, it implies the neglect of other views (e.g. other conceptual frameworks). There will always be also arguments against a selected view and other views are supposedly superior with respect to this or that issue. For example, distinguishing anatomy (structure) from physiology (function) might result in a lucid account of neuronal histology (in the anatomy branch) that is very useful for clinical neuropsychologist students learning diagnostic criteria of neural degeneration. This distinction might also result in a lucid account of the transmitter synthesis (in the physiology branch) that is important students in the Neurosciences for understanding the characteristics of a synapse. But students of computer science trying to assess the role of plasticity in neural networks might better be served with an account that does not differentiate between structure and function but, for example, works on specific animal models. The second disadvantage of a static directory is the omission of specific correspondences and relations between the different branches. This implies that these relations – on condition that they are actually referenced in the medium – only become overt in the second view as an extension of the respective concept. In hypertext, for example, the relation from synapse to calcium is hidden in the static directory view, but might become accessible when item 'synapse' is selected and a detailed view pops up that contains descriptions and hyperlinks. However, there are as well obvious benefits of a static directory. It mediates a standardized access to information as well as a map-like orientation, allows associating new concepts to nodes and branches of that framework and suggests a sequence of contents. A standardized, repeatable access to information (e.g. in the 'brain navigator' on the library computer) serves several purposes. It is important for sharing and communicating it with others. For example, consider a student saying: "I mean *that* meaning of adaptation as it is defined in the brain navigator." A standardized, repeatable access to information is also important for positioning anchors in memory retrieval. Consider the student's soliloquy: "Synapse was a topic under neuron and had the topics release, reception, reuptake, synthesis." So, the student learns to navigate in memory by retracing the access to information in the computer.

The benefits of a static directory can be explained by their property of reducing complexity. Given that recipients can only process a limited number of concepts and relations at a time (for an introduction into issues of limited capacity see e.g. Broadbent 1958; 1975), it follows that leaving out relations reduces the cognitive load and, thereby, makes possible a more general view. Furthermore, the conceptual framework the directory is based on, offers a strategy how to progress from one concept to the other. The decision of what to process next is not necessarily to be worked out at each step.

In general, two modes of navigation in the knowledge of the brain sciences can be differentiated: a thesaurus offers primarily a view on the 'nearest' neighbors of a given concept (concept-centered, first person perspective). A directory offers a view on a specific high-level structure (framework-centered, third person perspective). The former is comparable to landmark navigation, the latter to map navigation. Novices will probably prefer map navigation because it offers a reduced and standardized access to information. Experts will probably be nimble in both systems. They are able to abstract from a specific conceptual framework and 'dive' through the relations from one concept to another and they will be able to switch and plug concepts into several conceptual frameworks. The actual mode is determined by the task: the map navigation allows a categorical approach. For instance, a student using a 'brain navigator' on a library computer searching in the context of, say, synaptic receptors, could use the map mode, if a general categorization (e.g. ionotropic vs. metabotropic) is to be made. The landmark navigation with a thesaurus allows a step-by-step decision on where to look next that helps to check consistency of a concept against contextual details. The student could use the landmark mode for 'scanning' heuristically the topic for related transmitters. With each step the classifiers can be switched e.g. from 'ionotropic/metabotropic' to 'inhibitory/excitatory effect' to alphabetical order etc.

In conclusion, two general forms of representing information on brains can be distinguished: a knowledge base and a conceptual framework. A concrete representation of the former is a thesaurus, of the latter a directory. These forms are best understood as complementary, not as alternative representations. For example a ‘brain navigator’ on a library computer would provide both a map mode (conceptual framework, table of content) and a landmark mode (knowledge base, thesaurus/hypertext).

The following sections review existing resources for the neural, cognitive and computational domain. In order to add a concrete touch to the abstract character of this task, it is viewed with respect to the question: How should a ‘brain navigator’ (e.g. on a library computer) be optimally designed to represent the brain sciences? Three indispensable functions can be stated so far:

1. The knowledge base must be capable of generating a thesaurus, which implies the simpler cases of a keyword repository and a (referenced) glossary.
2. The system should generate multiple directories (e.g. show concept positions and concept ramifications) from the thesaurus.
3. The system should provide a directory, e.g. a table of content.

An ideal ‘brain navigator’ incorporating these expert skills would offer both a thesaurus inspired navigation and multiple directories that follow from different conceptual frameworks either in parallel or as exclusive modes. Of course, it seems sensible to offer novices a directory or table of content, and provide specified trajectories (courses) through the knowledge base, then confront advanced recipients with multiple directories and, finally, take experts to full functionality.

1.3.4. Knowledge bases for brains

1.3.4.1. *Universal indexing*

If there was an index with a universal scope (i.e. one that covers not only the brain sciences appropriately but also other knowledge domains) this should evidently be taken as a classification resource for a knowledge base for brains. In fact, there are many indexing systems, mainly coming from the bibliographic professionals that seek to classify universal knowledge. One very influential system is called the “Library of Congress Subject Headings”, short LCSH (Library of Congress 2002). It encompasses all knowledge domains, but not in an excessively specific manner. An alternative system, provided by the Online Computer Library Center (OCLC), is called Dewey's decimal classification (Online Computer Library Center 2002). As the LCSH, it is insufficiently specific. Both classification systems could serve as a general guideline for a knowledge base, but would definitely have to be supplemented. For example, the field of the Neurosciences is not represented explicitly, but distributed over many subjects, such as Science > Physiology > Neurophysiology and Neuropsychology. As another example, cognitive psychology is weakly represented, namely only as an extension of Philosophy/Psychology/Religion > Psychology > Consciousness/Cognition. In short, universal indexes are too broad and sometimes, maybe, too old to justify the claim for an actual account of universality.

1.3.4.2. *Thesauri*

A more specific resource developed and maintained by the US National Institute of Health is called Medical Subject Headings, short MeSH (National Institute of Health 2002b). It is well elaborated and widespread. For example, it is integrated in the major databases and indexing services such as BIOSIS, Medline etc. Thus, using MeSH for the 'brain navigator' would ensure easy recognition and reuse for indexers, bibliographers etc. and the existence of interfaces in digital systems. Furthermore, there are even theoretical approaches that employ the MeSH, e.g. a study that uses the MeSH to exemplify specific strategies for building dynamic classification schemes and ontologies (Kahng *et al.* 1997).

However, MeSH is developed in the field of medicine. Looking at the MeSH-Browser (National Institute of Health 2002b) reveals several oddities: for example, the term 'neural networks' does appear under 'natural sciences' and 'information sciences' but only as substructures of 'medical informatics computing', 'decision support techniques' or 'pattern recognition'. Similarly, in the field of cognitive science, 'attention' appears as a sub-term of 'arousal' but not as a sub-term of 'cognition'. Neuroscience is quite well represented, even though some specialties (for example 'dendritic spines') could not be found. All in all, with respect to its comprehensiveness, MeSH appears to be very suitable for a keyword resource for the brain navigator (not least because the NIH offers ready to use downloads.)

Noteworthy, too, is the AOD (Alcohol and Drug Abuse) Thesaurus that is hosted by the same Institution as the MesSH (National Institute of Health 2002a). At first sight, considering the AOD-T seems to be a categorical error, but – opposite to the purely medical approach always shining through the MeSH – the AOD-T obviously seeks to integrate biological and psychological concepts more carefully, quite similarly as it is usually done in the neural and cognitive sciences. For example, the section 'concepts in psychology and thought' exceeds the specificity and intuitive appeal of terms given in the MeSH many times over. However, the AOD-T is not so popular as the MeSH (e.g. search in www.google.com for "mesh+thesaurus" vs. "aod+thesaurus" resulted in 41400 vs. 967 hits, search performed 13.10.2003).

Beside these public thesauri, useful resources come from bibliographic indexing services. An indexing service is the sum of efforts (usually an organization or company) that seeks to integrate all relevant publications of a knowledge domain into one index by filing bibliographic data (and possibly abstracts) and supplying each data-set with index terms. The indexers make use of thesauri that provide definitions and controlled vocabulary helping to find the correct index terms. The results of these procedures are the typical services on library computers such as indexes integrated in databases and equipped with user interfaces for queries by scientists or other information seekers. BIOSIS, for example, provides an online thesaurus that serves as a guide in search results. This BIOSIS online thesaurus is based on 'major concepts', which are very broad. For example, concepts relating to the field of the Neurosciences are to be found in the branch 'neural coordination', which contains only one more neighbored node, namely 'nervous system'.

A similar coarse graining is found in the thesaurus for Zoological Records (BIOSIS 2002). In the Science Citation Index (ISI 2002) no standard indexing terms are assigned; it is limited to the author's choice of title and abstract words and the system's choice of keywords. A rather specific thesaurus for the Neurosciences is provided by Elsevier Science publishers (2002). It contains 10 classes with approximately 150 fields that are specific enough to account for the neuroscience field. The fields of computer science and cognitive psychology are not covered. A thesaurus for psychology is provided by the American Psychological Association (Walker 1997). The APA thesaurus is comprehensive, but appears to be oriented on application, not primarily on basic research. For example, the 'term cluster section' provides

Neuropsychology/Neurology but not ‘cognitive psychology’. On the other hand, the ‘relationship section’ contains sufficient terms for cognitive psychology, e.g. ‘cognitive process’ yields about 20 narrower terms and 20 related terms. A short description is also provided. There are many terms that actually do not fall in the neural-cognitive-computational cross-section. Thus, an extraction of the relevant terms is also missing. For the sake of transparency, only thesauri were taken into account here that were publicly accessible. There are likely to be many more thesauri and other indexing systems that are not publicly available (in the case of indexing companies evidently because they are an essential economical basis that is to be kept secret). But already the few examples considered here indicated that brain sciences are not appropriately represented by one existing thesaurus, but rather would demand a combination of specific thesauri.

1.3.4.3. *Glossaries*

An alternative approach for classifying knowledge is to make use of keywords. Rich resources for keywords are glossaries. As stated above, these can be extended to thesauri when hierarchical relations and related terms are provided. Several resources are accessible but – as it was already the problem with respect to thesauri – none was found that could be called comprehensive. In detail: a comprehensive glossary for Neuroinformatics could not be found within the publicly accessible resources. An alternative is the glossary at Principia Cybernetica Web (Heylighen 2002), which offers about 600 terms with definitions and relations. Moreover, the Principia Cybernetica Web offers several textbook like articles on general terms. Similar resources are “Nonlinear Dynamics and Complex Systems Theory” Glossary of Terms (CAN Center for Naval Analyses 2002) and an online glossary to the book Computational Beauty of Nature (Flake 2002). For general computing issues, the Free On-line Dictionary of Computing (Howe 2002) and the Mathematical Programming Glossary (Greenberg 2002) are recommendable resources. Cognitive science and philosophy are well represented on the internet: outstanding in the field of cognitive science are a glossary from Blackwell Science Publishers (2002), the glossary of the online textbook “Cognitive Psychology” (Medin 2002) and the University of Alberta’s Cognitive Science Dictionary (Dawson 2002), in the field of philosophy the Washington University (Eliasmith 2002) and the Meta-Encyclopedia of Philosophy (Chrucky 2002).

table 4. Aptitude of several resources to represent thesauri or directories in the neural, cognitive and computational domain. *The number of circles in the column 'thesauri' indicates how of the demands 'comprehensiveness', 'connectivity' and 'quality' were satisfied. Crosses in the column 'directory' indicate if a resource can provide templates for directories that represent the neural, cognitive and computational domain (large 'X'), only with reservations (small 'x') or not (missing).*

		thesauri			directories		
		neural	cognitive	computational	neural	cognitive	computational
1	NIH 2002a	OOO	O		X	X	x
2	NIH 2002b	OO	OO		X	X	
3	LoC 2002	O	O	O	X	X	
4	BIOSIS 2002	O					
5	ELSEVIER 2002	OO			X		
6	Heylighen 2002	O	O	OOO			
7	CNA 2002	O	O	OO			
8	Flake 2002	O	O	OO			
9	Howe 2002			OO			
10	Greenberg 2002			OO			
11	APA 2002	O	OOO	O	x	X	
12	Blackwell 2002	O	OO	O			
13	Medin 2002	OO					
14	Dawson 2002	O	OO				
15	Eliasmith 2002	O	OO				
16	Chrucky 2002		OO				

1.3.4.4. *Evaluation of resources*

The analysis of the existing information resources on brains indicates that, contemporarily, there is no appropriate knowledge base for the brain sciences. Rather a combination of resources is a more promising procedure. In the example of the 'brain navigator', it would be necessary to merge thesauri and glossaries in one database. Since there is a considerable overlap in the coverage of terms that could result in a massive redundancy on the one hand and uncontrollable terminological heterogeneity on the other hand, it seems reasonable to concentrate on few, comprehensive resources rather than collecting as much as one can. A justified selection is only possible if an evaluation of their aptitude to be integrated in a knowledge base is carried out. With respect to our hypothetical application context of the 'brain navigator' and for the sake of concreteness, an exemplary evaluation was actually performed.

According to the two forms of navigation in the brain navigator 'landmark mode' and 'map mode', the evaluation of the resources was differentiated for thesauri and directories. The aptitude of a given resource to determine a static directory, was only assessed with 'yes' (indicated by X), 'no' (indicated by a missing) or 'with reservations' (indicated by an X in brackets). The aptitude of a resource to act as a resource for a thesaurus could be rated with a score from zero to three points (indicated by the circles). Points were scored for a match on either of the following points:

- Comprehensiveness: Does the resource comprise one of the domains neural, cognitive or computational?
- Quality: Is the resource well formed and evaluated?
- Connectivity: Is it possible to merge the resource with other resources (e.g. in terms of data-structures)?

In general, the evaluation illustrates that it is in principle possible to assemble resources in a way that could result in a comprehensive knowledge base for brain sciences (see table 4). In detail, the analysis shows that resources deserving preferential treatment are: (1) MESH, (2) the APA thesaurus and (3) the Principia Cybernetica Web (PCW) glossary. A valuable alternative is (4) the AOD thesaurus - not least because it is also quite strong in the neural domain, it should be considered anyway. When cross checking the resources chosen so far against the disciplines involved, special keywords from the field of Neuroinformatics and philosophy of mind are strikingly missing. Therefore it might be recommendable to supplement (5) the online glossary to the book Computational Beauty of Nature and (6) the Philosophy of Mind glossary of the Washington University.

With respect to the design of a 'brain navigator' (see also the three tasks in 1.3.3), resources for setting up thesauri (task 1 and, to large parts implied, task 2) can be scooped from resources taken into account so far. Setting up static directories (sufficing task 3) is still missing - and it is far more difficult with the resources at issue. For example, candidates for setting up directories for the computational domain that refer to Neuroinformatics are completely missing and even the neural and cognitive domains are not well staffed. This is not so surprising, however, since predominantly resources for setting up knowledge bases were taken into account so far – and these typically provide dynamic, not static directories (see also 1.3.3). The issue of specific conceptual frameworks and static directories is topic of the next section.

The analysis of existing information resources on brains in this section should also be regarded with respect to the question whether it can tell us something about the current state of theoretical integration of the neural, cognitive and computational domains. Obviously, one would expect nowadays from a well-integrated brain science that a comprehensive knowledge base already exists, as it is the case for the medical sciences with the MeSH. This is not the case for brains sciences. On the other hand, a *dis*integrated science would not have to offer comprehensive and supplementary resources at all and these would not contain segments that already show integrated accounts of the neural, cognitive and computational domain. Taken together, the existing information resources on brains indicate that brain sciences are in an intermediate state of theoretical integration.

1.3.5. Conceptual frameworks for brains

An integrated brain science should ideally offer a single conceptual framework that spans the neural, the cognitive and the computational domain. Whether pursuing this objective is naïve or even undesirable shall be left open here. Rather, the quest for a single conceptual framework should inform us about the state of theoretical integration in the brain sciences. As already became clear in the previous sections, the quest clarifies that the task of designing a directory, as a representation of a conceptual framework, is far more complicated than providing a repository of keywords. In other words, developing a 'map' for the brain sciences is something that differs considerably from representing knowledge of the brain sciences as a dynamic and variable landmark-system. In the example of the 'brain navigator' on a library computer, the task of finding a map results in a standardized access to knowledge that is offered to students – a kind of table of contents for the brain sciences. As already pointed out above, such a single conceptual framework for brain sciences is not within reach. But is it impossible? How can we proceed in order to find out more about a directory that is congenial to the neural, cognitive and computational domain? The proximate task is to analyze existing conceptual frameworks from the respective domains. Typical cases for the application of conceptual frameworks are textbooks.

1.3.5.1. *Textbooks*

On the one hand authors of textbooks seek to integrate a wealth of the work done in a given knowledge domain, on the other hand they aim at imparting a digestible scheme. They try to construct a comprehensive and comprehensible format. They map a specific, individual conceptual framework onto a knowledge base. How can these conceptual frameworks be extracted from the textbooks? It can be assumed that they are reflected in the tables of content of the textbooks. Consequently, the next step is to review tables of contents of textbooks, to analyze and evaluate them. The ‘one and only’ integrated textbook for neural, cognitive and computational domain is missing. So, the question is: how can the various specific conceptual frameworks be used for finding a common conceptual framework for the neural, cognitive and computational domain? The first task is to decompose the specific conceptual frameworks of textbooks, and discuss their interrelations, consistency and coherence. The second task is to synthesize the results of the analysis.

The method applied was to select actual textbooks from the neural, cognitive and computational domain from the subject headings ‘neuroscience’, ‘cognitive psychology’ and ‘neural networks’, respectively, from the large online book store ‘Amazon’. The 30 best selling entries were searched for textbooks that offered a structured table of contents. As stated above, there is no fully consistent terminology. Thus, the main task was to identify the essential terms, dichotomies and other ordering criteria that represent the knowledge domain without being overrating, neglecting or misrepresenting (or being otherwise offensive). The rationale applied was: if there were a valid single conceptual framework for the brain sciences, it would probably result from an elaboration of the most commonly used classifiers. For this reason, it was analyzed whether a specific set of classifiers was characteristic for textbook accounts of a given knowledge domain.

Classifiers refer to the transition from one section to the next section.

- ‘temporal’ refers to an earlier/later or slower/faster relation, e.g. if a chapter on perception was followed by a chapter memory, a temporal classifier would apply because the proposition “perception precedes memory” is true. Another example for a temporal classifier: “learning is faster than evolution“.
- ‘spatial’ indicates if concepts are smaller or neighbored, e.g. neuron contains axon and neurons neighbor glia cells.
- ‘complexity’ indicates that one chapter refers to a more complex issue than the preceding chapter, e.g. feed-forward network is simpler than recurrent network.
- ‘functional’ indicates that different forms of functioning are worked off, e.g. visual, auditory, with or without learning, thinking or acting etc.
- ‘keyword’ indicates that sections are ordered by prominent keywords that do not show obvious relations to each other, e.g. experimental model systems, famous works, milestones etc.

Sometimes specific classifiers, dichotomies or keywords are remarkable (as stated in the column ‘comment’ in table 6). For example, clustered structures are quite common since they have the benefit of not implying a strong ontological claim and, thus, evade problems resulting from concurring approaches. Consequently, these structures are taken frequently when each chapter is provided by a different author. ‘Nodes’ indicate how many levels of hierarchy are used. ‘Neuron’ with subsection ‘synapse’ is one node – if ‘synapse’ additionally has a subsection ‘calcium’, there are two nodes. For providing a general idea about the contents of the textbooks table 5 contains summaries.

The results of the analysis are summarized in table 6. Textbooks on Neuroinformatics refrain from temporal and spatial classifiers. Instead ‘function’ as well as ‘complexity’ are the dominating classifiers in the tables of content. This is not very surprising since computational issues are not bound to the natural categories space and time. The predominant classifier refers to the type of network architecture. For illustrating what the term ‘architecture’ means in the computational domain, two exemplars (feed-forward and recurrent) are shown in fig. 12.

table 5. Summaries of the analyzed textbooks. The tables of content were analyzed with respect to the application of certain classifiers. The results are presented in Table 6. The selection of the textbooks is described in the text.

computational domain

Handbook of brain theory and neural networks (Arbib 1998) The book is primarily based on alphabetically ordered articles that do not contain assumptions about a conceptual framework – only as a comprehensive theme list. However, two introductory chapters are offered: 'background' introduces the neuron (basics, receptors and effectors, models, details), different levels of analysis and dynamics. The chapter 'road maps' encompasses connectionism (psychology, linguistics and AI), dynamics (optimization, cooperativity, self-organization) learning (deterministic, statistical), applications, biological systems (neurons, networks, mammalian brain), sensory systems (vision, other), plasticity (mechanisms, development, learning) and motor control (e.g. patterns).
Introduction to the Theory of Neural Computation. (Hertz et al., 1991) The book offers a structure clustered by exemplary models: Hopfield (and its extensions), optimization problems, perceptrons, multi-layer networks, recurrent networks, (unsupervised) hebbian and competitive learning and neural networks statistics.
Neural Organization (Arbib et al. 1997) The textbook introduces 3 different structures explicitly (structure, function and dynamics), extends for the most part in anatomical exemplary systems (olfaction, hippocampus, thalamus, Cortex, cerebellum, basal ganglia) and roofs these with an outlook on cognition.
Computational explorations in cognitive neuroscience (O'Reilly and Munakata 2000) The textbook is two-part. The first part deals with 'basic neural computational mechanisms'. Herein, concepts referring to individual neurons are distinguished from concepts referring to networks and from learning issues (hebbian, error-driven, model related and others). The second part aims at cognitive phenomena: after introducing the large-scale organizational structure, the subjects perception and attention as well as memory, language and higher-level cognition are worked off.
Neural and adaptive systems (Principe et al. 2000) This textbook is based on a structure that goes from simple to complex: after introducing data fitting, pattern recognition and different aspects of multilayer perceptrons (MLPs) are treated. Learning is divided into hebbian as well as competitive and Kohonen-Learning. More general chapters on processing (signal processing, adaptive filters, temporal processing) precede the closing chapter on the (most complex) recurrent networks.
Fundamentals of Neural Networks (Fausett 1994) This is a straightforward textbook that introduces neural networks by showing their fields and methods of application and a brief historical sketch. Then several networks for pattern recognition (Hebb, Perceptron, Adaline) and association (heteroassociative, autoassociative and other forms memory) are distinguished from competitive networks. The adaptive resonance theory and backpropagation are assigned separate chapters. A sampler of networks closes the book.
Computing the brain (Grethe and Arbib 2001) The textbook is primarily concerned with technical – not with thematic – aspects. It introduces 'Neuroinformatics' as a science of databases and tools. Modeling and simulation is described in chapters on the Neural Simulation Language (NSL), a modeling system (EONS) and PET. In the following, databases for time series (experimental data, design concepts, interaction issues), for atlas data (brain maps, 3d surfaces, rat brain) and – along with concepts for data management (federation of databases, ontologies, annotations, space management) – for models (repositories, BMW brain model on the web, knowledge management) are explained.

cognitive domain

Cognitive Psychology (Anderson 1980) Anderson's textbook goes from simple to complex, from lower to higher: the introduction is followed by the topics perception, attention and performance, representations, memory, problem solving, expertise, reasoning and decision and language. (Interestingly, learning is missing in the explicit structure!)
Cognitive Psychology (Best 1999) Best's textbook uses perception, (attention and object recognition), memory, knowledge (symbolic and connectionist), language and thinking (reasoning, decision, concepts, problem solving) as classifiers.
Cognitive Psychology (Medin et al. 2001) The structure of the textbook is somewhat different to that of Anderson (1999) and Best (1999). However, it builds on a key concept, namely 'information'. After this concept is deduced from the introduction, its acquisition (learning, perception, attention), storage (memory, knowledge, imagery) and its application in language as well as thinking (reasoning, problem solving, expertise, decision) is provided.
Connectionism and the mind (Bechtel and Abrahamson 2002) This book uses a "basics prime subjects" structure, i.e. two chapters build a fundamental understanding on architectures for modeling cognition and special subjects (learning, pattern mapping, representation, higher cognition etc) and closes with implications (Artificial Life, Brain Issues etc.).

table 5. continued

neural domain

Biophysics of computation (Koch 1999) This book is strictly oriented on the neuron. Nonetheless, it uses a heterogeneous structure beginning with introductory 'horizontal' chapters (the membrane equation, linear cable theory), then following the doctrinal information processing direction (passive dendritic trees, synaptic input, synaptic interactions in a passive dendritic tree, the Hodgkin-Huxley model of action-potential generation), where in-depth excursions are made (phase space analysis of neuronal excitability, ionic channels, beyond Hodgkin and Huxley: calcium, and calcium-dependent potassium currents, linearizing voltage-dependent currents). Then the logical extension in time (plasticity) is considered (diffusion, buffering, and binding, dendritic spines, synaptic plasticity) before again general issues are treated (simplified models of individual neurons, stochastic models of single cells, bursting cells input resistance, time constants, and spike initiation, synaptic input to a passive tree, voltage-dependent events in the dendritic tree, unconventional coupling).

From Neuron to Brain (Nicholls et al. 2001) The book is divided into an introduction and three parts: Signaling, integrative mechanisms and development. The first part has a strong low-level (biophysical, molecular and cellular) orientation, i.e. several chapters on ions and ion channels, electrical and biochemical bases of synaptic transmission and plasticity. 'Integrative mechanisms' is clustered by selected (well-known) systems such as, cellular basis of behavior in leeches, ants and bees, the autonomic nervous system, transduction (mechanical, chemical), processing (somatosensory, auditory), several chapters on vision (transduction, primal visual cortex, general cortical architecture, ocular dominance) and one on motor control. The first chapter of the third part (development) is structured temporally. The other parts are concerned with general developmental aspects (denervation and regeneration) and the visual and auditory model systems, respectively.

Principles of neural science (Kandel et al.1991) This is the best-known textbook for the neural sciences. The contents in brief offer a general introduction that relates biology to behavior. The structure then is from small to large: After neuron and synapse follows another introductory section on cognition. Then, as functional domains of neuroscience, perception, movement, arousal and emotion, development, and, finally, language, thought, mood as well as learning and memory are offered.

Essentials of neural science and behavior (Kandel et al. 1996) This is a 'digest' of the principles of neural science (Kandel, Schwartz and Jessel 2000). The structure is somewhat different: the general introduction is followed by a purist biological introduction (cell biology, anatomy, development). Then signaling within and, thereafter, between nerve cells is considered. Then, again an introduction of (non-purist) cognitive aspects is presented and followed by perception, action, genes/emotions/instincts and language/learning/memory.

Cognitive neuroscience (Gazzaniga 2002) This is a textbook that offers a historical introduction followed by basics (the cellular and molecular basis of cognition, gross and functional anatomy of cognition, the methods of cognitive neuroscience). Then, a complexity criterion (from simple to complex, from early to late) is applied to functional systems (perception and encoding, higher perceptual functions, selective and attention orienting, learning and memory, language and the brain, cerebral lateralization and specialization) that closes with a chapter on behavior (the control of action). Some general chapters roof the structure (evolutionary perspectives, development and plasticity, the problem of consciousness).

The New Cognitive Neurosciences (Gazzaniga 2000) This book is a comprehensive and very well known and edited handbook that is made of chapters of specialist authors. It begins temporally organized (development, plasticity), then functional (sensory systems, motor systems). The extensive sections on cognitive aspects are functionally ordered (attention, memory, language, higher cognitive functions, emotion, evolution, consciousness).

The Computational Brain (Churchland and Sejnowski 1992) This book has a non-hierarchical structure that is primed by a hierarchical structure that introduces several 'levels'. Readers have to map the functional themes computation, representation, plasticity and sensorimotor integration onto these levels. The structure could be called 'basics prime subjects'

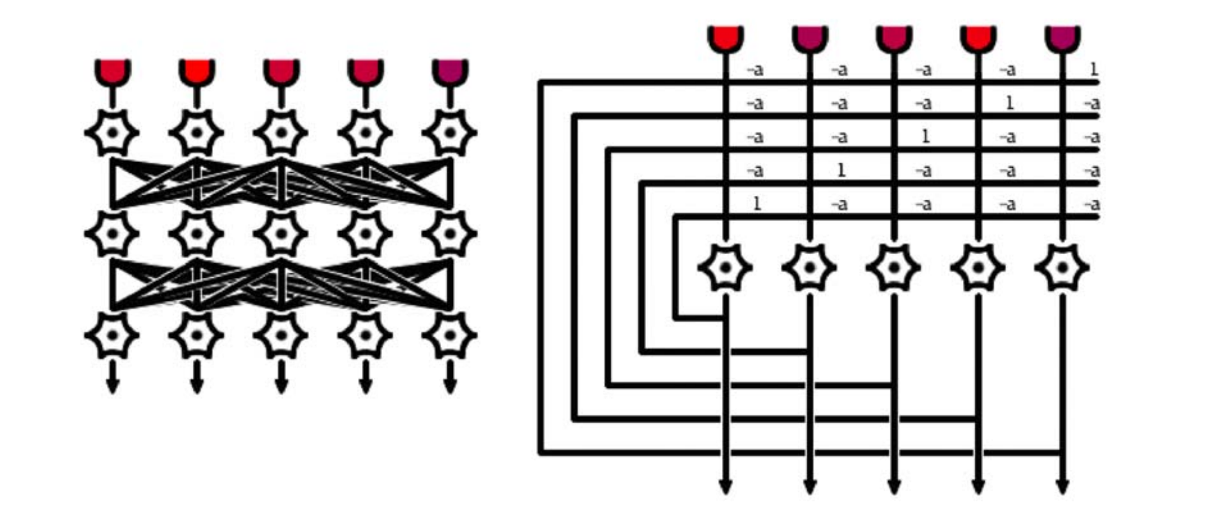
Neurowissenschaft (Dudel et al. 2001) This German textbook applies a straight sequence of chapters that uses a mixed set of classifiers, temporal (phylogeny, ontogeny) spatial (molecular, cellular) or functional (perceptual, hormonal). One after the other, evolutionary aspects serve as general introduction followed by molecular aspects. Then, ontogeny precedes biophysical basics of cells synapses and motor aspects. The vegetative and hormonal system conclude this general section. The remainder refers to sensory systems (general, chemical, thermo, mechanical, auditory, peripheral and central vision, electrical, magnetic, proprioceptive) and cognitive aspects (plasticity, learning and memory, rhythmics, sleep).

table 6. Summary of the analysis of textbooks. *Analyzed were classification criteria in the tables of contents. 'Nodes' shows how many levels of hierarchy are used. Within the classifiers' section, small letters (x, o) indicate 'is applied', large letters indicate 'major classifier', '–' indicate 'not applied'. For the classifier 'keyword' another symbol (o) is used because it belongs to an exclusive classification category. The column 'comments' contains special classifiers etc. For a detailed description of the procedure, see text. A summary of the contents is given in table 5.*

	nodes	classification criteria					comment (special classifiers etc.)
		temporal	spatial	complexity	functional	keyword	
computational							
Arbib 1998	1	–	–	–	–	–	alphabetical
Hertz et al. 1991	1	–	–	x	X	o	unsupervised/supervised learning
Arbib et al. 1997	2	–	x	–	x	O	structure/function/dynamics
O'Reilly and Munakata 2000	3	–	–	x	X	o	neuron, network, learning, cognition
Principe et al. 2000	2	x	–	x	–	O	recognition, learning, association
Fausett 1994	3	–	–	x	x	O	architecture
Grethe and Arbib 2001	2	–	–	–	x	O	technical
cognitive							
Anderson 1980	1	X	–	x	x	o	
Best 1999	1	X	–	x	x	o	
Medin et al. 2001	1	–	–	x	X	o	
Bechtel and Abrahamson 2002	2	–	–	x	x	O	
neural							
Koch 1999	1	x	x	–	x	o	neuron/synapse, passive/active
Nicholls et al. 2001	2	x	x	x	x	O	sensory/motor, clustered by model systems
Kandel et al.1991	3	x	X	x	x	o	neural/cognitive
Kandel et al. 1996	3	X	x	x	x	o	
Gazzaniga 2002	2	x	x	X	x	o	
Gazzaniga 2000	2	x	–	x	x	O	
Churchland and Sejnowski 1992	1	–	–	x	x	o	
Dudel et al. 2001	1	x	x	x	x	o	ontogeny/phylogeny

A typical textbook on Neuroinformatics will probably provide an introductory chapter on the neuronal analogy (i.e. the relation to biological neurons and networks), and then go through the different architectures: feed-forward and recurrent with or without (un-) supervised learning. An appendix offers mathematical and statistical prerequisites. As implied above, the specific classifiers are feed-forward vs. recurrent, learning/non-learning and unsupervised/supervised learning. The textbooks differ in their explanatory orientation, i.e. whether they seek to account for formal-mathematical, technical or natural (biological or cognitive) phenomena.

Textbooks on cognitive psychology are quite uniform. Probably, this is due to the consequent application of the so called 'information processing paradigm' that replaced behaviorism in the 50s and 60s (see e.g. Broadbent 1958; Neisser 1967; Atkinson & Shiffrin 1968): According to this notion (see fig. 13), external information causes sensation and perception. Selective attention routes information to short-term memory and elaboration and coding routes it further to long term memory. Actions are the result of the processing at the different levels. Beside this structural unanimity, the textbooks differ, above all, in the sequence and the focus on the major concepts of the information processing approach.



-a	-a	-a	-a	1
-a	-a	-a	1	-a
-a	-a	1	-a	-a
-a	1	-a	-a	-a
1	-a	-a	-a	-a

fig. 12: ***Typical architectures of neural networks.*** *The upper semicircles are input elements, the arrows indicate output interfaces, and the symbols indicate artificial neurons constituting 'layers' in a horizontal line. On the left side, a feed-forward network is shown in which it is assumed that the information is processed exclusively from the upper layers to the lower layers without feedback. On the right side a recurrent net is shown. The graphical presentation as a matrix shows that all neurons are completely inter-connected. The connections ('weight') to other neurons are indicated by '-a' connection to the respective neuron is indicated by '1'. © rubin (reprinted with permission).*

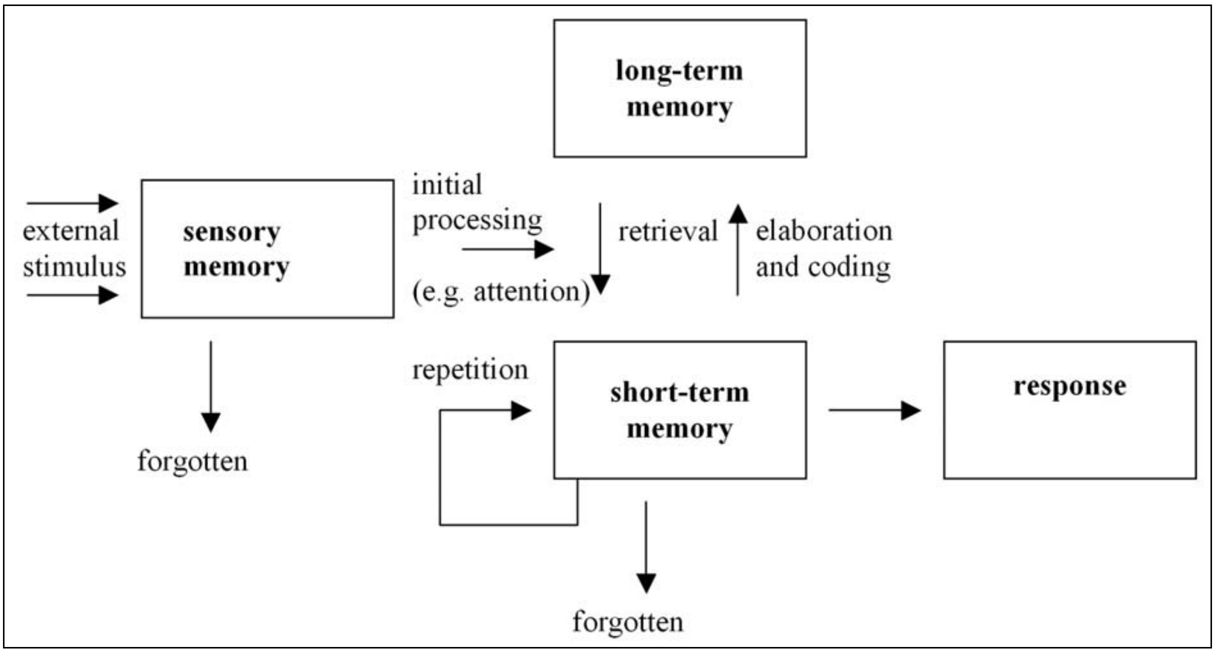


fig. 13: A sketch of the Information Processing Approach anchored on memory systems. External stimuli come into the sensory memory, are pre-processed and possibly routed to short-term memory. Then a response is chosen and/or coding in long-term memory is carried out.

A typical textbook on Cognitive Psychology would probably begin with sensation and perception followed by attention, memory and thinking. The topics representation and knowledge could be made explicit. The classifiers applied in textbooks on cognitive psychology are time (sensation precedes attention), complexity (thinking is more complex than sensation) and function that results from the information processing paradigm. The books usually differ in their focus on a specific processing stage.

Textbooks on Neuroscience show the most mixed application of classifiers. It is obligatory on a natural science that spatial and temporal criteria are applied – and so they are present in the tables of content. Additionally, they are mixed with the classifiers 'complexity' and 'function'. Consequently, most of the structures found, show a high level of conceptual integration, but are at the same time characterized by a mixture of biological classifiers such as ontogeny/phylogeny, sensory/motor, neural/cognitive, organism/environment, neuron/network, wired/plasticity, neuron/synapse vertebrate/invertebrate and others that were already mentioned above (simple complex, early/late). But there is a logic behind this 'strange brew' of classifiers. Usually, textbooks apply a spatiotemporal framework.

The spatiotemporal framework implies a notion of complexity in that large-scale processes are usually more complex than small-scale processes, if the small-scale explanations are applied to the large scale. For example, the transition between neurons to networks is simultaneously a transition from small to large, from short to long and – due to the combinatorial powers – from simple to complex. However, it is not clear that the large scale is explained by small-scale processes. But, generally, *it is conceivable* that the large scale was explained by the small scale. (Here the reductionist commitment of neuroscientists becomes evident.) Along the axis of time, space and complexity the *levels of organization* are to be found (see also 1.3.2). This notion is pervasive in the Neurosciences and presumes a hierarchical structure that extends from microscopic (e.g. electric, molecular, cellular etc.) to macroscopic (e.g. organism, environment etc.). A sketch of the levels of organization is shown in fig. 14. A typical textbook for the Neurosciences will appeal to levels of organization and focus more or less on one of the constituent classifiers (e.g. time or space). For example, excitability (biophysical and molecular basics) primes synaptic transmission and is followed by plasticity, sensory systems (sometimes all modalities), motor systems, cognition, learning, development and evolution. It becomes obvious when analyzing in this arrangement that a mix of temporal, spatial and organizational classifiers are applied and that the order is progressive: excitability-LATER-Synapse-LONGER- Plasticity-LARGER-Sensory Systems-LATER-Motor Systems-HIGHER-Cognition-LONGER-Learning-LONGER-Development-LONGER-Evolution (classifier determining the progression between topics in large letters). In conclusion, the guidelines for explanations in the Neurosciences implied in textbooks appear to be: “Progress on the scales time, space and complexity” and “Explain a given phenomenon with concepts residual on a lower level”.

It is concluded from the analysis of textbook structures that the neural, the cognitive and computational domains use different classification systems. Predominant classifiers in disciplinary exemplars of the domains can be extracted. These classifiers correspond to major conceptual frameworks: in Neuroinformatics it is *architecture*, in cognitive psychology it is the *information-processing paradigm* and in the Neurosciences it is *levels of organization*.

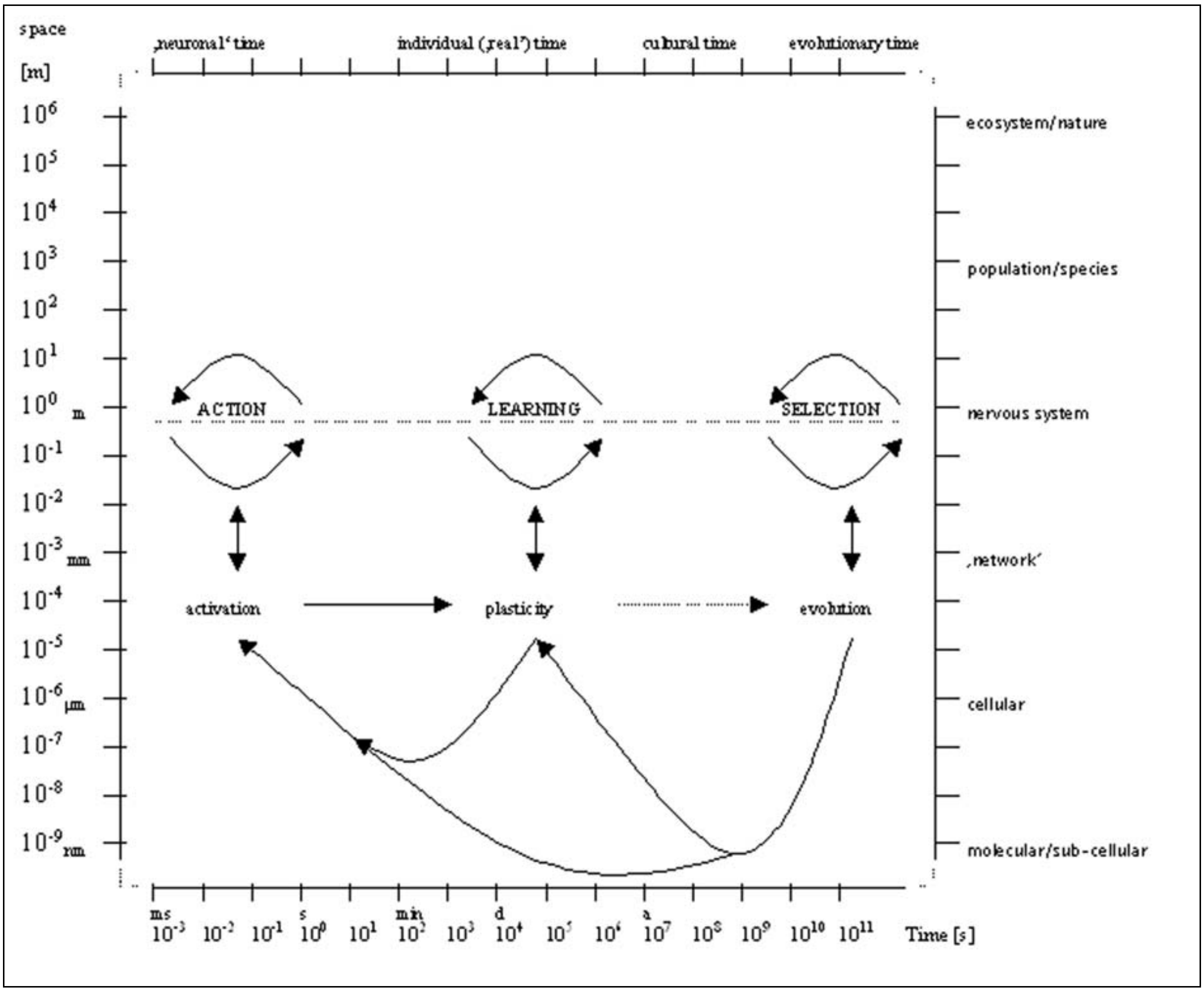


fig. 14: ‘Levels of organization’ as extractable from textbooks of the Neurosciences. The abscissa shows a time section, the ordinate a space section. Both are presented on a logarithmic scale. Common measures (e.g. seconds, meters) are shown as additional tic descriptors. The second axes (upper horizontal and left vertical) show conceptual descriptions of the respective measures. Inside the coordinate system, major processes are shown. On the level of the nervous system action, learning and selection are evident and determined by ‘underlying’ mechanisms of (electric) activation, plasticity and evolution.

With respect to the initial question whether resources in the neural, cognitive and computational domain can form a common conceptual framework for the brain sciences, the analysis of textbook contents has two important implications. First, there is no common conceptual framework that becomes evident in the analysis of textbooks accounts and that could be directly used, for instance, as a standardized access (e.g. table of content) in a ‘brain navigator’ on a library computer. Second, the specific conceptual frameworks used within the domains are homogenous to a certain degree and can therefore serve as starting point for further efforts. (The further analysis of commodity between the specific conceptual frameworks is started in the next section.) With respect to the general question whether the analysis of textbook accounts tells us something about the state of theoretical integration in the brain sciences, these two implications ("no common conceptual but already integrated specific frameworks") corroborate the conclusion already drawn from the analysis of existing knowledge bases: brain sciences are in an intermediate state of theoretical integration.

1.3.6. A common scheme for the brain sciences

It would presumptuous to propose a *common conceptual framework* for the brain sciences. This goal is far beyond the scope of this study, but will rather be the result of theoretical integration brought into being within the next years and decades by concrete scientific work in the contributing disciplines. Nevertheless, it can be tried to analyze the interrelations of the specific conceptual frameworks concerning their consistency and discrepancies. A more moderate goal is, thus, the development of an exemplary account that contains no major inconsistencies: a *common scheme* for the brain sciences. In a 'brain navigator' on a library computer, such a common scheme could act as a preliminary stage of a standardized access to information on brains (a table of content). A common scheme for the brain sciences should integrate all the specific conceptual frameworks of the neural, cognitive and computational domain, i.e. the levels of organization, the information processing paradigm and architectures, respectively. The questions are then: "Do the conceptual frameworks commute?" and "How exactly do they relate to each other?"

Elements of the information-processing paradigm are found frequently in Neuroscience textbooks. Logical sequences like dendrite-soma-axon-synapse or sensory-cognitive-motor are inspired by the flow of information. Thus, the idea about levels of organization is consistent with the information-processing paradigm. Of course, information processing is also an essential framework in Neuroinformatics since the basic computational element is a simple input-computation-output device. In sum, information processing definitely is a shared framework in all three domains. Levels of organization are not only a conceptual framework in the Neurosciences, but are also adopted by cognitive psychologists when they embrace neuroscientific explanations. It is easily possible to embed information processing in levels of organization, while it is difficult or, at least, not common to understand the levels of organization as information processing.

The architectures in Neuroinformatics can also be conceived in the framework of levels of organization (spatiotemporal structure), even though the notion of space and time that is usually meant in the Neurosciences, is sometimes very different to that used in Neuroinformatics. Two meanings of space and time can be distinguished in the Neuroinformatics domain: one refers to physical implementations, i.e. computers, chips, robots etc. The other refers to abstract temporal or spatial differences between data samples or units of computation, i.e. in time series analysis, computational maps etc. The former (physical space and time) is the same that is typically used in the Neurosciences, while the latter (abstract space and time) is usually only meant in the Neurosciences if the topics of neural representation and coding are addressed. A similar distinction holds for cognitive psychology: if the behavioral or the neuronal aspect is addressed, physical space and time (e.g. reaction time) are customary. But if theoretical aspects (e.g. PDP - parallel distributed processing) are addressed, abstract space and time is used. Additionally, the PDP paradigm ‘blurs’ the straightness of both, the temporal domain (parallel instead of before-after) and the spatial domain (distributed instead of before-behind). In sum, the architectures in the Neuroinformatics domain are effectively comparable to levels of organization of the Neurosciences, but it is important to differentiate between abstract (computational) architectures and physical (neural) architectures.

So far, the specific conceptual frameworks fit each other. The intersection between the Neurosciences, Neuroinformatics and cognitive psychology explains *information processing* in specific *architectures* at different *levels of organization*. Information processing can concern neural, cognitive and computational issues, the levels of organization allow a categorization in terms of spatial, temporal and complex quantity and the architecture denotes specific qualities of the system under scrutiny (e.g. learning, recurrent). The common notion of an outcome of these processes can be termed ‘behavior’: the system processes information in order to generate adequate behavior. The adequacy of the behavior is situated and not *a priori* given and is defined in terms of neural, computational or cognitive constraints. Levels of organization are not logically (or lawfully) deduced from theoretically grounded spatiotemporal framework, but rather clustered by general assumptions (paradigms, primitives or doctrines). These are, for instance, the ‘neuron doctrine’, ‘connectionism’ (PDP), the sensorimotor primitive, and the specific functional modes of learning and cognition. Taking these bits and pieces together, the basic concepts are: neuron, network, sensor, motor, learning, and cognition. The relationships between these basic concepts are not strictly hierarchical, but might be better comprehensible in a block diagram (see fig. 15). The basic container of the concepts can neutrally be called ‘system’. The system can have various instantiations: a human, an animal, a robot, a computer, a software, a theoretical model etc. The system is not self-contained, but constituted by specific functional domains: it processes information from ‘Sensor’ (input) to ‘Motor’ (output) and might have the modes ‘Learning’ or ‘Cognition’. Information processing is based on networks that themselves are constituted by ‘neurons’ (understood as basic computational units). Generally, the scheme should not be understood as an ontological claim, but rather as an elaboration of typical conceptions of the neural, cognitive and computational domain, i.e. what is in the head of, say, textbook-authors. With respect to a ‘brain navigator’ on a library computer, the standardized access that can be derived from the common scheme in order to represent the neural, the cognitive and the computational domain has the form of clusters around the basic concepts: neuron, network, sensor, motor, learning and cognition.

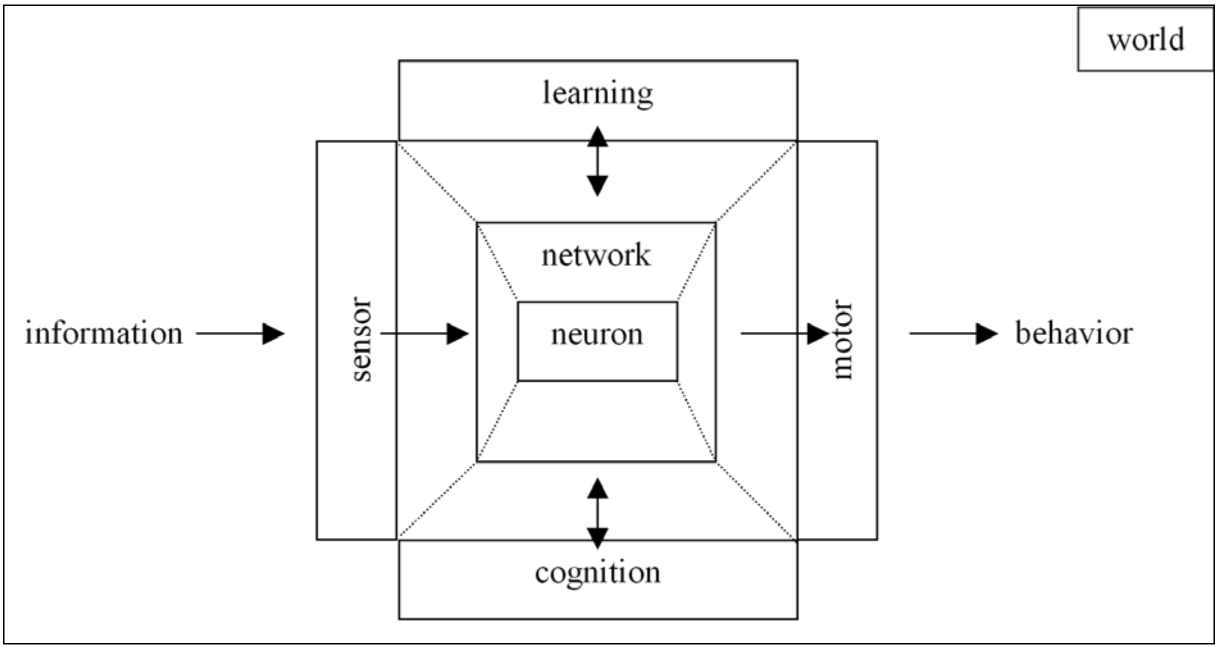
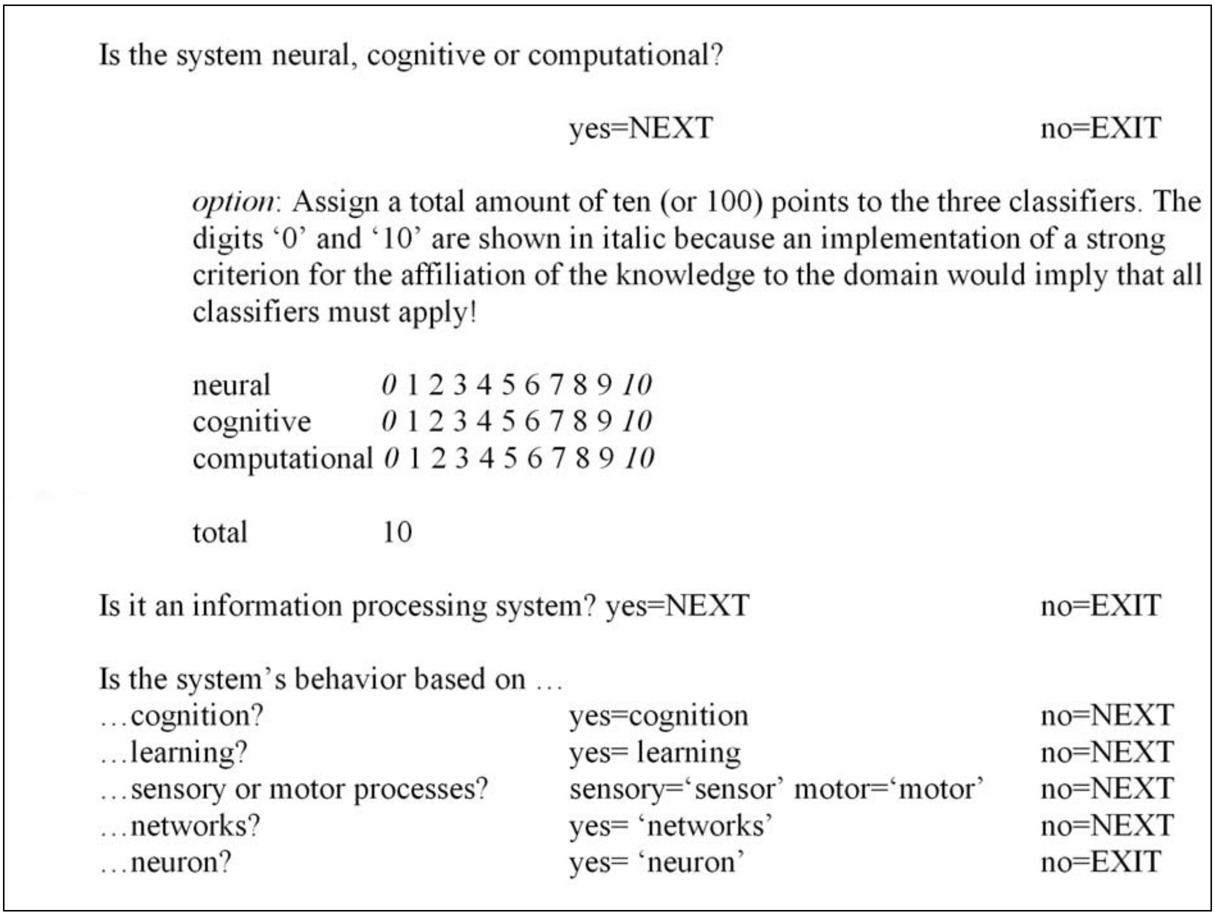


fig. 15: A common scheme in the neural-cognitive-computational domain. Information enters the system via sensors and is processed by neurons organized in networks. Processing is possibly accompanied by learning and cognition and leads to a motor response (an externally observable behavior) that is adequate with respect to the incoming information. The scheme is not proposed here as the adequate approach in brain studies but rather as minimal common denominator in the contributing disciplines.

For illustrative purposes, it might help to conceive the common scheme as an abstract device or creature that lurks behind the thoughts of a majority in the neural, cognitive and computational domain. The concrete form of this device might differ from domain to domain: in the neural domain it might be an animal, in the cognitive domain a human and in computational domain a software or a robot. The abstract incarnation of this creature that could be accepted in all the domains is the Braitenberg vehicle (Braitenberg 1984). In this sense, the common scheme is not so much a perfectly hierarchical structure that can be represented as a directory, but rather an explanatory strategy commonly adopted in the neural, cognitive and computational domain. Consequently, a standardized access to information on brains in a 'brain navigator' might not be adequately realized as a table of content, but rather a representation of an abstract creature. How this could look like shall, for the moment, be left to the reader's imagination...

1.3.6.1. *Applying the common scheme*

In order to assess the validity of the common scheme for the brain sciences, it shall be tested if it performs well as a guideline in a common knowledge base for the neural, cognitive and computational domain. Consider an author wishing to integrate a piece of information into the 'brain navigator'. The first task would be to determine the general affiliation of the piece of knowledge to the common scheme. In a questionnaire that is designed to determine the affiliation, the following question can be posed "Which of the themes are addressed: neuron, network, sensor, motor, learning or cognition?" If only one hit was enough for counting as 'affiliate', a rather *weak* notion of affiliation would be given. In this case, for instance, any educational theme (e.g. any basic school sports curriculum) would belong to the neural-cognitive-computational domain because learning is addressed. Such a weak notion of commonality would not be very informative. If two hits were necessary, the case in which 'learning' and 'cognition' are chosen illustrates that still great arbitrariness is present. For preventing this, the condition must be extended to: "If learning and/or cognition apply, another basic concept must also apply." In this case, two further degrees of affiliation can be distinguished: choosing either 'sensor' or 'motor' (input/output) would indicate a commitment to an information processing approach. This notion of affiliation could be called *moderate* and would imply any information processing system e.g. conventional AI systems. A questionnaire suggesting a *strong* notion of affiliation would also demand the commitment to the essential concepts ‘neuron’ and ‘network’ (PDP paradigm). The strong notion would exclude many systems from psychology and AI and would presumably restrict the potential of theoretical integration by reducing the knowledge base to poverty. (It should be noted that the relationships between the commitments enforced in the questionnaire are nested: a commitment to the PDP paradigm implies a commitment to a general information processing paradigm, while a commitment to a general information processing paradigm does not imply the PDP approach.) For the sake of completeness, the *ultimate* affiliation would be given if all basic concepts applied. However, such a notion was not the objective of the exercise since, for instance, making ‘learning’ or ‘cognition’ obligatory, would definitely restrict the scheme too much. In conclusion, the moderate affiliation appears to be a good approximation for deciding whether a piece of knowledge belongs to the common scheme or not.



Is the system neural, cognitive or computational?

yes=NEXT no=EXIT

option: Assign a total amount of ten (or 100) points to the three classifiers. The digits ‘0’ and ‘10’ are shown in italic because an implementation of a strong criterion for the affiliation of the knowledge to the domain would imply that all classifiers must apply!

neural	*0* 1 2 3 4 5 6 7 8 9 *10*
cognitive	*0* 1 2 3 4 5 6 7 8 9 *10*
computational	*0* 1 2 3 4 5 6 7 8 9 *10*
total	10

Is it an information processing system? yes=NEXT no=EXIT

Is the system’s behavior based on …

…cognition?	yes=cognition	no=NEXT
…learning?	yes= learning	no=NEXT
…sensory or motor processes?	sensory=‘sensor’ motor=‘motor’	no=NEXT
…networks?	yes= ‘networks’	no=NEXT
…neuron?	yes= ‘neuron’	no=EXIT

fig. 16: ***Categorization into the common scheme.*** *A forced choice dialogue determining to which basic concept of the common scheme a given piece of knowledge shall be assigned to.*

After heaving determined whether a piece of knowledge belongs to the scheme, it is necessary to specify the class it belongs to. The common scheme should allow a sound categorization for any piece of knowledge. However, the actual target class (category) "in the head of users" may vary. For example, memory issues can be classified to 'learning' as well as to 'cognition' because the borders between classes are not discrete but rather gradual. Thus, it is adequate to speak of 'degrees' to denote the affiliation of a given piece of knowledge to a given class. A simple multiple-choice questionnaire that only asks whether classifiers apply or not (but not what is the decisive classifier) yields a pattern of choices (e.g. 'learning' and 'cognition' apply for memory issues). This pattern of choices made in the questionnaire contains information on the piece of knowledge and represents relationships between the basic concepts. In a 'brain navigator' on a library computer, these relations could be provided as referenced glossaries or implemented as a "see also …"-function. But such a classification is not unambiguous.

An unambiguous classification of any given piece of knowledge into a directory, can be realized with a forced choice dialogue as it is shown in fig. 16. The basic assumptions of this dialogue are: "network *implies* neuron", "sensor, motor, learning, cognition *implies* network", "sensor *excludes* motor", "cognition *implies* learning", "learning, cognition *exclude* motor, sensor". Certainly, these assumptions contain flaws. For example, "cognition implies learning" seems as well questionable as "cognition excludes sensor". But these flaws are to be accepted if a standardized access shall be provided. So, any system that does not 'exit' the dialogue falls in a specific categorical slot of the neural, cognitive and computational domain. Since, there is a general assumption of information processing in the classification, the dialogue demands a commitment to the information-processing paradigm and, hence, implements the moderate affiliation to the common scheme. In conclusion, applying the common scheme for classifying knowledge is possible, but also reveals some oddities.

1.3.6.2. *Problems*

The common scheme for the brain sciences reflects the conceptual intersection between the neural, cognitive and computational domain as it can be synthesized from textbooks accounts. If there are any inconsistencies, oddities or flaws in the common scheme, they can very well lead to misunderstandings between members of the neural, cognitive or computational domain. This could impede theoretical integration, which was supposed to be supported. Thus, it is important to critically discuss the common scheme for the brain sciences.

A general flaw of the common scheme is the high degree of abstraction that allows for multiple meanings for major concepts being inherent. For example, there are inconsistencies between several terms: as already explained above (see also 1.3.6.2), there is no uniform notion of space and time in the different scientific disciplines studying the domains – a factor that undermines the 'levels of organization' scheme. Nor, is there a uniform notion of information, which would be desirable with respect to the aptitude of the information-processing scheme as a common concept. There are at least two notions: a general, qualitative notion of information that is customary in everyday scientific language has to be distinguished from quantitative information in the sense of Shannon because, unfortunately, the semantics differ substantially. Whereas theoreticians address Shannon information, cognitive psychologists and neuroscientists frequently use the term wrapped in teleological explanations, i.e. information necessary to generate a specific action. Of course, there are a lot more notions of information that shall not be exhaustively presented here (but see e.g. Maynard Smith 2000). In sum, the understanding of general concepts such as space, time and information differ substantially within and between the domains.

There are also disparities between understandings of 'architectures' that bear a given function. These become immediately evident when the term 'learning' is considered: neural mechanisms (LTP etc.), learning mechanisms (conditioning etc.), memory systems (implicit etc.) and formal rules (backpropagation etc.) are not refereed to each other, but are in many cases treated solitarily or separately from each other. As a consequence, researchers from Neuroinformatics (working on backpropagation) and researchers from the Neurosciences (working on LTP) could very well refer to something completely different when they use a term such as 'association'...

Other terminological problems are also prevalent. Recipients in the computational domain will probably have problems with the terms 'sensor' and 'motor'. In Neuroinformatics, 'input' and 'output' are customary. (Generally, in the computational domain the behavioral framework is not as present as in the neural and cognitive domain.) In cognitive psychology the dichotomy perception/action, but also attention/performance is common. Moreover, the class 'motor' can cause misunderstandings because there may very well be cases that do not at all imply motor issues, but are to be assigned at the output level. For instance, 'decisions' as output of cognitive processes or specific activation patterns of abstract networks are interpreted as output. Moreover, dynamicists will probably criticize that sensor and motor are separated and not treated as a closed 'sensorimotor' loop. (However, as an alternative, the principles of feedback can be treated separately from the issues of sensory and motor systems – what is an ever-continuing sensorimotor loop anyway?)

Another problem is that the neuron doctrine is heterogeneously applied in the domains. Generally, two meanings have to be distinguished, a biological cell and an abstract computational unit, usually defined as input – activation function – output unit. Cruse (1996), for example, uses the terms 'neuron' and 'neuroid' to emphasize the difference. In the neural domain, for the most part, biological cells are meant, but in theoretical and computational neuroscience, one can find abstract neurons. Cognitive psychology refers to both to the abstract unit in connectionism and to the biological cell cognitive neuroscience. In the very same sense as connectionism, Neuroinformatics refers predominantly to the abstract computational unit. The varying abstractness of neurons also affects the concept 'network' because neuron and network are interrelated by definition: neurons are the constituents of networks and networks are made up of neurons. Instead of networks, Neuroscience frequently uses the term 'circuits' but also 'tissue', 'ensemble' or 'region' as well as 'area'. The latter two are also customary in the cognitive domain, particularly in neuropsychology or cognitive neuroscience. Finally, cognition is probably the worst defined term (see also 1.2.7.2 and 1.3.6.2). However, for the present study it can be understood as anything that is 'higher', not directly covered by the other basic concepts and refers to sensation and perception, attention, memory, thinking and knowledge or related functions such as motivation and emotion. Cognition seems indispensable to imply functions such as selective attention, problem solving or categorization that otherwise would not be contained, but yet are believed to be explainable in neural and computational terms.

In sum, terminological problems as well as inconsistencies are present in the common scheme for the brain sciences. Such problems are, on the other hand, ubiquitous, even (or all the more so) within a single scientific discipline. Thus, in spite of the terminological and theoretical problems found, it can be concluded that the proposed scheme can be assumed to be common in the neural, cognitive and computational domain. Of course, it would be more instructive to provide a more elaborated scheme, for example to have more subclasses that allow for finer grained knowledge categories. But, for the moment – with respect to the problems that this simple scheme already introduces – it looks very much as if such a scheme is all that is within reach.

1.3.7. Outlook

From the practical side, it can be asked what is left to do for developing such a thing as a 'brain navigator' on the library computer. Presenting knowledge bases (e.g. thesauri) in an appropriate manner could yield a comprehensive knowledge 'landscape' through which the student can actively move or be informed about the actual position. But this mode of navigation might cause difficulties. The student only sees the nearest pieces of knowledge and can therefore only perform landmark orientation. Additionally, even landmark orientation might be difficult since the 'landscape' is not organized along explicit dimensions such as space and time, but determined solely by the relations between the pieces of knowledge. Therefore, a standardized access to knowledge in the brain sciences should also be offered that can guide the student and serve as a map. The common scheme for the brain sciences as proposed above (see also 1.3.6) can be understood as a first step towards a standardized access, but it is definitely too coarsely grained for acting as a guide. A finer graining of knowledge classification can be achieved by mapping the common scheme onto knowledge bases such as thesauri. Concretely, terms are to be assigned to the basic concepts (i.e. neuron, network etc.) and then grouped together by major concepts (e.g. neuron divided into passive, active and synaptic membrane). Ambiguities revealed during the mapping should be notified because they hint at transitions between basic concepts (e.g. memory could be assigned to 'learning' as well as 'cognition') or problems concerning theoretical integration (e.g. contradictory accounts of 'learning'). When the choice of major concepts is completed, a consistency check – similar to the textbook analysis provided above that yielded the basic concepts – should lead a step further towards a conceptual framework. If an even finer graining seems desirable (i.e. a deeper node in the hierarchy), the mapping and consistency check can simply be repeated for the major concepts. This procedure is similar to the above-mentioned mapping of a conceptual framework onto a thesaurus for generating a directory for the brain sciences. But it uses, for lack of a common conceptual framework, the 'homemade' common scheme as template. Additionally, subclasses are not deduced, but 'empirically' selected[6]. Irrespective of the actual realization of a standardized access to information on brains, its emergence appears realistic. As a conclusion of the practical thread of discussion, it can be stated that such a thing as a 'brain navigator' on the library computer is an achievable goal.

Seen from the theoretical side, the prospects of theoretical integration in the brain sciences, for example in the form of a common conceptual framework for the neural, cognitive and computational domain are the central question. The analysis of existing information resources on brains indicated that brain sciences are in an intermediate state of theoretical integration. There are no unified, comprehensive resources for the brain sciences yet, but parts in the domain specific resources show an integrated view on the neural, cognitive and computational aspects. Moreover, the specific conceptual frameworks, i.e. the levels of organization in the Neurosciences, the Information Processing Approach in Cognitive Science and the architectures in Neuroinformatics are consistent, in principle, and even supplement each other. However, the expectations are to be kept low: an integrated 'brain theory' might be hard to find. The relevance of high level theories for biological issues has recently even been challenged altogether by the claim that general laws (as the high level organization principles of theories) might not be adequate approach to explain all biological phenomena scientifically. Rather more specific regularities such as models and mechanism might be the appropriate approach. There is a continuing debate on the existence of general laws (Mitchell 1997; Sober 1997). Thus, if philosophy of science calls into question a lawful, theoretical basis for 'old' biology in general, it seems unwise to claim it for the 'young' brain sciences. This suggests that the claim for a common theory might better be dropped right from the beginning. On

[6] Here, the difference between the common scheme and a conceptual framework becomes obvious: a conceptual framework would allow of theoretically deducing subclasses that could be filled with terms of thesauri (top-down), while the common scheme requires the 'empirical' procedure of analyzing thesauri (bottom up).

the other hand, the brain definitely is a common subject of the disciplines concerned with the neural, cognitive and computational domain. What if the distinction between neural, cognitive and computational domains might turn out to be nothing more than a methodological artifact resulting from the different experimental approaches used, e.g. wet preparations in the Neurosciences, behavioral paradigms in cognitive science, and technical preparations in Neuroinformatics? Then, all the participating disciplines are actually targeting at a single phenomenological 'brain domain' and something like a theory is within reach. Thus, a common conceptual framework that characterizes the brain sciences should, of course, be pursued further on. An explication and optimization of a common conceptual framework should help to propel theoretical integration and mediate orientation for recipients. In spite of the general reservations concerning theories, it should be analyzed further, for example, whether major theoretical works are consistent with a common scheme for the brain sciences as it was proposed here (or any other common scheme). Cross checking the common scheme with existing general theories could corroborate it or call it into question – if Marr's Levels of Analysis (Marr 1982), the Atkinson-Shiffrin-Model (Atkinson & Shiffrin 1968), the neuron doctrine or even evolution theory would contradict general assumptions of the common scheme, it would be alarming. General works that can serve as sources for cross checks can be found within each of the domains (paradigmatic accounts), between domains (interdisciplinary works), below the domains (generative theories, such as mathematics, cybernetics, complexity theory, information theory), above domains (philosophies such as reductionism, materialism, supervenience) and through domains (methods such as imaging, statistics). Furthermore, explanatory frameworks or notions of theory, respectively, for the neural, cognitive and computational domain are brought about by the philosophy of science (e.g. Craver & Darden 2001; Machamer *et al.* 2001) and are to be considered. The multitude of possibilities for further studies stands in contrast to the simplicity of the common scheme for the brain sciences proposed here. This contrast indicates how much there is left to do. However, these tasks are beyond the actual scope and reserved for future work.

1.3.8. Conclusion

As the analysis of resources for information on brains indicates, there are no striking inconsistencies between the neural, the cognitive and the computational domain that could principally prevent their theoretical integration, i.e. that could prevent obvious (e.g. conceptual or disciplinary) boundaries between the domains to vanish. A common scheme for the brain sciences exists in the form of an abstract device applying neural processing principles in order to show adequate behavior in a given situation. On the other hand, it is somewhat unsatisfactory, if not alarming, that the major correspondences between domains are very abstract, sometimes not uniformly applied or even ill defined. There is no common conceptual framework or theory for the brain sciences. Generally, theoretical integration of the neural, the cognitive and the computational domain is likely to happen, but is, at the moment, in an intermediate state between emergence and maturity.

Reading Advice

The overall objective of this study is to examine the role simulation plays in explaining brains. In the first chapter (now finished) it was shown "what brains are about …". The first section of the first chapter introduced an explanatory framework that is applied throughout this study. The following two "Brain-Specials" extended this rather general framework in order to provide novices in the brain sciences and interested brain experts with a more detailed account of what it means to explain brains. The chapter on brains already pointed at some causes why simulation is important for explaining brains: simulations help to control dynamics and complexity found in brains. The following chapter will show "what simulation is about…" In order to understand that it is not trivial to get a grip on the concept of simulation, consider the two following 'everyday' cases: a child unpacking a flight simulator CD on Christmas and a high jumper simulating an attempt before actually starting. These cases are both referred to as cases of simulation, but they are based on completely different notions of simulation: simulation being something external (e.g. a program on CD) and simulation being something in our head (an imaginary situation). Some readers (simulation experts) might be able to easily reconcile these notions of simulation, but others may see no evident connection at all. I wish to unite these different groups of readers by providing a "Simulation-Special" for the non-experts directly at the beginning of this chapter: the text "Simulations as media" introduces different notions of simulation and critically examines the notion of simulation being something external by asking whether there can be a standardized simulation medium. The second section of this chapter "Simulation in science" then picks up the main thread of explaining brains (first chapter) and weaves in the thread of simulation: an analysis of the work of a (brain) scientist will reveal that there is a generic simulation scheme that can reconcile the various notions of simulation.

2. SIMULATION

2.1. Simulations as media[7]

What is simulation? Is it a medium such as a flight-simulator software or is it an imagination such as a high jumper preparing an attempt before actually starting? This question shall be answered by analyzing whether a simulation can be conceived as a standardized medium. Why should the consideration of standards help us to better understand what simulation is about? A standard is a specification of how to conceive something or handle it. Thus, if there is standard for simulation there is also a clear-cut notion of simulation. As the International Standards Organization ISO, for example, specifies standards for tennis rackets, audiovisual engineering and the determination of salt content in butter, organizations like the Moving Picture Experts Group (MPEG consortium) or several IEEE initiatives develop standards for interactive media such as simulation. In the following, the idea of standardizing simulations is critically examined. This brings to light not only the various notions of simulations, but also a sketch of a notion that can reconcile these.

2.1.1. Introduction

The venture of standardizing simulations shows a large gap between desire and reality. On the one hand nearly everyone concerned with media asserts that simulations play a prominent role beside films, texts etc. And nearly everyone claims to know what simulations are about (at least tacitly). On the other hand neither a common sense nor a binding formal specification is visible. Expressed in terms of computer applications, a “save as …” button for simulations has been scarcely realized. Without a common sense, the development of simulation-specific formats, authoring tools, metadata,

[7] The text, originally titled *Standardizing Simulations – “Uphill all the way!”* was written for the congress *Virtual Campus 2002* of the Association for Scientific Media in German speaking countries (“Gesellschaft für Medien in der Wissenschaft”) and provides an analysis of the venture of integrating standards for simulations as media (Horstmann 2002). Here is the abstract of the original abstract: *“Simulations are capable of representing complex and dynamic knowledge by being inherently functional. Despite this extraordinary capability – not realized in any other medium – no widespread standards for simulations as media have prevailed. Considering the conceptual difficulty, the semantic variety and the specialization in complicated content, however, the lack of standards is no surprise: the versatility of simulation takes a heavy toll on their potential of standardization! Moreover, the provision of inherent functionality necessitates that users decide what the simulation will be like and forces them to make the corresponding interventions. These active cognitive and behavioral processes inescapably introduce a human factor that cannot directly be included in standardization ventures. Since the principle of simulation is based on the human factor, it is concluded here that attempts to standardize simulations can only be successful if they focus on the human factor – a work that eventually implies enduring research and development processes.”*

ordered databases, quality standards, evaluation guidelines or instructional designs is hardly achievable.

A prominent illustration of the present situation is given by the Learning Object Metadata initiative IEEE-LOM (2003), probably the best known approach seeking to introduce classification standards ('metadata') for educational media: according to the LOM-Specification, simulations are conceived as a specific "Learning Resource Type". But beyond that, simulations just serve to exemplify learning objects showing an "active interactivity type", a "high interactivity level" and "high or low grades of semantic density". Hence, an unbiased reader of the LOM specification can take home the message that an important role of simulations is acknowledged. But apart from a characterization as 'something interactive' the conception of simulations is void.

But do we really need more specific standards? For answering this question, consider a teacher looking for usable simulations within the countless applets on the internet (e.g. simulations on the "traveling salesman", a famous formal problem asking for the optimal order to deliver goods to numerous recipients that is often mathematically solved by way of artificial neural networks). Suppose, the search yields about 100 different simulations (a minimalist estimation). Which one is the best? Which one to take for which instructional setting? Which one is evaluated? This case illustrates that there is a huge resource, but it is hard to utilize it without more specific standards for simulations. Standards could help to assess *in advance* what is to be expected when a specific simulation is first encountered. Serious efforts have been made in order to condense the medial aspect of simulation. There are several attempts to provide classification systems for simulations (see e.g. Schmucker & Apple Computer Inc. 2000; Fishwick 1995) and countless programming approaches and mark-up languages, but none of them did break through in a way that could serve as a guideline in standardization ventures. Continuing research traditions on simulations in psychology, education and artificial intelligence have been successfully pursued for decades. But, obviously, they didn't flock together. With respect to all these efforts the question is: Why didn't emerge a common sense for simulations in a way that there is common sense, say, about what a film or a text is? At least some kind of common sense, obviously, would be the minimum demand for any standardization venture.

In sum, the way to simulation standards is definitely explored, but in the present situation too many different paths sidetrack from a way straight-ahead. Therefore – rather than outlining yet another path – this text, in the first step, seeks factors that explain why the different paths do not effectively converge towards common sense. In the second step, it is attempted to peel out specific features that point the way towards the core of simulation. The third step previews how such core characteristics of simulations could be employed to yield a classification system that in turn could be applicable in standardization ventures.

2.1.2. Impediments for simulation standards

Three major factors impeding the emergence of common sense and standards are considered here: first, *conceptual difficulty* arises from the simulation's characteristic to be inherently functional (that is to be organizational open) and a non-trivial conceptual structure that encompasses three levels of meaning: simulated system (source, 'simulandum'), simulating system (model, 'simulans') and implementing system (simulator, see fig. 17). These difficulties are 'supported' by the closely related and no less complicated sub-concepts 'representation' and 'model'. As a result, multiple notions of what is meant by simulation in a given situation are possible. These different notions prepare the ground for the second impeding factor: *semantic variety.* At least five major accounts of simulation can be found when combing through scientific databases: 'social' often in the form of role-plays (Heitzmann 1973), 'gaming' (Crookall 2001), 'device' as in cockpit simulations (Kieras & Bovair 1984), 'model' (formal-mathematical) and 'cognitive' simulation (Johnson-Laird 1980; Johnson-Laird 1983; Gentner & Stevens 1983; Barsalou 1999). Third, to top it all, simulations are specialized in bearing *complicated content*; i.e. the represented knowledge is usually dynamic and complex – possibly exactly because otherwise the use of such a difficult concept would not be justified. In other words, the conceptual difficulty might be viewed as prerequisite that allows for the representation of complicated content. In sum, facing these impediments, the lack of common sense on simulations appears to be understandable. In the next sections, the difficulties shall be analyzed more detailed.

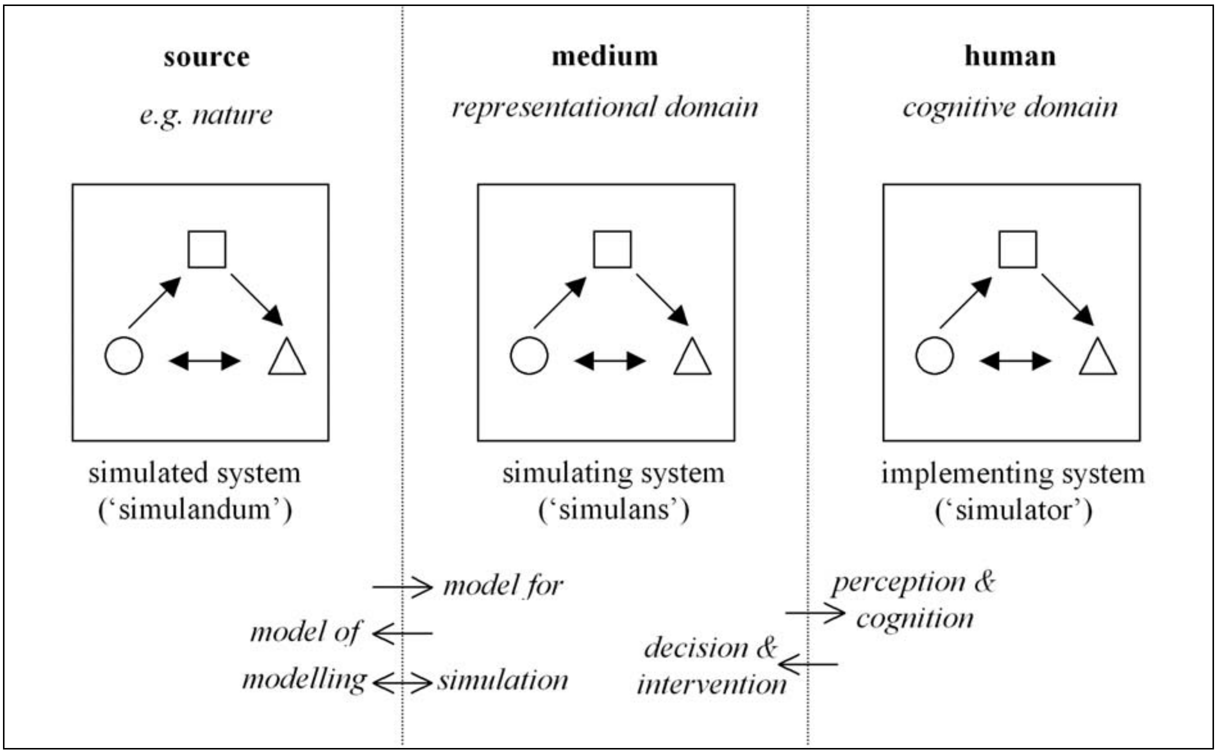


fig. 17: ***The concept of simulation.*** *Simulations refer to a certain system (box with symbols). The system shows inherent functionality, which results from the activities (arrows) of the constitutive elements (symbols) according to certain causal rules. At least three levels of simulation can be distinguished: The source denotes the simulated system ('simulandum'). The knowledge structure ('simulans') is located in an abstract representational domain. The relation between simulans and simulandum is that of modeling. Simulations depend on interventions and are therefore instantiated in a simulator, e.g. a cognitive system.*

2.1.2.1. *Conceptual difficulty*

Simulation is an abstract term and, therefore, particularly prone to misunderstandings. A very common misconception, for example, is to mix up simulation with animation. What are distinctive features of a simulation then? Merriam-Webster's dictionary (2003) paraphrases simulation as *"the imitative representation of the functioning of one system or process by means of the functioning of another*"[8].

An essential concept in the paraphrase is "*system or process*". Since a process, in this context, can be conceived as a certain order of the system's states (in time or in a logical order), I will only use the term "*system*" in the following. The representative aspect of a simulation implies that a simulation is not exclusively defined by one entity, but always refers to a second entity.

[8] Of course, there are many paraphrases and definitions. However, the concrete paraphrase can be chosen somewhat arbitrarily, because it shall not motivate a universally valid account of simulation, but rather serves as a starting point for discussions on terminological problems of simulations.

Thus, a simulation refers to two systems, in the following called system A and system B. Consider system A as the simulation and system B as the "original" system, i.e. the system to be simulated. In the style of the distinction between the "explanans" and the "explanandum", I will call A the "simulans" and the B the "simulandum" (see fig. 17). The relation between A and B is described as an "*imitative representation*". This relation can be conceived straightforward (like conventional media) as the relation between an entity and a picture of that entity. "*Imitative*" indicates that the representation has a specific purpose, namely not merely to depict, but to reproduce. Reproduced is the "*functioning*". Thus, the analysis so far provides at least one specific feature of simulations, namely "*inherent functionality*". Represented is not the system itself but the functioning of a system. The system's elements show certain activities according to specific regularities. Elements, activities and causal relations "in action" realize the functioning of the system.

Inherent functionality is a characteristic specific enough to differentiate between simulations and other media. Animations, for example, are not inherently functional. They do not aim at reproducing the functions of the system; they depict the functioning by applying dynamic images.

However, it is exactly this inherent functionality that gives rise to two further conceptual difficulties. The first can also be illustrated by the animation/simulation distinction: simulations must be based on models in order to reproduce the functioning of the system, while animations can depict the functioning just by showing the phenomenological behavior of a system that might be based on no model at all or even a wrong model. Thus, simulations imply models and, therefore, a theory of simulation necessitates a theory of models. Unfortunately, there has been continuing debate about the question whether there can be principally a theory of models and it does not seem is to find an end (see e.g. Magnani *et al.* 1999 for a general account). A review of these issues would definitely be beyond the scope of this article. But it is important to note these circumstances in order to understand the potential sources of the conceptual difficulties of simulations[9].

[9] *The Problem with 'Model simulation'.* It is suggested here that promising approaches to design simulations as media might probably be found if the aspect of models is put to the center. However, if we used the term "model simulation" we might have gained a more telling account, but we eventually land ourselves with new problems. First of all, it seems to be difficult to think of a simulation that does not imply a model! As already explained, simulation

The second conceptual difficulty resulting from inherent functionality relates to the issue of representation. As in the case of models, there is no unified theory of representation (see also 3.1.3.1). Like all other media simulations represent (i.e. "carry information") about something else. And, again like all other media, simulations only represent something in case they are used. (A picture of the Eiffel-tower lying in the desert without someone noticing it does not actually represent something.) But simulations depend on use in a special manner: the functionality inherent in the simulation is only represented if a user makes interventions, i.e. it does not only depend on perception and cognition but also on decided actions. This distinctive feature of simulations renders the attempt to find a common-sense concept of simulation extremely difficult because it introduces an undefined human factor (a "blank") into the concept of simulation that demands a triadic organizational scheme of simulation (see fig. 17). These relations between the three compartments of this organizational scheme will be discussed later (see also 2.1.3).

In sum, the conceptual difficulty might explain a large part of the question, why no common sense on simulations emerged. The term "simulation" invokes an organizational scheme that may encompass three different levels: a simulated system (the simulandum), the simulating system (the simulans) and a user that processes the simulations (the simulator) make up a simulation. The manifold possible relations within and between levels and the dependence on human interventions endanger the concept to be vague.

can be conceived as a simulandum that is represented by a simulans. On the one hand, the simulandum can be conceived as a model for the simulans, on the other hand the simulans can be conceived as a model of the simulandum. It should be noted that this relation is mutual but not reciprocal as indicated by the different prepositions of model ("model of" vs. "model for"). That means, the representation as such could be conceived as a model. (You can exchange "imitative representation" with "modeling" in the paraphrase used above.) Thus, it could be argued that the concept of model simulation is redundant or even circular. The second problem with the supplementary term 'model' relates to the semantics. Model might have even more meanings than simulation. Since even philosophy of science has a continuing debate about models (see e.g. Magnani *et al.* 1999), I will not go into detail here. Thus, it might be concluded that the combination of the term model with the term simulation multiplies the probability of misunderstandings. Nevertheless, it is understandable that the term is common: the extra use of the term model indicates explicitly that the simulation refers to a formal model. Furthermore, the process of modeling (that must have been passed before) is highlighted. As a side effect the deceptive notion is avoided. Finally, the term 'model simulation' is commonly used and, actually, in many cases media are meant. In this respect, the use of the term is justified by the use of the term. What is the recommendation following from all these pros and cons? The use of the term 'model simulation' does not help in the long run. But what is meant by the people using the term presumably leads to a precise account of simulations as media because the 'modeling community' has the most advanced and formalized approach that should help to design simulations as media. (Frankly spoken, I would be happy if I could omit the 'model' supplement, but I am afraid that more people will misunderstand what I mean when I just say 'simulation'.)

However, the feature of inherent functionality provides a means for differentiating between simulations and all other media.

2.1.2.2. *Semantic variety*

Almost inescapably, the aforementioned difficulties to conceptualize simulations result in a 'rich' semantic variety. In order to coarsely categorize different accounts of simulations, some of them shall be considered in the following. Since there are too many different accounts of simulations to deal with all of them, only selected domains that frequently "pop up" during inquiries on (various) scientific databases shall be considered. Moreover, I will focus on the last decades since before then the term was mainly used in a purely abstract sense (see e.g. Baudrillard 1988) or in the sense of deception. Both of these notions will not be considered since they do not lead to an understanding of simulations as media.

a. Social Simulation

An early use of simulation can be found in relation to social situations (see e.g. Heitzmann 1973), especially in the business or management and in the political domain. The objectives range from personnel training for specific (social) skills, assessment of applicants, conflict resolutions ('war games') and the control of complex decisive situations. This account of simulation is often designed as role-play or board game and does not necessarily rely on computers.

b. Gaming Simulation

The "gaming" domain is similar to social simulations insofar that role plays or especially training board games might count as the precursors of fun-oriented games. In its computerized form this simulation domain is probably today's most popular as indicated by the commercial success of the various computer products (e.g. Simcity™ to name only a 'classic'). Despite its 'funny' notion gaming can have serious implications for the development of simulations in general (see e.g. Crookall 2001).

c. Device Simulation

Another customary simulation account is, as social simulation, grounded in the training domain. Device simulations were used early and are still used for training controllers of all kinds for mobile devices (e.g. pilots, astronauts) or local devices (e.g. plants, machines, robots). Traditionally, device simulations are implemented as hardware (i.e. as a cockpit), but nowadays there are probably outnumbered by computerized devices (i.e. flight simulator). The example of the flight simulator illustrates that device simulations can be gaming simulations.

d. Cognitive Simulation

Any simulation has a correspondence “in the head of the user” (see fig. 18). Approaches to such cognitive representations of simulations can be termed cognitive or mental simulation. Approaches from cognitive psychology are most noteworthy in this context. Promising candidates for capturing the cognitive part of simulation are schemas (see e.g. Rumelhart 1980) or mental models (especially Gentner & Stevens 1983 but also Johnson-Laird 1983). Mental simulation is directly addressed and related to other cognitive theories by Barsalou (1999).

e. Model Simulation

A large part of ‘everyday’ simulation is covered by the formerly presented accounts. Nevertheless, the formally oriented simulation domain is missing. Consider an economist doing statistics in order to predict the gain of stocks, an engineer evaluating a circuit design for a power plant in a laboratory setup with resistors and capacitors or a biologist analyzing motion vision in an artificial neural network. All of these simulation types (more or less) explicitly refer to theoretical or formal models. The economist assumes a statistical, the engineer an electro-technical and the biologist a functional model. Since the economist’s situation can be conceived as a social simulation and the engineer’s as a device simulation it is evident that model simulations – like the other accounts – do not form an exclusive account of simulation. Model simulation the most promising candidate for the venture of designing simulations as media. It is broad enough to include a wealth of different simulation situations and demarcates from social or mental simulation (that do not directly refer to media), but rather include the media-related forms of, for example, gaming and device simulations. And it is based on formal grounds.

The accounts presented above illustrate the semantic variety of simulations. They do not form exclusive domains, but rather are possible *roles* of simulations that resulted in certain application contexts or development traditions. Thus, they can be related to each other or contained in another. However, as can be concluded from a missing common ground in terms of theory, methodology or terminology, the state of theoretical integration between the various accounts does not appear to be very advanced. In practice, people may very well find themselves talking at cross purposes (consider a meeting of an social simulation expert and a model simulation expert.) Thus, semantic variety provides another intelligible reason for the lack of a common sense on simulations.

2.1.2.3. *Complicated content*

The third major factor that impedes the emergence of a common sense on simulations refers to the kind of knowledge that is represented in simulations, i.e. the content. Consider the motivation to simulate: simulations are usually carried out because ‘reality’ (the simulandum) is not sufficiently tractable – it is too difficult, too complex, too dynamic …, in short too complicated (see below). Thus, simulations usually represent complicated content. Systems with simple functionality do not necessitate simulations because they can be readily explained in words (e.g. “a bread slicer functions by a applying a sharp edge to the loaf”). But especially systems showing dynamics and complexity call for simulations (see also 1.1.11). Their behavior often is not predictable and therefore examined by probing the system with parameter variations frequently accompanied by dynamic visualizations of the resulting changes in the system’s behavior. The result of the learning process might in large parts not be explicit knowledge about the world but rather implicit skills in controlling the system (a circumstance that complicates the assessment of learning success). On this view, the resulting knowledge is often non-declarative or procedural and hard to express in words (e.g. Anderson 1980). Thus, another factor impeding the development of common sense for simulations becomes obvious: It would be naïve to assume that a medium specialized in representing complicated content could be convincingly standardized by defining a few simple classification criteria. On the contrary, in order to represent complicated content, simulations have to be extremely versatile.

It should be noted that complicated content is not assumed to be characteristic for specific knowledge domains, e.g. physics, biology etc. Complicated content is assumed to result from general properties of the knowledge represented in simulations, e.g. dynamics, complexity etc. The ability to represent complicated content drives frameworks for simulations to represent any imaginable kind of knowledge (e.g. MatLab™). Not without good reason, simulation environments are endowed with all prerequisites for a 'production system' in the sense of Andersons Act-R (see e.g. Anderson 1993). However, versatility and arbitrariness are two sides of the same coin. The versatility of simulation takes a heavy toll on standardization.

In sum, three major factors impeding the emergence of common sense and standards are proposed: the conceptual difficulty refers to the triadic organizational structure of simulation with *simulandum, simulans and simulator* (e.g. nature, knowledge, cognition) that is - mutually potentiated by such complicated sub-concepts as 'representation' and 'model'. As a result, multiple accounts of what is meant by the term simulation in a given situation are possible. These different accounts, together with the domains to which simulation can be applied, prepare the ground for the second impeding factor, namely semantic variety. Finally, the content, i.e. the represented knowledge, is usually complicated (e.g. dynamic and complex). The capability of representing complicated content might be viewed as a cause for conceptual difficulty and semantic variety.

2.1.3. Towards the core of simulation

2.1.3.1. *Simulation and media*

The impediments explained above raise the question where to begin with standardization attempts? Usually, standards are to be applied to media. Consequently, simulations should be conceivable as a certain type of medium that runs on a device (as a film that runs on TV). On this view, the simulation would be what is left when the device is removed. If people were asked what device could be taken for realizing simulations, the answer would in most cases presumably be: a computer! The obvious reason is that the 'inherent functionality' of simulations (see above) depends on an implementation that goes beyond plain rendering. In order to provide all the facilities necessary to process and control a simulation, a versatile device as a computer seems to be indispensable. But the fixed focus on the computer can also hinder a clear view on the essentials. Think of simulations on cell phones or, particularly, on TV accessible via digital satellite receivers and used with remote controls.

Such applications might be tomorrow's standards[10]. However, simulation may even happen without any technical help, e.g. in social simulations (role-plays). In these cases humans provide the device and the medium is spoken language (sometimes combined with print media containing definitions and rules). Finally, simulation may also happen exclusively in the cognitive domain. For example, consider an athlete (e.g. a high jumper or a bob pilot) cognitively simulating the task before starting an attempt.

The case of cognitive simulation is most interesting for attempts of standardizing simulations. It shows that simulation can very well be given without any tangible representation as a medium. The representation can be exclusively cognitive. Of course, to a certain degree this situation applies to all kinds of media: a picture, a text or a film is only a functional medium when perceived and processed in some way. But a pure cognitive representation of films and texts is not easily conceivable. (It seems easier to conceive the cognitive form of films or texts *as* cognitive simulations.) On the other hand, films or texts as pure media have a straightforward meaning (videotape or book). Thus, films and texts are well defined by being a specific medium[11]. Certainly, a simulation can be represented on CD (e.g. SimCity™), but a simulation on CD appears not to be as complete as a film on videotape or text in a book.

A possible answer to the question what is actually missing could be: simulations are generally not consumed, like films and texts are consumed, but have to be *done.* Doing a simulation requires actions and actions require decisions. Simulations are incomplete as long as nobody decides what shall happen and carries out the corresponding interventions[12]. Consider a rehearsal of a text or imagery of a film and compare these to simulation. More specific, compare (perhaps facilitated by closed eyes) a mental film of the high jumper's attempt to a cognitive simulation of the high jumper's attempt. In the case of the film the trajectory is fixed and the result is known.

[10] It should be noted that the MPEG-Consortium (MPEG 2002) has acknowledged this challenge in their specification process of the forthcoming MPEG-formats 7 and 21 that attach more importance to interactivity.

[11] In simulations, the functioning of the medium as such is modified. Such decision processes are not possible in films or texts. Changing the functioning of films or texts would mean to intervene in the plot, e.g. by changing the character of a role. Of course, a DVD offering different ends for films or texts offering several strands of the plot might be conceived as marginal cases. However, they are not distinctive features of the respective medium.

[12] Since the term 'interaction' gives rise to uncertainties about the causal direction of a relation and, particularly, since it does not clearly express that a simulation is driven by the user's decisions (e.g. by parameter variation) the term 'intervention' is used.

In the case of a simulation, the trajectory still has to be determined – according to different hypothesis certain steps can be exchanged, varied, tried anew etc. – and the result of the simulation will depend on the decisions and interventions made in the runtime of the simulation. While the cognitive processes that accompany films and texts are primarily media-driven, cognition that accompanies simulations has to drive the medium.

In sum, the attempt to distinguish between media and device fails in the case of simulations because the device still has to specify what the medium will be like. On this view, simulations are characterized by a lack of specification. This raises serious questions for any standardization attempt: How can we standardize a lack of specification? How can we expect a self-contained format for simulations when there will always be blanks in simulations that have to filled in by human decisions[13]? On the other hand, the analysis above peeled out a *human factor*; namely decisions (and eventually interventions) carried out in runtime of simulations that distinguish simulations from other media. Thus, when standardization attempts come in at this point, there is a chance of finding adequate criteria for specifications.

2.1.3.2. *Standardizing humans?*

Standards usually refer to media, not to human decisions. How can a standard for simulations that incorporates human decisions then be accomplished? Even though there might be no clear-cut between medium and human, there is still possible distinction between a medial part and a cognitive part of the simulation. Since there is no way to standardize the human decision process itself, the place nearest to the decision process has to be chosen. This place is at the interface between medium and human. An adequate conceptual framework has to encompass the medial and human part and the respective interfaces (see fig. 18). Consider a user in front of a simulation-device (e.g. a computer with monitor) starting a simulation, say, of a thermostat with a control instrument (e.g. mouse). As stated above, the simulation shows an inherent functionality. In order to unfold that functionality the user has to act on the simulation. Interventions change the state of the simulation from $S(t_0)$ to $S(t_1)$ that might be monitored (in most

[13] Maybe, it is the difficulty of this task that hindered the most an emergence of a unified conception of simulation as media. Providing space for decision processes means to provide certain degrees of freedom – and providing degrees of freedom is something that directly contradicts the nature of standardization. It is hardly conceivable that this field of tension is easily overcome.

cases visually) and fed back to the user. Cognitive processes referring to the new state at *t₁* of the simulation close the circle. Another intervention establishes a feedback-loop. Then, the user is embedded in the simulation-cycle.

The means by which the user can inform the medial part of the simulation about the decisions is an intervention that can be transmitted through the input devices. The intervention is received at a specific interface between medium and human. Such *intervention ports* are usually realized as buttons, sliders etc. In the absence of methods that directly include the decisions, they can be constrained by providing a limited number of intervention ports, providing them at certain places, at certain moments in time etc.

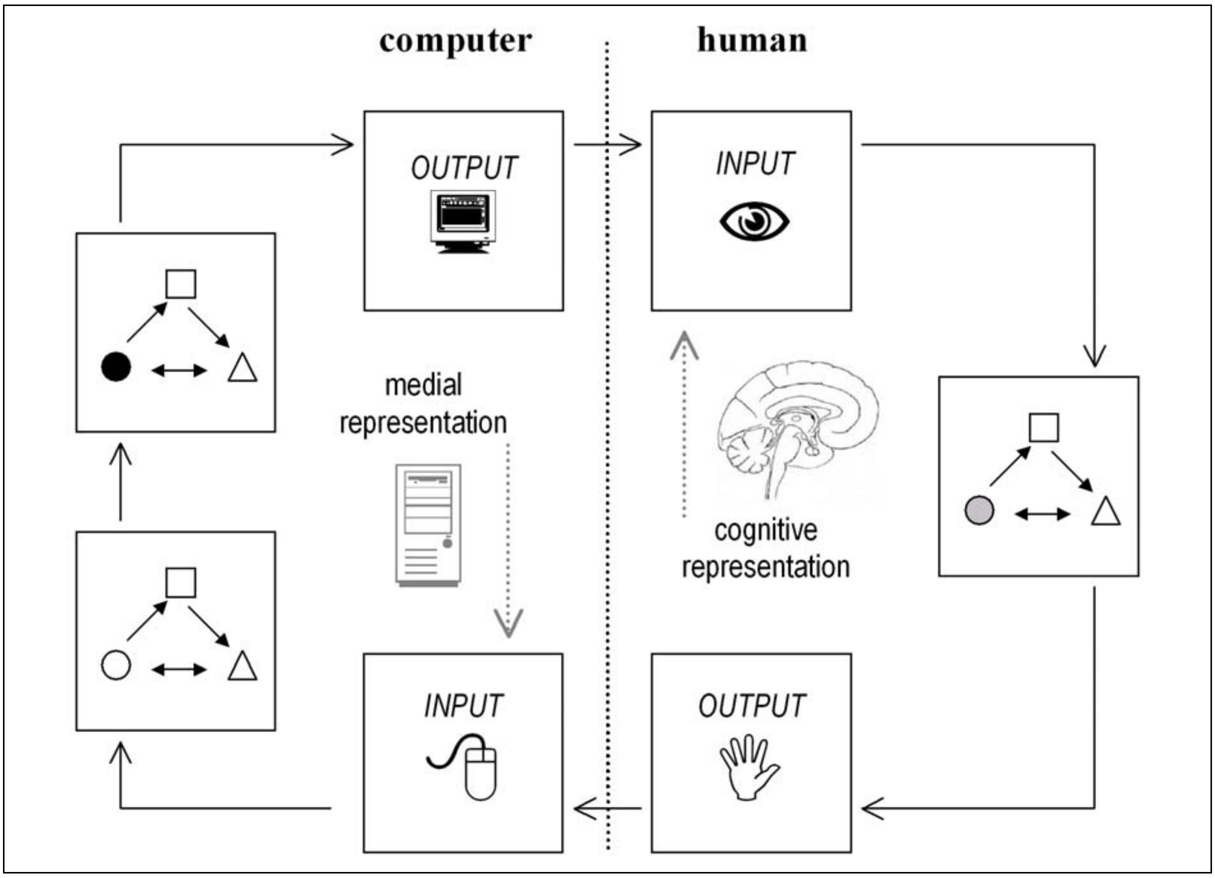


*fig. 18: **Simulation cycle as human-computer interaction.** The user receives sensory input of a system's medial representation, processes it and decides to start an intervention that is realized via behavioral outputs and device inputs. The intervention affects elements in the medial representation of the simulated system at a certain point in time (t_0). These changes cause changes in other elements of the system due to certain causal relations. The overall result is a new state of the system (t_1) that is presented on an output device and processed again. Arrows indicate temporal order. Dotted arrows indicate that a process is carried out only virtually.*

As prescriptions for the design of intervention ports, standards reflecting the human factor could well be introduced. Intervention ports have the neatest correspondence to human decisions in the medial representation of the simulation.

2.1.3.3. *A theory of cognitive simulation*

Practitioners might maintain that – even without any explicit consideration of prescriptions for intervention ports and a theory about the corresponding cognitive processes – there are many examples of well-designed simulations. Indeed, we have a tacit understanding of intervention design (i.e. we know where and when we have to place a button or slider)[14]. But in order to state *explicitly* and explain *causally* what factor enhances or weakens the simulation we need to test systematically (along the lines of a theoretical framework encompassing the corresponding cognitive processes). The analysis above showed that simulations represent complex and dynamic knowledge by providing systems with inherent functionality that are to be operated by specific interventions. According to this description, a cognitive theory must explain:

1. the representation of complexity and dynamics
2. how representations can be inherently functional
3. inferences based on these representations
4. how decisions and interventions are inferred and carried out
5. the general correspondence between medial and cognitive simulation (i.e. provide a representational framework)

Cognitive psychology offers several approaches of complex or ‘molar’ knowledge structures, e.g. schemas (Bartlett 1932; Rumelhart 1980; Mandler 1984), frames (Minsky M. 1975), scripts (Schank R. & Abelson R. 1977) and mental models (Johnson-Laird 1980; 1983; Gentner & Stevens 1983). According to Brewer (1987), the former three can all be subsumed under schemas, while mental models have to be distinguished from these. Brewer describes schemas as unconscious mental structures underlying the molar aspects of human knowledge and skill that involve ‘old’ generic information. Mental models, then, shall account not only for ‘old’ information but also for

[14] The design of simulations might be so familiar to us because cognitive simulation is a natural way to compile knowledge (cf. Barsalou 1999). Moreover, the designer’s (author’s) method to anticipate what the user will do is conceivable as that kind of mental simulation as it is used to explain folk psychology (cf. Gordon 1986).

situations we have never been in before, i.e. they demand imagery and inference. Concerning the differences between schemas and mental models, Brewer points out that schemas are precompiled generic knowledge structures, while mental models are constructed at the time of use. Thus, with respect to the demands 3 and 4, mental models clearly outperform schemas as a candidate for explaining simulations. (Even though schemas might not be the right choice for explaining inference, they definitely play a role in explaining the 'precompiled' parts of a mental model.)Inside the research tradition of mental models two different threads have to be distinguished: one referring primarily to Johnson-Laird (1980; 1983) and one referring primarily to Gentner and Stevens (1983). According to a distinction that Markman & Genter (2001) suggest, each approach might play its specific role in explaining simulations – Johnson-Laird's in the explanation of *logical* models and Gentner & Stevens' in the explanation of *causal* models (see also 3.1.2).

Obviously, there are many more sources to be taken into account for explanations of the cognitive aspects of simulation. For example the issues of implicit learning (Berry & Broadbent 1988), procedural knowledge (Anderson 1993), complex problem solving (Dörner & Wearing 1995; Funke 1992) or general cognitive architectures (Anderson 1993; Johnson-Laird *et al.* 1987) certainly provide rich resources. Most notably, there are accounts directly addressing the issue of cognitive (mental) simulation (e.g. Barsalou 1999). But, a comprehensive review of these theories and an evaluation in terms of the aptitude for explaining simulations would be beyond the scope of the present context (see, however, 3.1). In a first step, it is sufficient to put to the record that mental models provide a theoretical framework that can principally meet the demands of a cognitive theory of simulation and that numerous further specific theories can supplement the mental model theory.

2.1.4. Practical implications

In the absence of widespread standards, each project that deals with simulations can contribute significantly to standardization ventures in that it stringently integrates specific features of simulation into their architectures (databases, metadata-tools, experimental setups etc.) and test their practical value. An example how this can be accomplished is given in the following (see table 7).

table 7: Examples of criteria characterizing simulations. See text for detailed description.

<table>
<tr><th>feature</th><th>attribute</th><th>Comment</th></tr>
<tr><td>INTERVENTION FEATURES</td><td></td><td></td></tr>
<tr><td>intervention type</td><td>[passive | scalar | discrete | continuous | immersive | …]</td><td>ranging from minimal to maximal</td></tr>
<tr><td>intervention depth</td><td>[trigger | visualization | parameter variation | element design | system variation | system design | …]</td><td>ranging from superficial or deep</td></tr>
<tr><td>intervention ports</td><td>[0 | 1-5 | 6-20 | 21-50 | > 51]</td><td>number, classification arbitrary</td></tr>
<tr><td>…</td><td>…</td><td>…</td></tr>
<tr><td>SYSTEM FEATURES</td><td></td><td></td></tr>
<tr><td>system representation</td><td>[text | symbolic | graphic | …]</td><td>comprising sheer naming (e.g. filename), full-text description, formula, schemes, 3D-models etc.</td></tr>
<tr><td>system behavior</td><td>[digits | data visualizations | animated graphics | …]</td><td>e.g. digits in a command line, plots, oscilloscopes, dynamic 3D</td></tr>
<tr><td>update procedure</td><td>[static | stepwise static | stepwise dynamic | continuous dynamic | …]</td><td>describes what changes caused by interventions are computed and shown</td></tr>
<tr><td>variables</td><td>[1 to 5 (S) | 6-20 (M) | 21-50 (L) | > 51 (XL)]</td><td>number, classification arbitrary</td></tr>
<tr><td>connectivity level</td><td>[> 1 | ~1 | < 1]</td><td>quotient of variables and connections</td></tr>
<tr><td>connectivity type</td><td>[directional | mutual]</td><td>naming of predominant type</td></tr>
<tr><td>feedback</td><td>[0 | 1 | n]</td><td>order of feedback, n is given by the number of interconnected feedback systems</td></tr>
<tr><td>learning</td><td>[yes | no]</td><td>system stores previous states</td></tr>
<tr><td>…</td><td>…</td><td>…</td></tr>
<tr><td>COMBINED FEATURES</td><td></td><td></td></tr>
<tr><td>coverage</td><td>[0 < x < 1]</td><td>quotient of variables and intervention Ports</td></tr>
<tr><td>size (time of use)</td><td>[0 < x < n]</td><td>quotient of variables x connectivity level x … related to intervention ports</td></tr>
<tr><td>…</td><td>…</td><td>…</td></tr>
</table>

2.1.4.1. *Intervention features*

How can the interventions that drive simulations be characterized? The *intervention type* might be passive (start/stop of a sequence), scalar (slow motion, spatial resolution etc.), discrete (setting of initial-conditions/discontinuous parameter variation), continuous (effects of parameter variation visible without further operation) or immersive (parameter variation directly changes system representation). The *intervention depth* describes how the simulated system is affected. Ranging from external to internal, the system can be affected by way of trigger, visualization, parameter variation, element design, system variation, or system design. Intervention type and intervention depth are just two examples of simulation features that refer to the human factor.

2.1.4.2. *System features*

Independent from the human factor, but indispensable for assessing and classifying the simulation are the system features. A minimalist form of a system representation would be the sheer naming of the simulandum. Other forms are e.g. full-text, formula or graphics. Dynamics and therewith the *system's behavior* is mediated by process representations that can encompass digits in a command line representing the state of an element, data visualizations (e.g. plots or color-coded schemes) or animated graphics resembling real-world situations. These representations of processes can be further characterized by the *update procedure,* which can be: static (the behavior is visualized as a simple plot, e.g. representing an input output relation, no intervention possible), stepwise static (interactive plotting of states, one datum per intervention), stepwise dynamic (triggering one sequence after initialization, e.g. a 'sweep' shown in an oscilloscope) or continuous dynamic (effects of interventions are visualized dynamically in runtime). Complexity can be characterized by the number of variables contributing to the functionality and the connectivity between them. For practical reasons the *number of variables* could be classified: 1 to 5 (S), 6-20 (M), 21-50 (L), > 51 (XL). Several cases of *connectivity levels* could be taken into account: each variable affects each other variable (>1), one variable affects one (~1) or less than one (<1) other variable. It might be practical to distinguish *connectivity types* on the basis of specific architectures (hierarchical, serial, parallel, layered etc.). In the spatial domain, connectivity can be predominantly directional or mutual. In the temporal domain connectivity can be 'feed-forward' or 'feed-back'. If specific rules change the connectivity as such (and the changes are stored) *learning* takes place in the simulation. The type of operators used could further describe functionality: it can be qualitative, logical ('and', 'or' etc.) or relational ('more', 'less' etc.), or quantitative (numerical). However, the specification should be left to formal experts.

2.1.4.3. *Combined features*

The above named criteria describing intervention and system features can be combined to form further telling criteria. For example, an important feature of simulations is that not all of the variables are accessible to the user. In most cases – especially in the educational domain – the challenge of intervention design is to provide only 'relevant' intervention ports to specific variables, while the 'irrelevant' variables are hidden. This *coverage*[15] can be defined as the ratio of the number of intervention ports and the number of contributing variables (0 > x > 1).

In a similar combinatorial fashion, the size of a simulation can be defined as the state space of the system, e.g. by merging the number of variables with depth and type of spatial and temporal connectivity and relate this to the number of intervention ports. Such a feature (also to be properly designed by formal experts) could gain insight on the time-range a simulation offers: greater state spaces generally contain more possible trajectories a user can choose and the greater the number of trajectories the greater the time a user can spend.

2.1.4.4. *Things left aside*

Beside the intervention and system features that characterize simulations as media, simulations have a specific content, belong to a subject etc. Numerous criteria for describing the simulated system in this respect can be found, e.g. is it concrete or abstract, natural or artificial etc. However, those features are not under investigation here. It is assumed that every application context will have its own taxonomy for the respective subject area, probably borrowed from bibliographic databases (see also 1.3.4). Also ignored were the technical specification criteria: Which programming language is used (e.g. C/C++, Java, Delphi), is it a pre-specified format (e.g. Toolbook™, Shockwave™, Flash™), which platforms (Windows™, MacOS™, Unix-derivatives, cross-platform etc.) are possible etc.? Furthermore, the simulation may have a specific role, e.g. scientific, educational, economical

[15] The art of designing a simulation as a convenient medium is to make the necessary decision processes easy, to design easy intervention ports. It should be noted that – contrary to the widespread expectation of 'good' simulations being massively interactive – interaction might be heavily restricted in convenient simulations. In this sense, a simulation that shall be powerful is at risk of being inconvenient. On the other hand, a simulation being user friendly is endangered of being trivial. Thus, simulation design is always a power-convenience trade-off.

etc. For characterizing simulations as educational media, for example, it has to be specified which type of use (e.g. demonstration, individual, grouped) is possible, which prerequisites (prior knowledge, qualifications etc.) are given, what the context is (single unit, course, exam) etc. Here the LOM or projects like the Educational Modelling Language EML (Koper 2003) come into play since they provide metadata designed for this purpose. A comprehensive characterization of simulations as media somehow has to incorporate all types of criteria. Of course, it should be ensured that the resulting set of criteria is small enough to be manageable.

2.1.5. Summary and conclusion

The analysis given above can be summarized as follows. (1) Simulations have specific features: they represent inherent functionality of a (complex and/or dynamic) system that is to be operated by specific interventions. (2) The core of simulation is cognitive: simulations contain a human factor (i.e. decisions preceding interventions) that cannot be directly included in simulation standards, but can be approached indirectly by the design of intervention ports. (3) Standardization ventures should therefore refer to a (still not mature) cognitive theory of simulation (and corresponding experimental paradigms) that should protect them from ending halfway because having missed the human point. In sum, the venture of standardizing simulations is still in its infancy. The period of every expert clearing one's own path to simulation is not yet overcome. With respect to conceptual difficulty, semantic variety and the specialization in complicated content, the must of expertise is no surprise: the versatility of simulation takes a heavy toll on standardization! However, continuing effort will make this exclusive medium – at the moment primarily preserved to experts – finally fully accessible to the public. The venture of standardizing simulations goes uphill all the way – but it goes.

Reading Advice

The Simulation-Special provided an introduction into the various notions of simulation by asking whether simulation can be conceived as a medium like a film or a text is conceived as a medium. It was shown that simulations – contrary to films or texts – crucially depend on interventions in their inherent functionality. Simulations can therefore be successfully conceived as media if (and only if) the human factor is clarified and respected. But this also implies that focusing on media will bring us only halfway in understanding simulation. An approach to simulation that can reconcile the variety of notions has – as it was already suggested in the Simulation-Special – to focus on the human factor. In the next section, this human factor will be analyzed in the context of scientific work. Where else than in science can we find a more sophisticated and well-planned exertion of simulation as an explanatory tool!? Thus, scientific work will serve as a good practice scenario of simulation and therewith as a basis for the analysis of behavioral and mental processes accompanying simulation. This analysis aims at the extraction of a unifying account of simulation – a generic simulation scheme.

2.2. Simulation in Science

Simulation performed on computers has matured to an approved method in science. But its benefits of gaining explanatory power and new insights often are achieved at the cost of introducing new problems: *How abstract might simulation be? Is natural plausibility of the simulation mandatory? How to evaluate the explanatory value of the simulation? Who actually needs simulation?* Some of these problems are due to novelty effects that will sort themselves out in time. But the confusion also unveils serious inconsistencies in the underlying conceptions. The following section shall address some of the conceptual issues by providing a framework in which simulation can be understood as a genuine part of scientific work. First, a general characterization of a scientific workflow with the aspects 'theory', 'design', 'experiment' and 'evaluation' is provided. Particularly, it is discussed how the specific aspects of modeling and simulation relate to this workflow. It is argued that simulation – the external and the internal aspects – are best understood in a framework that focuses on mental simulation of the system under scrutiny. This framework allows comprehending why 'artificial preparations' such as simulations help to scrutinize natural phenomena.

2.2.1. Scientific work

The specific role that simulation plays in scientific work can be understood theoretically and practically. I chose a practical account. Therefore, I will not provide an exhaustive review of philosophy of science (but see e.g. Carnap 1995 for a concise textbook account and Bechtel *et al.* 2001 for a specific account of the brain sciences). Neither will I introduce any particular account for the aspect of modeling (but see Magnani *et al.* 1999 for a general introduction and see Webb 2001 for brain sciences). I presume that simulation in science is, in most of the cases, model simulation. Instead, I will provide a simple, practical account of scientific work that includes the aspects of modeling and simulation and extend this account to a more generalized (cognitive) framework. This framework shall help to 'dissect' scientific work and unscrew the specific role simulation plays.

2.2.1.1. *A scientific workflow.*

Consider the following four aspects of scientific work:
(1) Theory: system specification, problem analysis, hypothesis, prediction
(2) Design: methods, experimental protocols, problem-operationalization
(3) Experiment: concrete setup, preparation, data acquisition
(4) Evaluation: data analysis, statistics, conclusion, discussion

Obviously, these four aspects are interwoven and not clearly demarcated from each other. Even more important, they are additionally circular in that the experimental design (last aspect) leads consequently to the experiment (first aspect): the conclusion refers to hypothesis and implies suggestions for what to do next. In order to illustrate the different aspects they will be discussed in an example (the four aspects of scientific work are indicated by the numbers in parenthesis): A neuroscientist studies the visual system of flies. Assume that the fly is integrated in an experimental setup in which the electrical activity of single neurons in the fly's brain can be observed on an oscilloscope and registered on a computer while arbitrary visual stimuli can be presented. The neuroscientist – for the sake of legibility let us assume that the neuroscientist is female – can individually identify neurons in each fly preparation by the response to a motion stimulus probe. Thus, she knows that the neuron responds to moving stimuli. She hypothesizes that the neuron is involved in a figure detection task (1). For this to be true, the neuron must specifically react to a specific stimulus condition, e.g. a dark spot in front of a light background (2). The experiment (3) yields data showing that the activity change during presentation of the stimulus was above chance (4). Now, she knows that the neuron reacts to a moving spot, but she does not know if this stimulus is specific, i.e. if it can be called 'a figure' (1)? The cell could react to movements of large spots as well (2). So she decides to test different sizes of the spot (3). She chooses, say, ten stimulus sizes and tests each stimulus, say, 50 times (randomized) in, say, 20 individual fly preparations. She will digitally record the responses and build mean values of each response amplitude of the different stimulus sizes. As a summary of her experiment, she can plot the ten mean responses (and variance) against the corresponding stimulus size (4) (see fig. 19). The aspects of modeling and simulation will be considered extra in the following paragraphs for assessing the role simulation plays in this workflow.

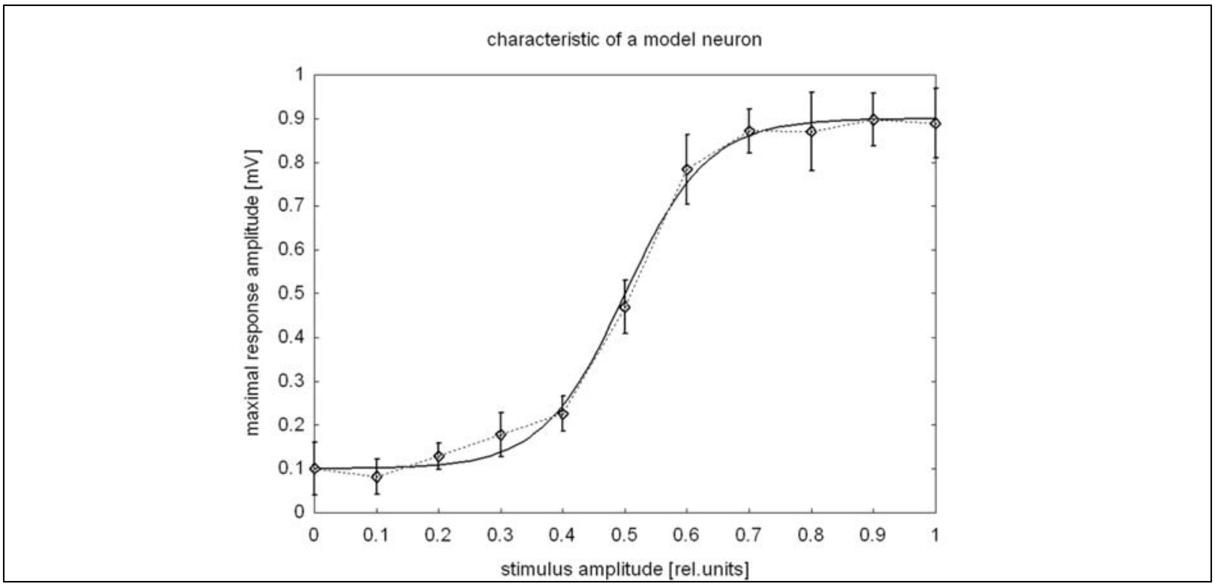


*fig. 19: **Characteristic of a model neuron.** Maximal response amplitude (ordinate) increases with stimulus amplitude (abscissa). Linespoints (dotted) show hypothetical measured values with standard variation. Line shows approximation of the relation between stimulus and response with logistic function.*

2.2.1.2. *Modeling and simulation*

Further consider our female neuroscientist studying figure detection in flies. By connecting the ten points graphically with a line, she can construct a preliminary stage of a simple model of the neuron (more exactly: of it's response properties), i.e. its *characteristic* for different stimulus sizes under the specific experimental conditions that she used. It should be noted that she has not measured the values on the line between the data points; but by graphically extrapolating these points, she *assumes* the relation between stimulus size and response to be continuous. (It might be the case, however improbable, that somewhere in between two measured stimulus sizes the cell does not respond at all or completely different from her assumptions.) The assumed relation in the characteristic can be considered a model of the neuron's response properties. She can predict a response for a given stimulus condition from the characteristic by reading the response for the corresponding stimulus size from the plot. She does not have to measure again, i.e. she does not have to go through the workflow again! (It should be noted that this is already a kind of mental simulation of the experimental situation.) Alternative to the graphical extrapolation, she can *fit* a curve to the data points. This mathematical method has the advantage of providing not only a graphical but also a formal model of the neuron's response properties. For example, she could take the logistic function $y=1/(1+e^{-ax})$

representing a sigmoid characteristic. In a simple case, she has to adjust the parameter *a* for receiving a fit accounting for the data points. Then, the response amplitude *y* can be predicted simply by computing the model for a given *x*.

Of course, models are frequently much more complex and comprise numerous series of experiments in which theory, design, experiment and evaluation are nested in a sometimes complicated manner. For example, our neuroscientist has to exclude errors and corroborate the specificity of the neuron's response properties by testing for further stimulus properties that could also cause the same responses of the neuron, e.g. contrast, color, patterns, motion etc. As a further step, by analyzing other neurons that feed into or are fed by the putative figure detection neuron she might by able to assign or even construct a model of the whole figure detection mechanism. For clarifying, whether the neuron plays a role in that specific figure detection mechanism, she implements it as a computer program and simulates the situation she has chosen in her experiments. By using comparable stimuli in both experiment and simulation, she can observe the response properties of the model neuron and compare it to her experimental results. Consistency between the response properties of the natural preparation and the model confirms her hypothesis, while differences point at problems. Now, she can use the model to simulate other stimulus conditions and predict how the neuron will respond. Consistency between simulations and experiment, again, corroborates her hypothesized mechanism. In this way, natural and artificial preparations supplement each other.

It should become clear from the aforementioned that modeling is embedded in the scientific workflow. It relates to evaluation (4) in that it results from data analysis, but also contains elements of theory (1) in that assumptions on the system under scrutiny are made. It relates to the design (2) and experiment (3) as well, but only theoretically: by accepting a model, it is assumed unnecessary to perform the experiment for each value because it will supposedly bring the result predicted by the model. In a sense, the experiment is carried out virtually in the cognitive domain. The model builds a short cut between theoretical analysis and data analysis.

Where is simulation in this conception? If the scientific workflow was considered as a circuit with the aspects theory, design, experiment and evaluation and if the model is considered as that component, which realizes the short cut between theory and data, then simulation can be considered as the entirety of processes going on in the short circuit. Thus, simulation is more than the model. It implies:

i. rules for applying the model that result from the stage of theoretical analysis, i.e. hypothesis or problem definition,
ii. a prediction in the form of choosing initial parameters for the model (experimental design),
iii. operating the model (experiment),
iv. an evaluation of the output (or at least a preliminary stage of evaluation).

In the case of the neuroscientist simulating the model of the figure detection mechanism she has to define the problem (i): "Does the model mechanism respond specifically to figures or also to other stimuli?" Further, she has to parameterize the model (ii), and run the model on the computer (iii). Finally she has evaluate the results, i.e. determine if the response to figures is specific (iv). Seen from the perspective of the scientist, it does not make a difference if she interacts with the model or with the natural preparation (i.e. the fly) – in both cases she performs scientific work as it was defined above.

In the case of the neuroscientist applying the simple graphical characteristic of the neuron just by looking at it and analyzing it (without a computer), speaking of simulation appears to be less adequate at first glance. But she also has to define the problem (i), e.g. determine responses for stimuli she has not measured. As a form of theoretical analysis, the neuroscientist (implicitly) instantiates participating elements and activities (setting up the model), i.e. that there is a neuron, that it receives an input value, that it generates an output value (and, usually, that it serves a specific function, e.g. figure detection). She has to specify the parameters (ii), e.g. by looking at the abscissa for a given stimulus size. She presumes that the model will behave in certain manner if she imposes certain conditions on the model, i.e. an input value of *x* will result in *y*. Thus, choosing a certain case (selecting a value of a variable, defining a parameter set) reflects the principle of experimental design. She has to operate the model (iii), e.g. look at the ordinate for obtaining the response value. She has to generate the appropriate eye movements to produce the sensations necessary for concluding the neuron's response for a given input amplitude. Finally, she has to evaluate it, e.g. accept it as plausible and assess the implications for the question posed. So, again, all aspects of scientific work are implied in this case. If simulation is seen as the short-circuit of scientific work introduced above, applying the graphical characteristic is also a kind of simulation.

Another case, even more 'decoupled' from practical scientific work, is given when the neuroscientist simulates the characteristic (e.g. the logistic function) exclusively mental. Visualization (imagery) can cast the inner eye on the sigmoid form of the characteristic and help the experienced neuroscientist to find the correct response amplitude for a given stimulus size. She might not be able to determine the exact value, but she will be able to give an approximated answer in terms of ranges in which the response might occur. For example, she might chop the characteristic into the three parts: 'pre-threshold', 'dynamic range' and 'saturation'. She can conclude from this partitioning that increasing the stimulus size at large values, say, relative values between 0.9 and 1, will result in only minimal changes of the response amplitude. Moreover, if she knows the steepness of the quasi-linear dynamic range she can approximate it by a linear function $y=a\cdot x$. Without any external media or devices, she can even make a good guess for the exact value of a response amplitude y for a given stimulus size x. This case illustrates that scientific work can be performed also independent of external events, i.e. a fly, an experimental setup, digital data etc. All conditions necessary for scientific work are already 'in the scientist'.

The cases provided above show that simulation comprises the four aspects of scientific work introduced above. Thus, simulation can be conceived as virtual experimental work, hypothesis testing or *scientific reasoning* (see Klahr 2000 for a similar account). On first glance, this notion of scientific work appears very general. For example, data analysis is usually conceived as acting on externalized data (e.g. files on paper or in a computer) and not as a brain process of the scientist that accompany eye movements directed to data. Moreover, simulations are usually known as computer simulations, not as mental states and affairs. But it is also evident that any data analysis and any computer simulation is based on cognitive processes that can also be performed isolated from external media and devices, such as graphs, oscilloscopes or computers. Nevertheless, it might particularly seem strange to see simulation as an integral part of scientific work (and mental simulation as the core of any simulation). But the analysis provided above (particularly the case of the thought experiments with the characteristic) make clear that an account of mental simulation is needed to understand the issue of simulation in scientific work (e.g. computer simulation). In this way, mental simulation introduces itself as a cognitive mechanism used in scientific work. This view is not new. Particularly brain scientists propose compelling argumentation and evidence in favor of an even more general conception of mental simulation. Jeannerod (1994; 2001), for instance, argues that cognitive processes become possible for humans by suppressing motor activity. The internal representations of motor activity can be handled more freely without actual execution and result in the ability to mentally simulate future actions. Jeannerod presents a wealth of evidence from neural and behavioral sciences in support of this hypothesis. A similar account, but with focus on lingual processes is provided by Hesslow (1994; 2002). A very comprehensive presentation of mental simulation covering also the field of cognitive science is provided by Barsalou (1999).

It can be concluded so far that theory, design, experiment and evaluation describe the aspects of scientific work comprehensively. Modeling can be described as a shortcut between theory and evaluation because it serves as a surrogate for experiments. But a model alone does not serve any scientific purpose as long as it is not operated. Operating the model means to perform experiments with models. Thus, operating the model re-introduces the aspects of design and experiment. A simulation is such an operated model and, consequently implies all aspects of scientific work. Thus, considering simulation poses the question where to draw the line between scientific work and simulation? Scientific work can be conceived as simulation if cognitive processes of the scientist are put to the center of analysis. Mental simulation is pervasive in scientific work. However, mental simulation is not scientific work – it is a cognitive mechanism used in scientific work. A closer look to mental simulation in the next section will help to reveal the differences and consequently help to further clarify the role simulations play in scientific work.

2.2.1.3. *Mental simulation in scientific work*

Mental simulation is a general cognitive process that helps to answer questions, to find explanations. Consider this very simple case: a person (let him be male this time) seeing a moving twig on a tree during a walk asks himself why the twig moves. Assume that, eventually, the man will come up with an answer to that question. Now, mental simulation can be described as the entirety of processes in between question and answer. The moving twig example shall be analyzed in more detail in order to understand how mental simulation helps to find explanations. The following analysis is a (quite raw) conceptual specification of mental simulation. A more detailed account is provided elsewhere (see section 3.1.; Jeannerod 1994; 2001; Barsalou 1999).

It should be noted that all states and affairs described below could happen in parallel and with mutual causal connections. The whole is only partitioned for a clearer presentation. For instance, having perception and posing a question might very well be same process since perception is already an interpretation of a given situation and might only arise if some inconsistency or question is detected[16].

0. Perception
Premise of mental simulation is the perception of arbitrary elements and activities, i.e. perception of the moving twig. Generally, it does not matter if elements and activities are genuinely natural or artificial because both end effectively as a perceptual element or activity. Even an observation of inner phenomena (e.g. an idea) is part of this perceptual domain.

1. Question
Perceptual activity alone makes no observation. But ongoing perceptual activity during the walk can become an observation by posing a question, e.g. *"What caused the twig to move?"* If there were no question there would be no observation, just ongoing perceptual activity. The question triggers the extraction of specific perceptual elements or activities from the ongoing activity. The raw version of the question might be described as: *"Mmmhhh?"* or *"What was that?"* The verbal specification will presumably happen at later

[16] Even though it shall not be considered in depth here what actually causes the question to come up, it might help to think of a cognitive mechanism that detects inconsistencies and tags them for scrutinizing, such as a 'checking for cheaters' module (see Cosmides 1989 but see also Johnson-Laird 1999 for critical remarks).

instants. An elaborated version of the question could be: "*Was it the wind that moved the twig or was it something else?*"

2. Model

Instantiating an appropriate model for answering the question and clearing inconsistencies comprises several parts. (a) Elements and activities must be distinguished from the ongoing perceptual activity, e.g. 'twig', 'motion', 'wind'. Obviously, memory processes can very well come into play at this instant, e.g. names for elements (verbalization), personal experiences with wind etc. Certain attributes can be specified, e.g. size and weight of twig, force and direction of the wind. (b) Causal relations 'stage' the elements and activities (for obtaining a mechanism), e.g. "wind is a force that acts on the twig and causes motion". Elements and activities and the causal relations between them make up the model. (c) Rules must be found for how the model shall be operated, e.g. "transfer a non-moving twig (state#1) to a moving twig (state#2) by applying the force that wind causes to the twig". This 'design' or 'plan' can be very implicit, but should be explicitly notified at this point for not being confused with the rules inside the model ('causal relations') that put the model to stage. Model and operational rules build the mental 'setup' or 'preparation'.

3. Operation

Enacted elements and activities, causal relations (the model) and operational rules yield a running model - a simulation. Due to the serial character of thoughts it seems hardly conceivable that these enacted models are smooth Hollywood-like pictures running through our head. Rather, enacted models might better be conceived as checking single configurations of elements, activities and causal relations, one at a time. (In non-controlled, non-thoughtful simulation configurations might rather be checked in parallel.) For example, the state space of the model is sampled step by step in a controlled fashion when the twig model is envisaged without wind as one configuration 'state#1' and with wind as the next configuration 'state#2'. At this point, again, broadcasting prior knowledge (memory) to the running model seems obvious (e.g. the personal experience with wind).

4. Evaluation

The results of the running model must be analyzed in order to assess their implications for the question, i.e. their explanatory value. This is not necessarily an explicit reflection, in the sense: *"What does it mean?"* The mere transfer of the model from 'state#1' to 'state#2' can be seen as a first step of evaluation: the running of the model as such is the first indicator for explanatory value of the simulation. The transfer from 'state#1' to 'state#2' could also fail, e.g. due to the wrong wind direction. Then, a new inconsistency would be thrown up and the game has to start anew. Thus, successful simulation makes explanations conceivable and the failed simulations render explanations inconceivable. Something being conceivable can be sufficient for being a successful consistency check. Something being inconceivable is to be seen as an inconsistency in the simulation that hinders the model from running ('malfunction'). Eventually, depending on the priority of the question, several outcomes are possible: the twig issue can be settled by instantiating a running model, e.g. *"The wind moved the twig!"* But the running model can also serve as the starting point for further questions, e.g. *"OK, the wind is a possible candidate for moving the twig, but what about an animal?"* Malfunction of the simulation throws up a new inconsistency and can cause a new iteration of the simulation with changed parameters or new elements and activities, e.g. *"Was the twig moved by an animal?"* Finally, the question can be considered not to be worth answered, e.g. *"That shall not be my problem!"* Consistency is the premise for something being a possible explanation, i.e. it indicates high explanatory values. Inconsistency, on the other hand, points at low explanatory values (which is better than none, because one can hook off that configuration of the model). In sum, the phase of evaluation ends with a "decision" for an explanation (an adoption of an answer to the question) and, consequently, implications of how to proceed.

The moving twig example illustrates what mental simulation is about. But aren't there other concepts that also describe what happens? The entirety of processes between question and answer might be also termed, depending on the context, thinking, reasoning, modeling, problem solving or mental simulation. What is the correct term, then? Reasoning has no reference to the external situation (i.e. the moving twig perceived during a walk). Modeling bears a too generic and explicit notion, i.e. developing something for applying it in another context. Similarly, the real problem is missing for calling it problem solving. Thus, mental simulation appears to describe best the mechanism that helps the man taking a walk giving an answer to the initial question: it relates to something external (i.e. the simulated system, the moving twig) and has an ongoing, casual and automatic character. Mental simulation is that something in between question and answer.

Comparing this account of mental simulation with scientific work, i.e. the examples of the moving twig and the neuroscientist, respectively, reveals striking similarities: The peculiarity in the ongoing perceptual activity that poses a problem in mental simulation corresponds to theorizing in scientific work, the staging of elements and activities corresponds to design, operation of the model corresponds to experimental work and the assessment of the outcome corresponds to evaluation. The striking and obvious difference between mental simulations and scientific work is that the former is predominantly internal and the latter is also external. Moreover: mental simulation can be very implicit and automatic, while scientific work is usually explicit and controlled. For example, there is not necessarily an explicit prediction or hypothesis in mental simulation but rather an inconsistency between perceptual activity and internal representation.

In sum, there are strong correspondences and obvious differences between the accounts of scientific work and mental simulation described above (see fig. 17). The comparison indicates that mental simulation can be assumed to be a crucial cognitive mechanism used in scientific work. Consequently, mental simulation shall now be applied for describing the role simulation plays in science, generally. Whether or not mental simulation is a tenable cognitive framework of scientific work should become clear when applied to the different situations of scientific work.

2.2.2. Domains of scientific work

Several situations of scientific work can be distinguished. The ones already considered were: experiments in the laboratory, computer simulations and thought experiments. These situations can be classified according to the domains in which they are performed, i.e. experiments in the natural domain, computer simulations in the artificial domain and thought experiments in the cognitive domain. However, these domains show considerable overlap in the situations of scientific work since, for instance, mental simulation is involved in any case. So the question is: Which aspect in a given situation of scientific work belongs to which domain? How are the domains arranged so that they help to solve capacity problems? In order to demonstrate that almost any aspect of mental simulation (e.g. model, operational rules etc.) can refer to any domain in scientific work, the different situations of the neuroscientist studying figure detection in flies (e.g. characteristic, curve fitting, thought experiments) shall be reconsidered in the following sections and discussed with respect to domains of scientific work.

One broad distinction for domains is natural-artificial. In the sense of the term 'artifact', the artificial domain might be conceived as a man-made construction inside the natural domain. It definitely comprises digital computers and simulation environments with computerized models, and might, depending on the concrete case, additionally be true for experimental environments. The case of experimental environments makes clear that the application of the natural-artificial distinction to organisms is not trivial: while it seems clear to call wildlife animals natural, it seems at least questionable if mice with genetically knocked out behavioral traits are natural. However, for the sake of simplicity, I propose to refer to computer simulations and robots as being artificial, while animals and biological preparations are termed natural. Experimental environments (setups) are considered to refer to natural phenomena as long as the relevant data are extracted from biological preparations.

In the lab situation of our female neuroscientist, the fly fixed in the experimental setup is a biological preparation. It belongs to the natural domain. Operational control of the biological preparation (i.e. the experiment) is realized in the cognitive domain by way of mentally simulating the experimental setup including the biological preparation. In another situation the neuroscientist developed a characteristic. The graphic characteristic of a neuron that is used to predict a response amplitude for a given stimulus size is an artificial preparation. The actual application of the artifact can be diverse: while the novice has to perform the procedure of applying the artifact explicitly (and probably with great effort), the expert simply 'knows' the result. Thus, the artifact can become more implicit by repetition of mental simulation and can eventually become exclusively cognitive: there is a transition between artificial and cognitive. The advanced scientist (in between novice and expert) can develop and apply a 'recipe' to generate a prediction, e.g. *"note given stimulus value, look x-axis, find stimulus value on x-axis, fixate, move straight up to curve, fixate cross-section, go from cross-section to y axis, read y-value, make prediction with* *y-value"*. Similarly, a formula can be applied. A linear function $y = a \cdot x$ is a simple case: *"note given stimulus value, take as x-value, multiply with 'a', note result, take as y-value".* In this situation, not the model but the operational rules for using the model are learnt (and become cognitive). The logistic function, for instance, might pose more problems so that the neuroscientist realizes it as a computer program. Then, the 'recipes' for simulation are implemented in an automatic program that offers predefined intervention ports, e.g. a field in which a value of a stimulus size can be entered for computing the corresponding response amplitude. The model (the formula) as well as operational rules are outsourced to an artifact. Only the aspects of perception and evaluation are exclusively left to the scientist. In the case of the neuroscientist performing curve fitting for finding a characteristic analytically, the role of computer simulation is even more obvious: she specifies the elements and activities that describe the neuron (modeling) and defines operational rules for computing the model in the form of input, output and evaluation functions. She might be able to do this in her head or on paper, too, but it is more economical to externalize parts of these tasks to a computer program, particularly computing model and evaluation functions. These cases illustrate that computer simulations are mental simulations outsourced to the artificial domain.

By acting in the artificial domain, it is possible to work scientifically independent of a natural preparation. For example, if the scientist changes from laboratory to computer, the 'medium' in which scientific work is performed changes from a spatiotemporally larger and structurally more constrained environment to a more compact environment. The artificial preparation (e.g. the computer simulation) offers a variable amount of degrees of freedom. In order to get models running in the artificial domain, elements and activities affecting the natural preparation have to be translated into artificial elements and activities that obey formal rules[17].

The last situation of scientific work that will be considered here are thought experiments. Seeing scientific work as 'thought experiments' means to move the final step towards the cognitive domain: here, the experiments are

[17] It should be noted that, in this conception of scientific workflow, it does not matter how similar simulated elements and activities are with respect to the natural situation (as one might feel tempted to think) – the only thing that matters to understanding is how similar elements and activities in the artificial domain are with respect to those in the cognitive domain. This elucidates that cognition drives the understanding of the natural domain. However, cognition itself is also driven externally, for instance, by the causal rules it is confronted with when experiments are carried out and perceived.

carried out exclusively in the cognitive domain. Genuine thought experiments do neither necessarily refer to natural environments nor to artificial simulation environments and, therefore, are less subjected to natural or formal constraints. Hence, even more degrees of freedom than in the artificial domain are possible. The boundlessness is also illustrated by the “Martian” touch of typical thought experiments frequently found in philosophy (e.g. Putnam 1975 for the famous 'twin earth' example).

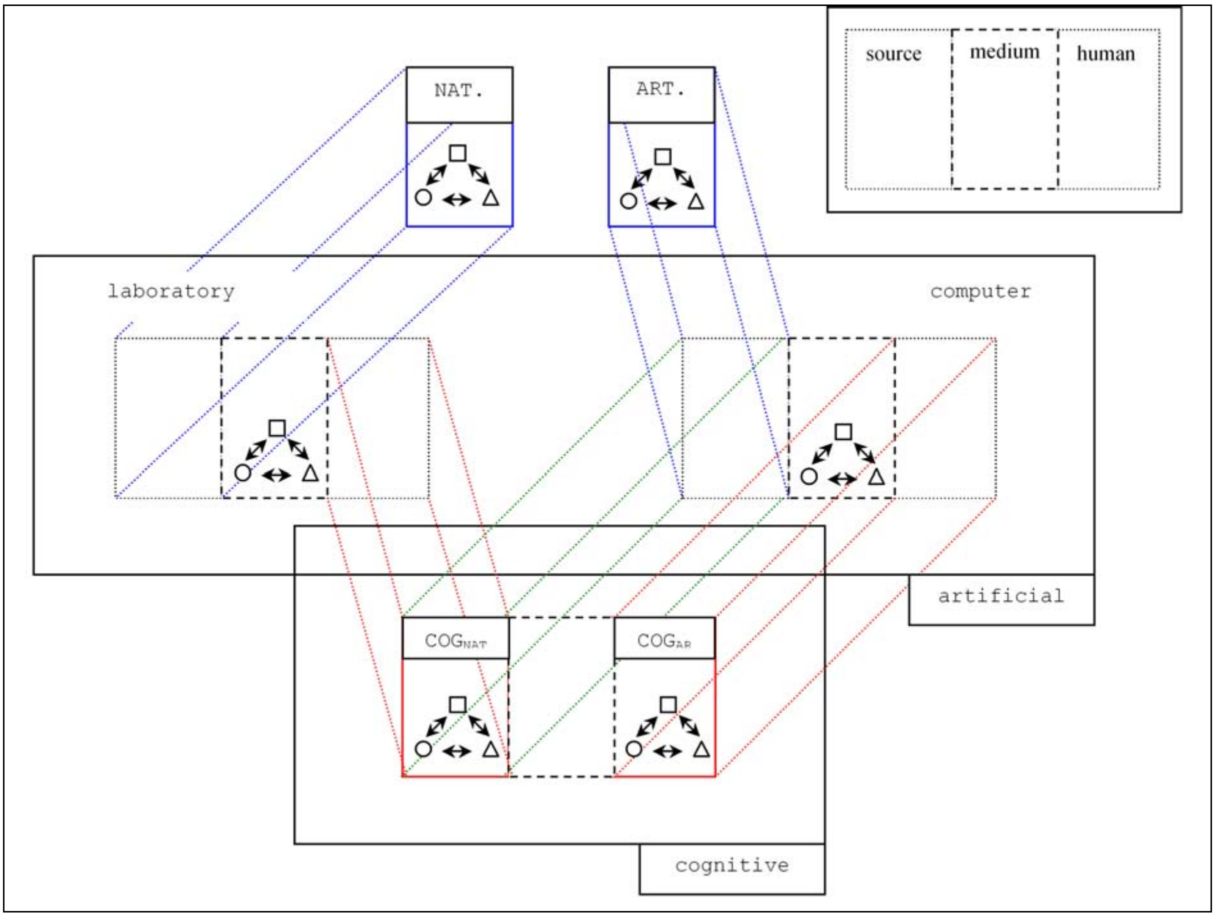


*fig. 20: **Domains of scientific work.** The natural domain comprises all other domains. Natural preparations (in experimental environments) are used in scientific work to ‘feed’ the cognitive domain (mental simulations) with data of external world. Mental simulations in the cognitive domain, in turn, operate the natural preparation. Similarly, artifacts, e.g. computer simulations are used to feed mental simulations of the artifact. The difference between both systems, seen from the mental simulation stance, is that natural preparations provide inherent functionality (a ‘momentum’) by themselves, while artificial preparations have to be created. The inset in the upper right corner indicates that the modeling relation refers to three different instances: source (the preparations), medium, i.e. the representational domain (e.g. laboratory and computer) and human (mental simulations).*

In sum, three different domains of scientific work can be distinguished: natural, artificial and cognitive (see also fig. 20). If mental simulation is assumed as the underlying mechanism, a clear view on the relation between the domains can be provided: an external preparation – be it an experimental device (e.g. natural preparation or animal model), a computer simulation or a verbal description – is an extension of mental simulations that can aid handling the mental simulation. Put another way, the external preparation can guide or *control* mental simulation. The natural, artificial and cognitive domains relate to each other in a specific manner. The natural domain contains both the artificial and the cognitive domain. Inside the cognitive domain are representations of the natural and artificial domains (call it N' and A'). These representations allow mental simulation of systems that are inside these domains. Mentally simulated systems can be investigated by way of "scientific work ". Scientific work can be targeted at a natural preparation (e.g. lab experiment), to artificial preparations (e.g. computer simulation) or to cognitive preparations (e.g. thought experiments). Natural preparations can be transformed to artifacts by externalization of cognitive preparations ('modeling').[18]

The account of scientific work introduced so far might cause the quite unusual impression that anything in scientific work is simulation (Jeannerod 1994; 1999; Hesslow 1994; 2002; see, however, Cruse 2003 for similar accounts). But seen from the scientist's point of view this is true: since the principle of mentally simulating external states and affairs is propagated in the experiment, the experiment can itself be regarded as a simulation of the natural states and affairs ('natural simulation'). The experiment and the natural preparation were built after the construction plan of the mental simulation. This might sound as if there was no nature at all in scientific work: *"Everything is simulation? Why bother about the natural preparation? Why don't we do it all in our head?"* This notion is definitely not meant, here! As will be shown in the next sections, there are very good reasons to focus on natural situations – but also that it is difficult to control them.

[18] As the attentive reader might have noticed, the domains of scientific work, i.e. natural, artificial and cognitive correspond nearly exactly to the domains of the brain sciences, i.e. the neural, computational and cognitive domain, respectively (see also 1.1.3). This is not surprising since the activities of scientists form the domains of the brain sciences. It could be stated that they refer to the same ontological category.

2.2.2.1. *From natural situations to natural preparations*

Presumably, no one seriously doubts that a neuroscientist should aim at the most natural possible situation for studying the brain. In the example of the neuroscientist (remember, she was defined to be female) investigating the figure detection mechanism in the visual system of the fly, the most natural situation the scientist can seek, is probably a field experiment, in which a wild living fly (coming from a wild population) performs spontaneous figure detection (e.g. chasing mates) during free flight. If she is lucky, she can derive approaches for explaining the free flight performance with her observational data from the field. However, she is unlikely to make substantial progress concerning the neural mechanisms underlying the observed behavior. For analyzing neural mechanisms, she has to measure the neural activity. Optimally, she would like to implant an electrode with radio transmitter into a figure detection neuron of a wild fly and measure in a wild situation. But various factors can (and will) render this plan futile. First, realization will be difficult. For example, the technical equipment needed for transmitting the neural activity is many times heavier than the fly! Second, interference with other neural factors will be a problem. Even if she succeeded in technical realization and registers neural signals during chasing, the signals she registers can be superimposed with many other signals that arise from the various other tasks the fly has to perform in a natural situation. For ruling out the influence of other sources, the stimulus must be reduced. Moreover, for ruling out the influence of chance (noise, spontaneous activity) the same sequence must be repeated several times. These changes would already restrict the natural character of the situation.

Most of the factors that are decisive for success or failure of such an experimental venture are ultimately economical. They have to be assessed by cost and benefit. Consider the following hypothetical premises: the technical equipment needed for the experiment is realizable in a development process taking 10 years and the sequences she needs to record might be reachable if she measured another 10 years with a team of 100 scientists. Usually, she will conclude from these premises that a more promising way to study figure detection in the fly is in the laboratory. So, she designs a natural preparation: she fixates the fly in the middle of an experimental setup and defines what the fly shall see and how long, how often etc. it is presented. In other words, she simulates the natural situation in order to gain control over it. The cost-benefit rationale underlying these decision to go to the lab is: she presumes that the time and effort it takes to get the free flight data is better invested in laboratory work since she can gather so much data in the laboratory that she will outperform a colleague attempting the free flight project by lengths. This way to a successful observation is based on reduction of complexity and dynamics of the natural situation. The example illustrates that the control of dynamics and complexity is important for scientific work. Two aspects of gaining control have to be distinguished, one that is external of the scientist and one that is internal. Externally, dynamics and complexity of the natural is reduced. For example, the fly cannot fly where it would spontaneously fly, but is fixated. It does not see, what it would see during free flight, but is presented artificial stimuli. Internally, the dynamics and complexity of the mental simulation the scientist has to perform is reduced with respect to various aspects. For example, reducing the natural dynamics and complexity means to get rid of unmeant alternative explanations. Interfering factors are eliminated.

But is it really desirable to reduce internal dynamics and complexity? Why not consider the whole problem? What if money is no object and the scientist could do both: the field experiments and the lab experiments? For understanding the necessity of reducing complexity more thoroughly, imagine an extraterrestrial intelligence that has the same basic prior knowledge on brains that our scientist has, but that has superior perceptual and cognitive abilities. Observing a single (natural) chasing sequence of the fly might be enough for determining the neural mechanism because its superior cognitive apparatus enables it to mentally simulate all possible models and compare it with the observations. On the basis of a characteristic difference between a trajectory observed in the natural free flight situation and a trajectory generated in a mental simulation the extraterrestrial intelligence can conclude that only one specific neural mechanism is possible. This becomes possible because it can mentally simulate all possible mechanisms and predict the trajectories and compare these simulated trajectories to the observed trajectory. In contrast, the scientist does not have this cognitive ability and has to reduce complexity by analyzing neural activity in experiments. Thus, one reason (and an economical constraint as well) that necessitates reduction of internal complexity is the limited capacity of the human ‘simulation engine’ (see e.g. Broadbent 1958; 1975) Seen from the mental simulation stance, the goal of reducing internal complexity is a reduction of cognitive load.

But which aspects of scientific work can be reduced? Comparing field situation to lab situation reveals that the spatiotemporal coherence of the participating elements and activities is increased, e.g. the fly cannot fly anywhere, the light source is not the sun etc. Interactions with a given experimental environment have to be planned and controlled in the cognitive domain by way of representations concerning elements and activities in the respective environments. Thus, the operational rules and, therewith, the sensorimotor coordination and planning necessary to control the situation is reduced. Reducing the environment means to reduce the cognitive load while the model is operated. The model is 'cleaned' from sensory-motor representations of environmental constraints. Assuming a limited capacity process underlying the cognitive handling of the model, reduction of environmental complexity can explain why the scientist is able to increase the amount of manageable complexity. Choosing the lab situation reduces cognitive load in almost any aspect of scientific work – alone by reduction of representations.

Ultimately, a maximum of cognitive capacity is achieved by omitting the external world altogether, i.e. in the case of pure mental simulation (thought experiments). Here, the model can be enacted almost undisturbed from external constraints. Given that cognitive capacity is maximized in thought experiments, why, then, not do it all in our head? Supposedly, as long as it goes it is actually done in our head because it is the most economical way to solve problems. But mental simulation is not secure for two reasons:

1.Thought experiments can be inappropriate if the question necessitates the natural preparation to show its inherent functionality. For example, it would be a categorical error to try to 'imagine' a neuron's response in order to assess it's intrinsic noise amplitude.

2. Thought experiments can fail because complexity and dynamics become too strong: When one factor in the model influences multiple other factors and, hence, multiple parameter configurations are possible for processing a model, no clear result might be obtained. This can be termed the 'problem of complexity'. Similarly, when the final state of the simulation is difficult to

predict because each state determines the next intermediate state. This can be termed the 'problem of dynamics'. Note that dynamics can be seen as resulting from a great number of activities, while complexity can be seen as resulting from a great number of elements. Both 'masses' are causally interconnected (see also 1.1.6).

It is assumed here that thought experiments can fail because processing the whole thing at a time exceeds cognitive capacity and does not reveal the rules necessary to understand elements and activities of the phenomenon under scrutiny. Cognition is too slow and too limited in capacity and temporal resolution. Without partitioning, slicing and chopping the simulation and thereby reducing cognitive load, the simulation engine obstructs. Here, not only the model itself, but also its operation is a key problem of processing dynamics and complexity. The model can be "well formed", while it is not clear which is the initial, the final and the intermediate state and what are functions that determine the transitions from one state to the other. Serialization fails. Since the model cannot be operated, no validation can take place: the model is not comprehended.

In conclusion, observing natural situations can be regarded as the optimal approach to explain natural phenomena. But natural situations are often too complex or dynamic to be analyzed efficiently. Therefore natural situations are reduced to experimental situations that contain natural preparations. Economical factors might hinder a scientist to investigate natural situations. But economical factors are not a principal impediment – rather the limited cognitive capacity of the scientist makes reduction of the natural situation to a natural preparation in an experimental situation inevitable. The next section will concretely show solutions to capacity problems.

2.2.2.2. *From natural to artificial*

In scientific work, solutions to capacity problems can be found in all domains considered here: the natural, the cognitive and the artificial domain. The lab situation can offer a solution to these problems caused by dynamics and complexity by reducing the cognitive load in the mental simulation. As explained above, one approach is the reduction of the natural situation by developing a natural preparation. Another approach that will be analyzed in more detail in the following section is to develop experimental protocols that include operational rules for the natural preparation. In an experiment, the natural preparation is operated in a predefined way and the cognitive system is fed with only that amount of perceptual activity that is needed to drive the mental simulation. In the example of the neuroscientist, these operational rules have the form of specific stimulus conditions and specific metrologies that extract specific activities of the natural preparation. For example, electric activity is measured while, say, motor activity is not measured or even suppressed. Each running model, i.e. each state transition of the model that does not reveal inconsistencies is a step towards validation of the model and its comprehension. (Whether the model is 'really' understood, is never certain, of course. It is just demonstrated that the model can be operated – it 'functions'.) Within the various possibilities to conceive the natural preparation, experimental protocols show one specific trajectory through the state space of the natural preparation that guides attention in a predefined way. Experimental protocols are reproducible, i.e. operable on other cognitive systems (e.g. colleagues) another time, another place. This is why experiments are well suited for communicating models between scientists.

The example of the scientist communicating experimental procedures to the scientific community also points at the role of artifacts in solving capacity problems. Any externalized version of the operational rules, be it a verbal description, a journal paper or any other medial representation, is to be seen as an artifact. Scientific work is full of artifacts: the experimental equipment can be seen as an artifact that implies operational rules, especially if it is computer controlled. The preparation itself can be an artifact, e.g. when the neuron is modeled on the computer. This artificial preparation can be used to control the experiment with the natural preparation. Consider that our female neuroscientist presents motion stimuli to the fly, e.g. an object 'passing by'. She measures activity of neurons in the figure detection circuit while the fly is fixed in the experimental setup and cannot move (because the neural activity would otherwise not be recordable). The neural activity that is elicited by the object passing by would be used in the natural situation to trigger a motor response, e.g. a chasing maneuver. But the fly cannot generate a motor response because it is fixed. The object passes by without any approach of the fly towards the object. (Since the natural sensorimotor loop is 'cut' this condition is called 'open loop'.) Now, consider that the neuroscientist has developed a computer program that simulates the motor system. She can feed the neural responses into the computer program and simulate that motor response, which would be generated by the fly on the basis of the neural activity as a response to the object passing by. Finally, she can feed that motor response to the program that produces the stimulus of the object passing by and generate the stimulus that the fly would have seen if it had moved. The resulting situation is as if the fly would have moved. (This condition is called 'closed loop'.) In this example, natural preparation and artificial preparation are interwoven. This is not an unusual method (see e.g. Kern *et al.* 2001).

In the final step towards the artificial, a natural momentum is reproduced with artifacts that substitute the natural preparation. In the example of the neuroscientist, this would be a computer program that also simulates neural activity in the figure detection circuit. Thus, irrespective of the concrete realization with a machine, a robot or a computer program, a simulation is characterized by an artificial momentum. Note that the artificial momentum can only be achieved by operating the model, not by the model alone. A machine, a robot or a computer, which is not operated will not show any momentum. (Neither will a graph or a journal paper show a momentum as long as it is not mentally simulated.) As a consequence, the model will not help the scientist solving capacity problems. The model alone does not help – only together with the operational rules a helpful situation is generated. Then, the operation of the model is successfully outsourced and the scientist can concentrate on observing the results of the operation. The capacity problem is solved in the artificial domain.

For the sake of completeness, it should be noted that experiments with natural preparations not only help to solve the capacity problems of the scientist in the lab situation by guiding the experimental protocols – the natural preparation also does some of the operations itself that must be performed by the cognitive system of the scientist in the case of thought experiments. For example, a neuron shows signals (momentum) without any effort of the scientist. In a thought experiment, these signals must be mentally simulated by the scientist. When utilizing the natural momentum of the neuron, just the results of the operation are perceived. Generally, the external preparation, be it natural or artificial performs the operation of a part of dynamics and complexity. (Exactly this is the momentum!) Thus, natural preparations and artificial preparations serve the function of an external operator.

In sum, natural preparations and artificial preparations are used to solve capacity problems of the scientist in that they reduce cognitive load of the mental simulation (e.g. perceptual input) to the amount that is needed to drive the mental simulation in way that produces intelligible results. Natural preparations and artificial preparations can be used solitarily or mixed. The difference between them lies in the origin of the momentum they have. If it is inherent in the preparation without adding human design, it is natural. If humans have designed the momentum, it is artificial. This account of scientific work implies that the natural momentum is the only thing that can *not* be termed "simulation". Also the preparation of the natural, i.e. the experiment as such is to be seen as an artifact that serves the function of simulating nature. Concerning epistemological power, however, both are equal. There is no principal superiority of the natural preparation over the artificial preparation. Both can only be evaluated in terms of their explanatory value concerning a specific question (see also 1.1.12). If the artificial preparation gives better answers to a question concerning nature than the natural preparation, it is superior for that question. Thus, the notion of doing science with an obligation of a natural preparation appears ignorant and not constructive[19]. On the other hand, following this conception of scientific work is the point where a functionalist / computationalist commitment is implicitly made. Even if one might have in mind to use computers exclusively as explanation tools, it has to be assumed that a telling explanation is not principally rendered impossible once the natural domain is left. This is a radical change in approaching the subject, from specific natural science to general nomothetical science.

2.2.2.3. *Generating the artificial*

The principal difference between a 'genuine' experiment and a computer simulation is that the former has a natural momentum and the latter has an artificial momentum. The situation becomes a bit more complicated, though, when it is considered that the artificial is derived from the natural. Where

[19] There is no computer simulation about nature. There is only a computer simulation about the scientist observing nature. Computer simulations do not directly refer to nature – they refer to the cognitive processes that the scientist has about nature. There is no direct representational relation between nature and artificial simulation. There is just the representational relation between natural and cognitive, between artificial and cognitive and between cognitive and cognitive.

from does the artificial come? It has to be assumed that mental simulation is a generative mechanism for the artificial momentum.[20]

When the scientist changes the investigatory focus from a natural preparation in the lab to the development of an artificial preparation on a computer, a typical case that implies this generative mechanism is provided. On the one hand, it seems to be the same as in the nature/lab-transition (see also 2.2.2.1): the mental simulation is streamlined since more elements and activities are sorted out and less sensorimotor representations are needed to operate the model. Just think of a skilled scientist sitting in front of a computer just moving hands for keyboard and mouse control and moving eyes for feeding back sensory inputs into the mental model. The scientist in front of the computer sets himself in a nearly direct feedback loop with the model. This close connection between the system under scrutiny and mental simulation is usually much more difficult to achieve with a natural preparation in the lab or field. But beside these quantitative differences there is also a categorical difference between the nature/lab-transition and the transition from nature/lab to the computer: a complete redesign of the elements and its possible interactions is necessary in order to omit the natural by substituting it with something artificial. All the elements and activities that constitute the natural have to be converted to objects and rules in the artificial. Omitting natural constraints implies a definition of artificial constraints that are, in principle, arbitrary. (Although there are means to measure the aptitude of the model, mathematical soundness, for example.) If the scientist aims at exploring the model heuristically, the complexity of the model can be even increased in the artificial preparation. Handling this increased complexity becomes possible because some of the scientist's cognitive capacity is relieved by externalization of operational rules (see also 2.2.1.3). However arbitrary the definition of the model might be, if the model is operable, an artificial momentum has been generated by the cognitive system. Thus, the artificial is generated by a mental simulation of the natural.

2.2.3. A generic simulation scheme

Concluding from the previous sections, the following picture reveals: the

[20] A “first order” simulation directly refers to the natural preparation, i.e. it refers to something that did not receive the treatment of design by cognition. An experiment can also refer to an artificial preparation (e.g. testing a robot). If artifacts are the subjects of the study, an “nth order” simulation is given.

scientist performs natural simulation in the case of experimental work with a natural preparation (e.g. in the lab) and artificial simulation in the case of working with artificial preparations (e.g. with a computer). Mental simulation is performed in parallel in both of the former cases and isolated in the case of thought experiments. These three forms of simulation build the scientific simulation framework. The generic scheme underlying all forms of simulation is that operational rules are applied to a model so that a 'momentum' is generated, which means that the model changes autonomously from an initial state to a goal state. Since this generic simulation scheme is a key to understand the role of simulation in science a more detailed account will be provided in the next section by specifying the modeling relation the simulation scheme is based on. Since 'modeling' is the main topic in the following paragraphs it should be kept in the back of the mind that a model, in the generic simulation scheme proposed here, is rather understood as a post-hoc defined structure of an ongoing simulation than as a necessary or sufficient condition for a simulation to happen. Simulation contains more 'ingredients', namely operational rules and some kind of activation, These are not considered in the following paragraphs for obtaining a clearer view on the underlying modeling relations[21].

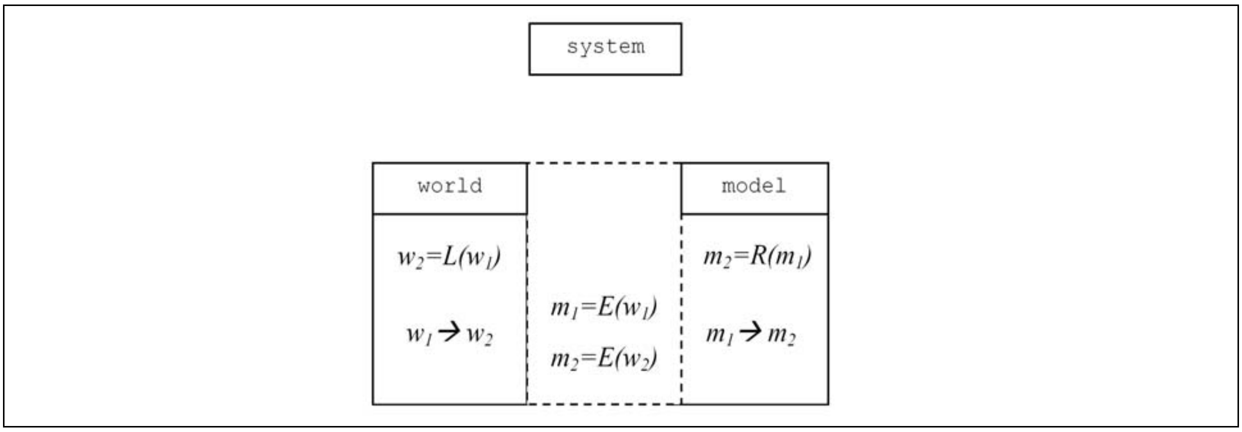


fig. 21: The modeling relation. *A state transition in the world from w1 to w2 accords to a rule or law L. The representation function E 'transfers' these states to the model domain in which the state transition from m1 to m2 is brought about by a model-rule R.*

[21] Scientific work is merely a specific case of the general principle of simulation that was proposed in 2.1 with its threefold architecture: there is (i) a source or simulandum, i.e. the situation the phenomenon originates from, (ii) a representation or simulans, i.e. the natural or the artificial preparation and (iii) a representing or implementing system, i.e. the scientist. If no explicit representation of the source is necessary, because it is *as it is* (e.g. in the case of observing natural situations), the situation during observation is the implicit representation in which the source manifests. The momentum in scientific work corresponds to inherent functionality in the general account.

"The most immediate kind of a model is a metasystem, which implements a homomorphic relation between states of two subsystems, a modeled system and a modeling system. Formally, a model is a system

$S = <W, M, E>$ with:

A modeled system or world
$W = <W, L>$ with states $W = \{w_i\}$
and actions or laws $L: W -> W$.
For example, W could be the set of key presses of a computer operator or the physical world, while L is the behavior of the operator or natural law;

A modeling system
$M = <M, R>$ with internal model states, or representations $M = \{m_j\}$
and a set of rules, or a modeling function $R: M -> M$.
For example, M could be a set of symbol strings or neural signals, while the rules R are the activity of a computer or a brain;

And finally a representation function
$E: W -> M$. (Remark: the original text used R instead of M!)
For example, E could be a measurement, a perception, or an observation.

When the functions L, R, and E commute, then we have:

$m2 = R(m1) = R(E(w1)) = E(L(w1)) = E(w2)$.

$= E(L(w1))$ The representation function is based on the neural mechanism (law) L that operates on the inactive state $w1$ in the neuron

$= E(w2)$ The representation function shows the active state of the neuron w_2.

Thus, when these conditions apply, the mental model can predict the neuron's response properties and mental simulation can be successful (see fig. 22).

Additional modeling relations come into play when an artificial preparation is considered (see fig. 22). On the basis of the mental model of the natural preparation, an artificial preparation is designed and produced, e.g. on a computer. The artificial preparation is a second world that has another mental model M_a (a= artificial).

$S_a = <W_a, M_a, E_a>$ with:

> The modeled system is an artificial world
> $W_a = <W_a, L_a>$ with states $W = \{w_a\}$
> and actions or laws $L_a: W_a \text{ -> } W_a$.
> For example, W_a could be a variable in a computer program, while L_a is an algorithm or method;
>
> The modeling system
> $M_a = <M_a, R_a>$ with internal model states, or representations $M_a = \{m_{aj}\}$
> and a set of rules, or a modeling function $R_a: M_a \text{ -> } M_a$.
> M_a *and* R_a are analogous to the natural modeling relation, i.e. cognitive processes;
>
> The representation function
> $E_a: W_a \text{ -> } M_a$.
> Corresponding to the observations in the experimental setup, E_a can be the observable computer output.

It should be noted that this approach offers no way to determine directly whether the artificial preparation commutes with the natural preparation or the mental model of the natural preparation – the only way to determine this, is to compare the mental model of the natural preparation with the mental model of the artificial preparation. This can be achieved by the third modeling relation of the cognitive model that ‘reuses’ the former two models, but necessitates a new, cognitive representation function.

$S_c = <M, M_a, E_c>$ with:

> The modeled system is identical to the model of the natural preparation
> $M = <M, R>$ with states $M = \{m_j\}$

and actions or laws $R: M \rightarrow M$.

The modeling system is identical to the model of the artificial preparation $M_a = <M_a, R_a>$ with internal model states, or representations $M_a = \{m_{aj}\}$
and a set of rules, or a modeling function $R_a: M_a \rightarrow M_a$.

The representation function E_c is exclusively cognitive:
$E_c: M \rightarrow M_a$. If M and M_a commute, W and W_a should also commute.

Thus, the cognitive domain is the only possibility to evaluate if (natural) world and artificial world commute. Put more generally, assessing the aptitude of a model is strictly observer dependent. And the use of artificial preparations is an elegant means for performing this check of a model for oneself. It closes the modeling loop from the side of the abstract version of the natural preparation. Simultaneously, the externalization W_a of the mental model M is a way of presenting the model M to other observers without the necessity of using the natural preparation. In a sense, verbal reports on M (e.g. talks or journals papers) can be regarded as restricted versions of artificial worlds W_a that can serve as an instruction for others of how to setup M. What is missing, however, is the artificial momentum with all its benefits (see also 2.2.2.1), which is only given in the case of artificial simulation. The momentum has to be generated during mental simulation. Thus, rather than providing a complete artificial modeling relation, such restricted versions of W_a serve to initialize the cognitive representation function E_c.

Setting up a cognitive system S_c without artificial modeling is the case of performing thought experiments with a second (abstract, maybe often formal) mental model M_a. This is the case of analogical reasoning (see e.g. Markman & Gentner 2001). Finally, thought experiments without the modeling relation to the world W are also conceivable and can be seen as the basis of designing an artificial model.

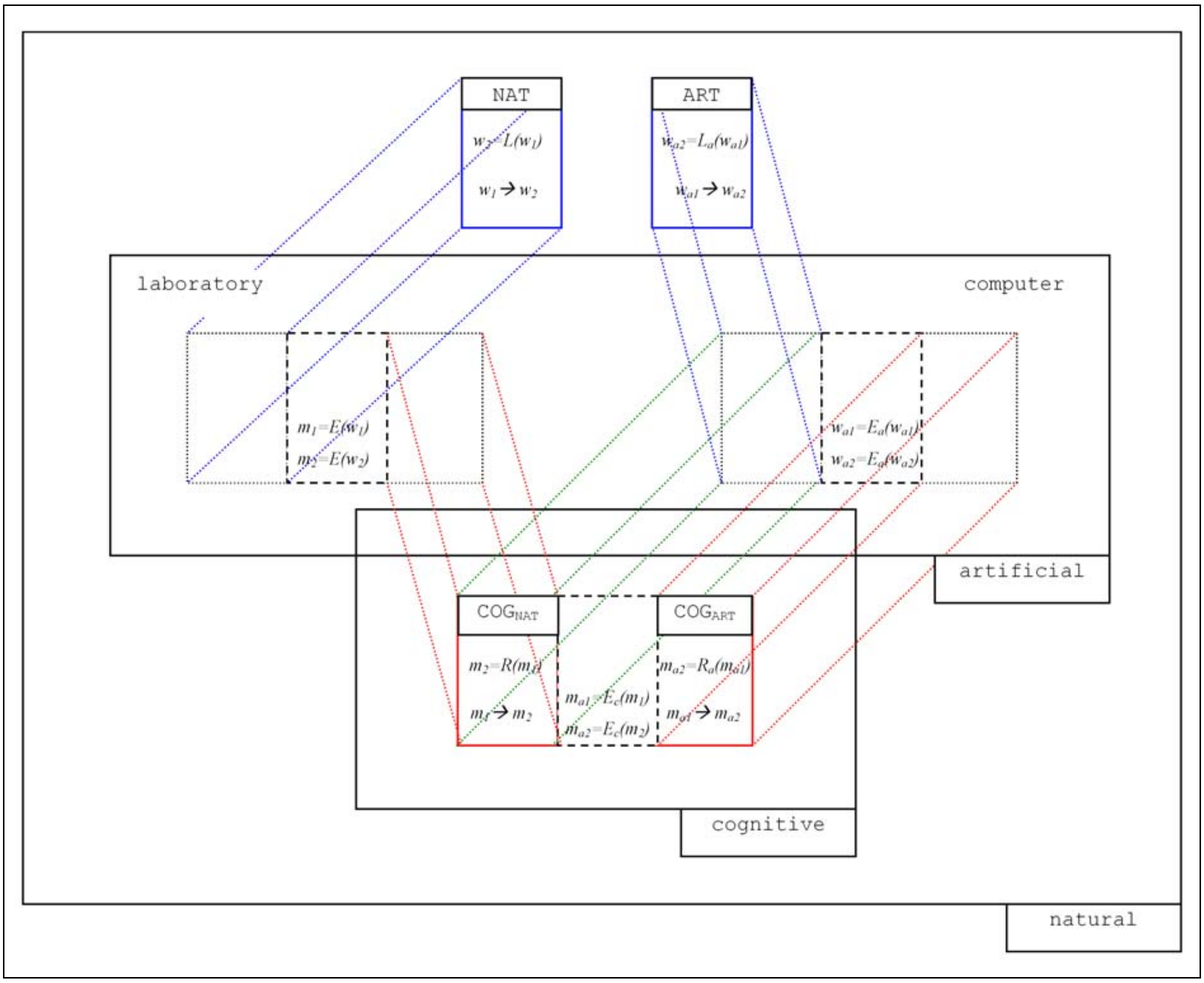


fig. 22: The modeling relations in scientific work. The scientist (the cognitive domain) models both the natural preparation and the artificial preparation. The artificial domain with laboratory and computer provides the representation function between preparation ('world') and model. General function of simulation is to find commuting mental models of the cognitive and the artificial preparation COG_{NAT} and COG_{ART}.

It can also be stated that the experimental situation (e.g. the neuroscientific experiment) is a model for a general natural mechanism (e.g. figure detection). In biology this is often called 'animal model' (see e.g. Schaffner 2001 for a detailed account). Here it is termed 'natural preparation'. This modeling relation can be understood with the same approach as artificial modeling: the experiment shifts to the position of the artificial and nature takes the place of the experiment. Correspondingly, the artificial model has a natural dimension in terms of a computer hardware that represents the artificial model and the mental model has a natural dimension in terms of a brain representing the mental model. (And the fly models the object during figure detection…) However, all these relations add nothing essentially new to an understanding of scientific work as it is proposed here.

2.2.4. Strong and weak Simulation

The modeling relation between natural, cognitive and artificial provides three simulations: natural simulation, artificial simulation and cognitive (mental) simulation. The former two realize a domain transition from external (natural and artificial) to internal (cognitive) with the respective representation functions. Whereas the natural domain is *as it is*, the artificial domain can be seen as an interaction layer between natural and cognitive that is designed to cause perceptions and to receive actions and, therewith, realizes the representation function. Both natural simulation and artificial simulation represent pre-existing external states. Mental simulation, on the other hand, generates its own, internal, representational domain. Thus, there is a principal difference between natural or artificial simulation and mental simulation. Strong mental simulation is given if the mental model M is modeled as another mental model M_a by the cognitive representation function E_c. The other modeling relations (to natural and artificial preparations) apply representation functions that involve perceptions and actions relating to the world, be it W or W_a. These worlds show a momentum that is modeled (or represented), but not simulated in the cognitive domain. Only E_c generates a genuine cognitive momentum that makes out the case of mental simulation. Since E_c makes possible an exclusively internal operation of representations it can be called the *realization function*.

Since this realization function does not have to comply with the constraints of the natural or artificial preparation, it is easier to change the direction of the modeling relation. For example, in the case of natural simulation, the scientist's cognitive processes M are the model for the neuron in the fly's brain W, but the neuron is not the model of a cognitive process. But in the cognitive domain it can very well be the case that the initial modeling relation is reversed. Consider that the scientist has a mental model M of a neuron in the fly's brain and an artificial (abstract) version of that model M_a, e.g. a formula for the response behavior of that neuron. Now, computing the formula can (and shall) very well make predictions for the behavior of the natural neuron. Then, the mental model of the artificial becomes the source in the modeling relation while the mental model of the natural becomes the model. Thus, first the artificial mental model M_a is model for natural mental model M. But then, M_a becomes the source for M. In this case, the representation function is bi-directional: E_c: M -> M_a and E_c: M_a -> M. This relation is the prerequisite for generating an inner word or, as Cruse (2003) puts it, a system "having internal perspective".

Put more generally, this approach arrives at a distinction between a strong and a weak sense of simulation. In the strong sense, simulation is only given if the model, the modeled system and the operational rules reside in the same domain ('internal'), e.g. the cognitive domain. It is not principally excluded that an artificial system can also realize a strong simulation, but this question has been sufficiently discussed (Searle 1980) and is left to experts. In the weak sense of simulation, model and source reside in different domains.

In the account of scientific work provided here no other strong simulation than mental simulation can be found. (Definitely, there is no simulation relation between the World *W* and the Artificial World W_a as one might be tempted to think.) Weak simulation applies also to the notion of a simulation being a running model: the scientist mentally simulates the natural or artificial preparation by operating the model in the lab or computer (see also 2.2.2.2). In these weak cases, sensorimotor coordination is usually necessary to bridge between the domains, e.g. the scientist has to perform actions on the natural or artificial preparation. The system is "reactive" rather than genuinely "cognitive" (cf. Cruse 2003). Additionally, the natural or artificial preparation itself contributes its momentum to the simulation. This implies that the operated model, i.e. the rules *R* applied to the states *m* ("enacted elements and activities") are caused and controlled by external events. In strong simulation, the model must be operated inside the domain. Generating these causal powers is, literally spoken, 'making sense'. The operational rules, e.g. the parameterization of the model can be controlled from the cognitive domain in both weak and strong simulation. Artificial simulation is closer to strong simulation because a mental simulation had to be performed to create the artificial simulation. Natural simulation is the weakest form of simulation and, therefore, is not necessarily conceived as simulation but rather as a natural (sometimes misleadingly called "real") interaction.

2.2.5. Summary

Scientific work aims at explaining a phenomenon that relates to a system having certain elements and showing certain activities that are causally related. An explanation typically implies a mechanism that causally relates elements and activities. Scientific work can be described as a four-step process:

1. Theory: definition of elements and activities in the scrutinized system and hypotheses concerning causal relations.
2. Design: operationalization of the hypothesis by staging elements and activities.
3. Experiment: operation of the design and recording of data.
4. Evaluation: checking data against hypothesis.

If a theoretical implication is concluded from the evaluation, a new iteration of the process can be initiated, providing new conclusions …

The means by which the scientist manages scientific work can be described as a mental simulation that shows also four aspects (coarsely corresponding to the four aspects of scientific work):

1. Problem: observation (phenomenon) causes a question (inconsistency or ambiguity).
2. Model: elements, activities, causal relations and operational rules are assembled so that an answer becomes possible.
3. Operation: the model is active.
4. Analysis: observation of the model's behavior is tested as a possible answer.

Even though simulation primarily applies to the operated model (3.), the whole process (1.) – (4.) can be termed mental simulation since the steps are difficult to separate from each other. Mental simulation is a limited capacity process that forces the scientist to prepare and organize investigations in the different domains:

1. Natural: the scientist uses a natural preparation (typically in a laboratory experiment) or observes a natural situation (in the field).
2. Artificial: the scientist uses an artificial preparation (typically a model in a computer or a robot).

3. Mental: the scientist uses a mental model and generates a representational domain independent from the external world, typically, the case of scientific reasoning and performing thought experiments.

Natural preparations and artificial preparations and mental simulation can be mixed in a given scenario of scientific work. There is no other principal difference between natural and artificial preparations, but that the natural shows an inherent functionality – a momentum – that is *as it is* (not designed by cognition), whereas the artificial has a momentum that was designed by cognition. Operating mental models for representing the external domain (natural or artificial) is termed weak simulation since the momentum is caused by external elements and activities. Operating internal models by way of internal representations is termed strong simulation because the momentum is caused by an internal generative mechanism.

This analysis of scientific work was started to describe the role simulation plays in science. Beginning with a notion of simulation as computer programs it became obvious almost immediately that computer programs are not to be understood as media but rather as cognitive processes of the scientist. This introduced the concept of *mental* simulation. Seen from the 'mental simulation stance', computer simulations are artifacts that are made for serving a guiding function for mental simulations. If the concept of mental simulation is applied consequently, investigation on non-artificial, natural situations is to be considered, too. So it turned out that there it is no way too account for the role simulation plays in science without treating science as simulation!

As already mentioned, it might seem strange on first glance to stress the simulative aspect in experimental scientific work (or thinking in general) – experiments are usually thought to relate to "real" nature. But it is much easier to conceive all scientific work as simulation than conceiving some aspects as real and others as unreal. Simulation offers an alternative perspective of scientific work (or thinking in general). In brain sciences, discussions on biological plausibility are frequently observable (see e.g. Crick 1989; Webb 2001). The question that is frequently tried to be answered is how the natural relates to the artificial (nature vs. model)? In the present approach this question of 'reality' is ill posed: there is neither a natural nor an artificial phenomenon without mental simulation of the scientist. Therefore, the question how the relations of both preparations to the mental simulation can be described should be the primary objective. *"How does a given preparation help to explain the phenomenon?"* Of course, this introduces an epistemological aspect in the discussion that can be annoying, too. However, this aspect is not necessarily philosophical, but rather can be targeted towards a practical or even pragmatic discussion. The question is: *"How can we find good explanations for the phenomenon under scrutiny (and what are these explanations good for)?"* Admittedly, this implies to abandon ontological questions such as: "The nature of figure detection is ..." But instead of an often futile wish for answers to the ultimate nature of something, another aspect is put to the center (where it belongs): function and value of human explanations of nature.

2.2.6. Implications: "Simulation, media and explanation"

Rather than providing general notes of the role of simulation in science for concluding this chapter, implications for the enveloping sections, namely the account of simulations as media (see also 2.1) and the following chapter on explanation (see also 3.1) will be provided.

Simulations as media have a straightforward meaning in the scientific simulation framework proposed above: they are the externalized version of a mental model, the artificial world W_a. It can be stated that the typical case of 'modeling' is particularly this process of externalization. According to the generic simulation scheme, the externalization ('modeling') can only be assessed by setting up a second modeling relation that again is related to the first modeling relation by the cognitive simulation function E_c. In this sense, *there is no serious modeling without simulation*. (Of course, there is no serious simulation without a model, too.)

The section on simulation as media contains also a general sketch of simulation (see also 2.1.3). What was termed 'source' there is the modeled system W, the 'representational domain' is the representation function and the cognitive system corresponds to the model M. Generally, the sketch of simulation from the section on media corresponds and complies to a single modeling relation provided in scientific simulation framework. The section on simulations as media emphasizes the topic of interactivity ('intervention ports') because media are otherwise conceived as external elements that do not necessitate active intervention. Stressing intervention ports should clarify that the external character might be true for other media but not for simulation. Simulation necessitates an active cognitive (and/or sensorimotor) component. This characteristic of simulations as media can also be understood in the context of the scientific simulation framework. A simulation (and this applies to all of the approaches named in 2.1.2) entails not only the artificial domain W_a (medial part) but also the cognitive domain. Furthermore, the medial part of the simulation W_a cannot build the realization function E_c. Since there is no direct relation between the artificial preparation (e.g. a computer program) and the natural preparation (e.g. a fly), the user must build the realization function (e.g. relate computer program to fly).

But the medial part of the simulation can contribute instructions for building the realization function. Consider the difference between artificial modeling and artificial simulation. In artificial simulation the momentum is in the medial part, whereas in artificial modeling (film, a painting, a text etc.) the momentum is not in the medial part, but added by cognition. Thus, the first step towards the realization function is to demonstrate the functioning of a system in artificial simulation (a computer program with a neural circuit of the fly). It provides an artificial momentum that helps to set up the artificial modeling relation. In a second step, the artificial momentum is manipulated. Intervention ports (e.g. sliders, buttons that change stimuli or response properties) are such manipulations. These manipulations (e.g. a parameter choice) are operations on elements, activities or causal relations in the artificial domain – it is demonstrated how to operate elements, activities and rules of the mental model (e.g. the computer simulation). Thus, part of the operation is taken away from the artificial domain and transferred to the cognitive domain. Precisely, the operational rule is not controlled by the artificial but by the cognitive. An operational rule, for example, is the assignment of a value to an ambiguous (polyvalent) attribute (e.g. assigning *3* to *x*). Applying the operational rule means to determine the consequences of the assignment (e.g. computing a formula with $x = 3$). Controlling the operation by defining and performing operational rules for the mental model is a step towards strong simulation in which, eventually, elements etc. are operated independent of the external world *W*. The final step after the internalization of the operations is the generation of a cognitive momentum.

States of the system can be predicted and operated internally. The realization function is built.

It becomes clearer at this point, why simulations as media can help to manage dynamics and complexity. Serialization and decomposition are the major strategies (see also 1.1.12). In a first step, the operational rules can be specified and outsourced to the medium. The artificial momentum relieves internal processing capacities and simultaneously provides patterns for mentally simulating the model. In a step after outsourcing, reintegration of the operational rules can be done by internalizing ('learning') these patterns. The mental model can be partially operated with these patterns. Ultimately, the goal of mental simulation is to internalize the operation of the model completely, i.e. in terms of automating operational rules so that the simulations can be performed as an inner film rather than a slide show involving a choice of each slide and its presentation parameters between each slide. In the case of automation, a cognitive momentum can be realized: the 'cognitive preparation' has an inherent functionality like the natural is autonomous or the artificial operates automatically.

An *explanation* can be conceived as such a pattern, as a serialized operation of a model, as a specific trajectory through its state space. Spoken or written language is one (widespread and effective) form of explanation. But explanation is not principally lingual, but can be conceived as mental simulation. Generally, an explanation has the same structure as a simulation: there is an explanandum (phenomenon) and an explanans corresponding to the modeled system ('source', 'natural preparation') and the model, respectively. The relation between explanandum and explanans corresponds to the representation function. The elements, activities and causal relations are equally present in an explanation and in a simulation. The difference between simulation and explanation is that an explanation has a serialized character by applying specific operational rules and it uses specific values for the attributes. Thus, an explanation corresponds very well to a specific simulation. Of course, there is another striking difference between simulation and explanation: explanation shall cause understanding. Thus, an explanation is a simulation with the somewhat explicit ('aware') intention to comprehend. Understanding takes place if a specific (parameterized) modeling relation is commutative. In this view, an explanation can be present in the natural or artificial or cognitive domain, i.e. can be dependent on a laboratory experiment, a computer simulation or a thought experiment. Verbal explanations can be conceived as specific cases of artificial simulation (except speaking soundless to oneself, which is rather a thought experiment). Explanations, as the target of scientific work, are internalized representations of the world. Mental simulation can be conceived as the mechanism that produces explanations by staging elements, activities and causal relations according to specific operational rules. Artificial simulations can help to get mental simulations running.

Reading Advice

The analysis of simulation in science led to framework incorporating mental, artificial as well as natural simulation and it led to a generic simulation scheme describing an activated model driven by operational rules that trigger an autonomous change from an initial to a goal state ('momentum'). It was shown that the generic simulation scheme reconciles the various notions introduced in the Special "Simulations as media" and provides a clear-cut conception of simulations as media. Consideration of the scientific simulation framework allowed elucidating the role of simulation for understanding, learning and explanation in science. It was also shown how simulation helps to handle dynamics and complexity, which are a major impediment in explaining brains and a major motivation for applying simulation in brain science. Thus, the overall objective of this study – examining the role of simulation in explaining brains – is achieved to the part that concerns science: It was shown that mental simulation is crucial to scientific reasoning and that artificial (as well as natural simulation) can guide mental simulation and relieve cognitive capacity by providing external input (operational rules, perceptual activity, momentum etc.). But explaining brains by simulation is not necessarily a scientific issue! Science is rather a good practice scenario of simulation. Thus, one question to be answered to complete the examination is left: Is simulation also a general explanatory mechanism? Does the generic simulation scheme also hold in contexts other than science? An answer to that question will be given in the next section that provides an approach for treating explanation as a natural, cognitive phenomenon and simulation as a general explanatory mechanism. The "Explanation-Special" that concludes the study demonstrates that this approach even allows to empirically analyze the issue of explaining brains by simulation.

3. EXPLANATION

An explanation is something that answers questions concerning a phenomenon. In the scientific context, explanation is often regarded as a rather general, primarily logical issue, particularly in the philosophy of science (see Carnap 1995 for an introduction). In this formal-logical notion, it is implicitly assumed that explanation is something independent from an explainer (e.g. a scientist) and the mental states: it has an existence independent from a specific temporal or spatial context. It is not viewed as a natural phenomenon. But there is another notion of explanation that relates to the actual performance of a scientist or any other cognitive system, e.g. a scientist lecturing students or a parent informing children. In this notion, explanation can be understood as a natural phenomenon because it is embedded in a concrete spatial and temporal context. Of course, both notions correspond and it is possible without further ado to use the term explanation without specifying which notion is meant. However, for arriving at a more detailed understanding of the role simulations play for explanations, the concept of explanation has to be dissected. Both the logical notion from philosophy of science and the natural notion used in everyday language contribute to the concept of explanation: logical explanation implies precise specifications for form and content of an explanation e.g. in terms of formal logics (Lemmon 1987) and natural explanation describes what is actually happening during an explanation.

From the naturalist's point of view, logical explanation is, of course, a specific case of natural explanations: not every natural explanation is logical, while every logical explanation has a nature. There are several approaches that aim at naturalizing scientific explanation (see e.g. Giere 1992; Carruthers *et al.* 2002). In order to understand the role of simulations in scientific work, a similar (natural) approach was already implicitly applied in the chapter on simulation (see also 2.1). Accordingly, in the following section *logical explanation will be treated as the final, well-formed and valid result of the process of natural explanation*. Arriving at a logical explanation in a naturalization procedure involves a variety of highly complex phenomena such as verbalization, writing, discussing, peer-reviewing etc. These issues shall altogether be put aside in the present context in favor of focusing on the role of simulations as a generative mechanism in explanations. Thus, for avoiding lengthy philosophical discussions in this introduction, details of logical explanation shall only be applied when necessary – more relevant for describing the role simulation plays in explanation is to analyze the cognitive (natural) domain.

The situation of natural explanation comprises an explainer, an explanans and an explanandum (see also 2.2.6). The relation of explanans and explanandum can generally be conceived as a modeling relation (see also 2.2.3). The explainer uses the modeling system to represent the modeled system. Mental states represent elements, activities and causal relations of the explanandum. The neuroscientist, for example, uses a mental model of a natural preparation to represent the natural preparation (e.g. a fly's neuron). Applying operational rules to this mental model allows for a mental simulation of the model that reveals specific behaviors. In sum, an explanation contains all aspects of a simulation: a model (elements, activities and causal relations) as well as operational rules. This also implies that a model alone will not suffice to construct an explanation, while simulation does. On this view, simulations can be conceived as a mechanism for generating natural explanation. The task of analyzing natural explanation corresponds to the task of analyzing the nature of simulation. As a consequence, explanation is not necessarily a verbal description, but can also be purely mental, e.g. explaining to oneself. These cognitive processes are necessary and sufficient and the minimalist case of explanation.

In conclusion, the task of analyzing natural explanation reduces to the task of analyzing simulation as cognition. Consequently, the strategy underlying this chapter is to introduce an account of simulation as cognition. In the first section, simulation will be described as a cognitive phenomenon by applying the *generic simulation scheme* that was developed along the lines of the *modeling relation* (see also 2.2.3) in the previous chapter. Cognitive theories that are important resources for an account of simulation as cognition are reviewed in the following sections. Candidates that can explain the generic simulation scheme as a cognitive mechanism shall be filtered out and analyzed in more detail.

3.1. Simulation as cognition

The information processing approach, prevalent in cognitive science (see also 1.3.6) offers a simple and well-known entry point to cognitive theories.

Regardless of whether it perfectly matches everyone's demands (see 1.3.6.2 for a discussion), it can be taken as a framework in which the role simulations play for explanation can be extended in the cognitive domain. The information processing approach generally assumes stimulus, response and information processing in between. This can also be described in terms of the modeling relation (see also 2.2.3). A stimulus corresponds to a world and a modeled system (explanandum), information processing corresponds to representation function and modeling system. Finally, a response corresponds to the behavior of the system that can enter further processing. Information processing typically begins with sensory processing. For example, a man observing a moving twig receives sensory input on the visual scene that is encountered during a walk. These sensory representations already contain a raw model of the world since certain elements and activities are separated from other aspects of the sensory representation (cf. Marr 1982), but will fade away almost immediately unless they are gated (e.g. by selective attention) for 'central' processing. This was already explained to be the case when inconsistencies turn up (see also 2.2.2.1). Central processing is typically associated with short-term memory that allows operating on sensory representations for a considerable but limited time. For example, the observation of a moving twig can be reviewed in short-term memory with respect to its possible causes, such as wind, bird, monster etc. Elements, activities and causal relations build a model of the situation. Operating on the model can be conceived as simulation in terms of hypothesis testing (see also 2.2.2.1). Simulation may lead to elaborated versions of the model that can enter long-term memory that, in turn, can reentry the ongoing simulation and provide elements, activities and causal relations, e.g. a bird in the mental simulation of the moving twig scene, which is actually not there.

An explanation is a consistent simulation, i.e. the case when world and model commute. For example, both the wind and the bird are possible explanations for the moving twig. They are not necessarily competing because a decision for this or the other possibility does not necessarily have any severe consequences for the observer. But the possibility of the monster must be resolved because it would imply danger. A resolution could be the categorical exclusion of the possibility due to the believe that monsters do not exist. This evaluation of the model is an example for further operations on models, i.e. elaborations. According to the information processing approach, elaborations increase the probability of recall, i.e. involve long-term memory. Elaborating observations therefore corresponds to *learning*. If elaborations are pursued with more effort, still more alternative explanations can be taken into account and (thought) experiments can be performed in order to assess these alternatives. These efforts will possibly lead to a logical explanation that has a more general form such as: "A twig can be moved by application of a force that exceeds its elastic resistance." An ultimate result of the process of elaborating an explanation is *knowledge* about an issue that is expressed as the ability of the explainer to recall the elements, activities and causal relations and integrate them into ongoing simulation.

Taken together, the information processing approach offers an understanding of simulation in cognitive terms. It becomes clear that simulation involves all basic aspects of a cognitive system: thinking or reasoning, attention, perception, learning, memory, knowledge, representation etc. This is no surprise since simulations tend to come into place when cognitive processes relate to complex and or dynamic situations ("knowledge rich domains"). These situations demand all that cognition has to offer.

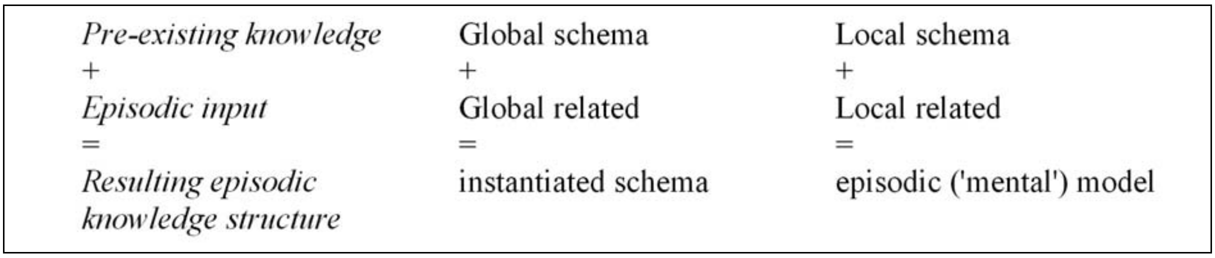
3.1.1. Cognitive theories

3.1.1.1. *Knowledge*

For an understanding of the cognitive aspects of simulations, it should be clarified what the cognitive items are that are processed in a simulation. Thus, an adequate concept of knowledge structures has to be found. As has been pointed out before (see also 2.1.2.3), the type of knowledge to be dealt with in simulations is complex and dynamic. Cognitive psychology offers several approaches of such complex or 'molar' knowledge structures, e.g. schemas (Bartlett 1932; Rumelhart 1980; Mandler 1984), frames (Minsky M. 1975), scripts (Schank R. & Abelson R. 1977) and mental models (Johnson-Laird 1980; 1983; Gentner & Stevens 1983). According to Brewer (Brewer 1987) the former three can all be subsumed under schemas, while mental models have to be treated separately (see fig. 23). Brewer circumscribes schemas as "unconscious mental structures that underlie the molar aspects of human knowledge and skill". They involve "old" generic information. An instantiated schema is a "specific cognitive structure that results from an interaction of the old information of the generic schema and the new information from the episodic input". (Episodic input can be conceived as a kind of sensory representation.) Mental models, as opposed to schemas, shall account not only for old information, but also for situations we have never been in before, e.g. imagery and inference. Concerning the differences between mental models and schemas, Brewer points out that schemas are "precompiled generic knowledge structures", while mental models are constructed at the time of use. Accordingly, the methodological distinction is: inquiries into schemas involve items known before the experimental situation began, while mental models aim primarily at knowledge generated in the experimental situation. In addition, schemas relate to more global knowledge, while mental models relate to more local (specific) knowledge.

Similarly, Markman (1999) describes schemas as 'general belief structures' and scripts as 'schemas that wrap event sequences'. He adds naive theories or folk theories encompassing larger domains such as biology. In a description of molar knowledge structures as simulators, Barsalou (1999) comes to a similar conclusion about the way memorized perceptual symbols are organized. (Perceptual symbols are specific representations that are taken in to account in 3.1.3.2). Unlike Brewer (1987), Barsalou chooses the term "frame" for denoting the ordering structure in which 'old' information is represented, but, like Brewer, he states that frames are similar to schemas and scripts. Frames integrate perceptual symbols and are contained in simulators so that potentially an infinite number of simulations can be constructed. Like Brewer, Barsalou refers to a mental model as non-generic but rather a specific structure. However, in his view, "mental models tend not to address underlying generative mechanisms." Barsalou's account can be summarized as follows:

perceptual input + simulators (or 'frames') -> simulation (~ mental model)



Pre-existing knowledge	Global schema	Local schema
+	+	+
Episodic input	Global related	Local related
=	=	=
Resulting episodic knowledge structure	instantiated schema	episodic ('mental') model

fig. 23: Concepts on molar knowledge structures. *"Instantiated schemas are specific knowledge structures derived from generic knowledge represented in global schemas, while Episodic models are the specific knowledge structures that are constructed to represent new situations out of the more specific generic knowledge represented in local schemas" (after Brewer 1987).*

Recently, the cognitivistic approaches to understand (the generation of) knowledge introduced so far have been challenged by the claims of situated cognition proponents (e.g. Greeno 1989), who put forward the idea that contextual factors rather than cognitive operations dominate learning processes. There has been a continuing debate between Anderson et al. who criticized the situated cognition approach (primarily as empirically invalid) and Greeno (Anderson *et al.* 1996; 1997; Greeno 1997). The debate led into a partial settlement in a common description of 'territorial properties' (Anderson *et al.* 2000). Situated cognition has originated from research on Artificial Intelligence and attempts to account for puzzling phenomena, such as that the children in Brazilian street markets that having little or no schooling are able to perform nontrivial operations, e.g. determining complicated prices of lottery tickets. Knowledge based, cognitivistic approaches have difficulties explaining such results. Greeno describes the general problem as "the insulation of symbolic knowledge" (the loss of context) and proposes a framework for a more relational epistemology. But the role he attributes to mental models is similar to those attributed by Brewer (1987) or Barsalou (1999): they serve as the most valid approach to explain the generation of novel knowledge. Other recent cognitive theories are similar to that of Greeno in that they propose a domain-specific organization of cognition rather than content-independent mechanisms: 'the modularity-theory' (Hirschfeld & Gelman 1994), 'evolutionary psychology' (Tooby & Cosmides 1989) and 'embodied cognition' that focuses on the role of sensorimotor processes (Glenberg 1997). It would be beyond the scope of this study to review all of these approaches that could contribute to a theory of knowledge, even though they might contribute interesting aspects. But it already became obvious that the most interesting approach for describing such 'rich' knowledge structures – resembling those found in simulations – is mental models. Since mental models, therewith, become the most promising candidate describing simulation as cognition, theories of mental models shall be reviewed in more detail in the next but one section. Before, a brief review of theories on the cognitive processes that generate knowledge, namely 'learning' is provided.

3.1.1.2. *Learning*

Cognitive psychology offers approaches that take into account not only the learning outcome, namely 'knowledge', but also the learning process that leads to knowledge. For the sake of brevity, only approaches that deal with complex systems are taken into account.

The concept of 'Implicit Learning' refers to phenomena that elucidate discrepancies between measured performance and verbalizable knowledge. In a typical case, learners perform better after a training phase, but are unable to explicate their acquired skill or knowledge[22]. Interesting with respect to learning with simulations is the approach "control of complex systems" (e.g. Broadbent 1977; Berry & Broadbent 1988). In these experiments, subjects had to control parameters of a system (e.g. "city transport", "sugar production" or "person interaction"). The results consistently showed that the subjects' performance to control the system improved with practice, but practice had no effect on their performance in answering written questions after the test. However, there is a continuing debate about the existence and validity of implicit learning (Haider 1992; Berry & Dienes 1993) that is additionally expressed in a confusion on meaning and use of the implicit/explicit distinction in combination with the concepts of knowledge, memory and learning (Dienes & Perner 1999). The research tradition of implicit learning that is important for the present study, namely control of complex systems appears not to be continued consequently. The majority of the continuing research on implicit learning has applied the paradigms of "artificial grammar learning" (e.g. Reber 1967) and "sequence learning" (e.g. Nissen & Bullemer 1987). "Control of complex systems" is still applied (Cleeremans *et al.* 1998) but apparently less frequently.

'Complex Problem Solving' is a strong German research tradition and is dominated by the work of Dörner (see e.g. 1989; 1998). The focus of Dörner's work lies on the analysis of users handling large simulation systems with more than 100 variables, e.g. controlling the virtual city "Lohhausen" (Dörner *et al.* 1983). Most of Dörner's work differs methodologically from the aforementioned studies in that idiographic approaches (e.g. single case studies) are applied. Demarking from Dörner's work (but simultaneously neatly related to it) researchers in Germany developed several approaches for nomothetical, experimental studies (for reviews see Funke 1991; Frensch & Funke 1995). These studies apply small systems with less than 10 variables that are completely described mathematically. Measures are post task tests,

[22] Note, by the way, that implicit learning was found to go along with decreased brain activity (Buckner *et al.* 1995).

e.g. the ‘goodness’ of verbal reports or of causal diagrams. However, most of these studies had no measurement of performance during the task so that an operation of the model during the task (the simulation) was not monitored.

“Learning Decomposition” introduced recently by Lee and Anderson (2001) is to be seen more as a keyword than as an established approach. However, the study is worth paying special attention because it integrates many aspects one would propose for an extensive experimental analysis of learning with simulations and can, thus, serve as an exemplar. Subjects had to work on the Kanfer-Ackermann-Air-Traffic-Control-Task, a simulation in which airplanes have to be landed depending on several parameters like wind, other airplanes etc. It combines measuring learning and behavioral performance in one simulation task and integrates various aspects in an elegant manner: paradigmatic approaches of cognitive psychology such as Anderson’s theoretical framework (Anderson 1980) and formal models, such as ACT-R (Anderson 1993) are combined with actual accounts such as the ‘information reduction hypothesis’ (Haider & Frensch 1996) and corroborating behavioral measures such as eye tracking.

The approaches to learning as cognitive processes taken into account here, all referred to the learning of complex systems. Mental models were not considered so far. But the research tradition on mental models deals with complex systems and provides a rich resource on learning as well. Thus, theory of mental models provides both substantial approaches to knowledge and to learning complex systems. Therefore, it is the most promising approach for developing an understanding of simulations as cognition. Consequently, the focus of the following sections is adjusted to mental models. This should not imply that the other cognitive theories could not be valuable alternatives or supplements – but that is left to be shown by future work.

3.1.2. Mental models

In order to understand the underlying cognitive structure of simulation, mental models seem to be the most important approach. They are to be understood as the *actual* construct of elements, activities and causal relations that are to be modeled. Schemas, on the other hand, are specific (precompiled) sources from memory (and could be the format in which an elaborated model is represented). Mental models, however, seem to be more relevant for understanding the role of cognition in simulation because they incorporate the concept of schemas and can well be described as a generative mechanism for developing an explanation.

Inside the research tradition of mental models, two different streams have to be distinguished: one referring primarily to Johnson-Laird (1980; 1983) and one referring primarily to Gentner and Stevens (1983). While Johnson-Laird focuses on language comprehension, Gentner and Stevens focus on understanding devices and simple physical systems. As Gentner & Stevens put it in their introduction (1983), (their) research on mental models is characterized, first, by a specific domain of simple physical systems. These systems are well suited for analysis because they are based on explicit normative models. Second characteristic is the application of AI-theory (e.g. constraint networks, production systems) rather than mathematics. Finally, they propose an eclectic methodology (e.g. a mixture of protocol analysis, experimental cognitive psychology and simulations).

More recently, Gentner and coworkers (Markman & Gentner 2001) defined mental models as a representation of some domain or situation that supports understanding, reasoning and prediction. Again, it is differentiated between Johnson-Laird's 'logical mental models' as working-memory constructs that support logical reasoning and 'causal mental models' as the characterization of knowledge and processes that support understanding and reasoning in knowledge-rich domains. ('Knowledge-rich domains' and 'complex systems' presumably denote similar things...) Two distinctive features of a causal mental model are that their 'tokens' (i.e. what is processed) correspond to elements of a causal system (rather than the 'algebraic symbols' in logical mental models, see below) and that they involve long-term memory structures. 'Mental simulation' is considered as a prominent example for how people employ mental models. It can be circumscribed as the imagination of a future trajectory of a system given a set of initial conditions that are qualitative (relativistic) in nature (e.g. Forbus 1983) and might be strongly coupled with motor movements (Schwartz & Black 1999).

Indeed, as can be read from the more recent publications (e.g. Johnson-Laird 1999), Johnson-Laird's focus in the last years was inference (especially deduction) oriented on formal logics and investigated through text comprehension tasks. As Payne (1992) points out, Johnson-Laird's theory encompasses the basic theoretical commitment shared also by all authors in the Gentner & Stevens volume: that people's existing knowledge has a considerable influence on their reasoning about a new problem, phenomenon, device or idea. Payne sees Johnson-Laird's theory of mental models as more developed in that it specifies the format of representations and procedures, which are used to operate them. But Payne also points out that Johnson-Laird's theory is essentially a mental models theory of interaction with a particular artifact, namely text. Predominantly because of this focus on text (as opposed to the focus on physical systems in causal mental models), understanding the cognitive aspects of simulations appears to be better supported by the approach of causal mental models (Gentner & Stevens 1983). Even though text plays a dominant role in reasoning, it seems unlikely that it explains the operation of an experimental situation or a computer simulation. Two examples will illustrate what causal mental models are about.

A typical piece of research on (causal) mental models is given by Kieras and Bovair (1984). They use the term 'device model' to distinguish it from Johnson-Laird's use of the term mental model. They presented an unfamiliar piece of equipment (a phaser controller borrowed from StarTrek™) to subjects and tested whether and how the knowledge of the device's underlying principles that were mediated in training, affected the performance compared to a group not having that knowledge. The trained group was superior in terms of time, errors and inferential behavior. As further experiments indicated, the specific effects of these general benefits can be referred to advantages carrying out inferences (Experiment 2) with specific information (Experiment 3) on the model's functions.

Their practical suggestions for mediating a mental model were:

1. Mental models support inferences about specific (rather than general) tasks.
2. The knowledge can be incomplete because the user is able to carry out inferences in order to operate devices.
3. In case of easy devices or simple operations on difficult devices no mental model is necessary (task dependence).
4. Incorrect knowledge will impair performance.

This example of Bovair & Kieras illustrates that the approach of causal mental models nicely matches the typical situations in scientific work, i.e. device operation.

Another typical piece of research on causal mental models refers to the understanding electrical circuits (Gentner.D & Gentner.D.R. 1983). This is particularly interesting in the context of explanation in neuroscience because electrical circuits are not only physic's and engineer's basic lecture – also neuroscientists use these circuits to model electrically responsive membranes of neurons (see e.g. Rall 1989). Consequently, these circuits are potentially part of a curriculum for brain sciences. Primarily, they focus on an account of "generative analogy" as a basic process in scientific reasoning. The two analogous systems 'base' B and 'target' T are described as a propositional network typical for the schema theoretic representation of knowledge (e.g. Rumelhart & Ortony 1977). Nodes of the networks are concepts $b_1, b_2 \ldots b_n$ and $t_1, t_2 \ldots t_n$ that can have predicates A, $R, R`$ etc. (A = attribute = predicate with one argument, R = relation = predicate with two arguments). Analogy is described as a structure mapping procedure between B and $T (M: b_i -> t_i)$ in which relations $R(b_i,b_j)$ are preserved $M: [R(b_i,b_j)] -> [R(t_i,t_j)]$, while attributes are not preserved in any case. For example, in the analogy between the solar system and the Rutherford model of the hydrogen atom, the relation between 'central' (sun vs. nucleus) and 'peripheral' (planet vs. electron) and 'more massive than' or 'revolves around' are typically preserved, while the attributes $mass = 10^{30}$ kilograms or $temperature$ = 25.000.000°F are not. The principle of systematicity expresses that higher order relations, such as cause $R`[R(b_i,b_j), R(b_k,b_l)]$ are more strongly predicated than isolated relations in the structure mapping $M: [R`(R(b_i,b_j), R(b_k,b_l))] -> [R`(R(t_i,t_j), R(t_k,t_l))]$.

Gentner & Gentner tested two analogies of electricity for their aptitude of explaining electric circuits, namely 'water' and 'moving crowd'. In the water analogy, a simple electrical circuit with battery and resistor corresponds to a hydraulic system with a pump or reservoir and a constriction of the pipe. In the moving crowd analogy, the circuit is racetrack with gates (resistors) with mice (or vehicles) that obey a loudspeaker (battery). The tests asked for predictions of current and voltage in four circuits: serial batteries, parallel batteries, serial resistors and parallel resistors. The rationale of the study was: if analogies were actually used to solve the problem, differences in the analogies should result in different inferences and, consequently, different performance. For example, parallel resistors in an electrical circuit double the amount of passing current rather than reducing it. The moving crowd analogy should result in better predictions because two gates for mice or vehicles easily explain an increase in current, while a second constriction in the water analogy may also suggest another impediment that reduces current flow. Gentner & Gentner conclude that their account of generative analogy is supported by the results and that "analogies help to structure unfamiliar domains" and "can indeed serve as inferential frameworks".

Obviously, the approach to mental models presented by Gentner and Gentner relates closely to the generic simulation scheme (see also 2.2.3) proposed here. The structure mapping procedure corresponds almost perfectly to the modeling relation between the world model M and the artificial model M_a. Thus, the mutual interaction between two mental models (rather than one) that is put forward in the generic simulation scheme is assumed for an approach to scientific reasoning (even though Gentner & Gentner do not explicitly note it). In sum, the two concrete examples of research on causal mental models illustrate that they convey a well suited approach – conceptually as well as experimentally – for understanding the knowledge structures and the inference simulation is based on.

3.1.3. Mental simulation

Mental models describe comprehensively the knowledge structure underlying the generic simulation scheme (see also 2.2.3). For a description of mental simulation, though, the notion of a model 'in action' is still an open issue. Particularly, two aspects are strikingly missing in the approaches introduced so far. First, the principle of how the models enter the cognitive system and come out again is left unclear. How can the relations between the tokens of the natural, cognitive and artificial domains be described? Thus, an account of representation is missing. Second, the concrete operation of elements, activities and causal relations, i.e. the inherent functionality and the momentum of simulation were not covered so far. How are causal relations applied to entities in order to obtain a prediction? How do mental models relate to the stages and mechanisms of the information processing approach, e.g. attention, working memory etc. These questions call for an integrative understanding of simulation as a mechanism in a 'fully functional conceptual system', i.e. a system that can represent, perceive, analyze and, finally, explain. These two aspects of representation and conceptual systems are topic of the next sections. The final section, then, probes a conception of 'simulation in action' as a cognitive mechanism.

3.1.3.1. *Representation*

Mental simulations are operated (or enacted) mental models. They are internal ongoing activity. But the elements and activities of the simulations are determined by external ongoing activity. The neuroscientist probing different stimuli on a visual neuron, for example, has to perform a mental simulation of the characteristics of the visual neuron that provides clues for deciding which stimulus to probe next (see also 2.2.2.1). This internal ongoing activity depends also on external ongoing activity, namely changes in electric activity of the fly's brain audible via audio-monitor or visible via oscilloscope. Thus, the actual sensory input of the neuroscientist shapes and changes the mental simulation and determines motor output (e.g. the next stimulus probe) that determines further sensory input (e.g. the next response of the fly)… Thus, the ongoing character of simulation exceeds the purely cognitive domain and introduces the necessity to consider the whole situation of the explainer (e.g. the neuroscientist) in an environment (e.g. the experimental setup and the natural preparation). Therefore, it is not sufficient to simply presume elements and activities in the model as given representations (e.g. as being something in the head of the explainer) without specifying how they can be conceived as a natural phenomenon in the situation the explainer resides in. This question calls for a natural account of representation, i.e. an account that allows embedding mental simulation in an actual situation and specifies the relation between internal and external elements and activities.

Interesting for approaching a natural notion of representation is the account of representation provided by Bechtel (2001b). He suggests applying the concept of representations also for 'simple' physical systems. According to his account, the state of a part of a mechanical device (e.g. bike's shifting lever) represents ("carries information about") an object or event (e.g. gear) if it is used for behavior. Y carries information about X for Z, which uses Y in order to act or think about Y. The state of the shifting lever represents the gear in the drive train that uses the shifting lever to adjust the gear of the drive train. Similarly, the state of the scientist's mental model represents the characteristic of the fly's neuron that is used to determine the stimulus probe. This account of representation is use-dependent. It should be noted that this account is consistent with most uses of representation in the Neurosciences, e.g. a neuron's firings represent features in order to use them for realizing a given function[23]. So, Bechtel's account even allows to further naturalize the representation of the fly's neuron: the neuron's firings in the head of the scientist represent aspects of the characteristics of the fly's neuron that are used to specify the parameters of motor activity needed to show the next stimulus probe.

It should be noted that Bechtel's account of representation differs substantially from many approaches to representation that are used in the cognitive sciences – first and foremost Fodor's account (Fodor 1975; 1987). Due to its abstract, amodal character Fodor's account or, put more generally, the propositional account of representation begs the question "...of how representations might be embodied in brains." Bechtel proposes to consider the approach of perceptual symbols, developed by Barsalou (1999) that mediates between Fodor's and Bechtel's account of representation.

Barsalou puts forward a modal (e.g. bound to a sensory modality) account of representation that is supposed to avoid the problems of the amodal accounts favored by Fodor and also by many connectionists (McClelland & Rumelhart 1986). The amodal account assumes that representations do not imply references to the perceptions they originate from. This makes them flexible ('arbitrary') in use and, therewith, the favored candidate for designing conceptual (e.g. cognitive) systems (artificial intelligence) in the last decades. As the main problems of the amodal account Barsalou names the little "direct empirical evidence", "the symbol grounding problem" (How are amodal symbols mapped to perceptions, if they contain no reference?) as well as "unfalsifiability" that results from their ability to explain virtually any finding post hoc, while failing to make predictions. In a modal account, symbols are represented in the same mode as the perception that produced them. This account is, in his view, unrecognized: since modal systems satisfy

[23] However, according to this view, cells should not be viewed as feature detectors but rather as "...filters with a representational profile." Bechtel stresses the neuroscientist's need for such an account of representation. They need it to make sensible statements about the function of neural systems for controlling the organism's behavior. A challenge following from this view is that the use of representations is "...often many steps removed from any behavior" (Bechtel 2001a).

the conditions of being inferential, productive and supporting propositions, required for building a conceptual system, they should no longer be conceived as pure recording systems that only provide input for a central processing stage.

In conclusion, Bechtel's account of representation provides an easy to grasp notion for representations as elements and activities (corresponding to neural activities, for example) that are operated in a simulation. Barsalou's perceptual symbols provide a first idea what representations actually are. Naturalized representations are the premise for integrating elements, activities and causal relations in the framework of natural, cognitive and artificial domains (see also 2.2.2). How these representations are actually operated in a fully functional system will be shown on the next section.

3.1.3.2. *A conceptual system*

In order to understand how models are operated and become inherently functional, simulations have to be embedded in a fully functional conceptual system (e.g. a 'cognitive' system / an explainer). This implies the questions:

i. What are the demands of a conceptual system?
ii. How does the generic simulation scheme (see also 2.2.3) relate to the conceptual system?

Rather than developing yet another account of a conceptual system that is based on the generic simulation scheme, an existing account shall be described in detail and compared to the simulation scheme. Fortunately, Barsalou's 'perceptual symbol systems' (1999) are not only a good approach for naturalizing representations (see also 3.1.3.1) but are also to be seen as a full conceptual system that incorporates most cognitive processes such as attention, memory etc. Moreover, it directly refers to simulation. Therewith, it is the optimal candidate of serving as a pattern for relating the generic simulation scheme to a conceptual framework. This exemplary analysis of a conceptual system shall show if the simulation scheme can be connected to other cognitive processes. Connectivity would be a sign to be on the right track, while inconsistencies or contradictions show where still problems are.

Barsalou suggests six core properties of perceptual symbols that will be described in the following. The empirical evidence he provides will not be reviewed. Paraphrases (not necessarily citations) of Barsalou's account are shown *in italics*. Comments that relate Barsalou's account to the generic simulation scheme are added afterwards as plain text.

1. *Perceptual symbols are record of the neural states that underlie perception. They can be processed consciously or unconsciously.* This is the notion of representations being natural phenomena (see also 3.1.3.1). Perceptual symbols correspond to the elements, activities and causal relations that are operated in the generic simulation scheme.

2. *Perceptual symbols are schematic. They represent coherent aspects of a brain state (not the whole brain state) 'chosen' via selective attention and stored in long-term memory. Furthermore, they are dynamic (as an attractor) and componential.* This statement shows how perceptual symbol systems relate to the 'classical' information processing approach, i.e. how they relate to concepts like attention, memory etc. Additionally, the extraction of elements, activities as coherent aspects of a brain state is implied. The storage in long-term memory would be possible, but not a necessary criterion in the generic simulation scheme.

3. *Perceptual symbols are multimodal. i.e. they include sensory modalities (audition, haptics, olfaction and gustation) as well as proprioception and introspection (representational states, cognitive operations and emotional states).* This property makes clear that all representations of a given situation the explainer (e.g. the neuroscientist) resides in are integrated in one domain, namely the cognitive domain. And this property illustrates that it is not necessarily an abstract element or activity that is to be operated in a mental simulation – it can very well be concrete such as a switch of an oscilloscope.

4. *Perceptual symbols "become organized into a simulator that allows the cognitive system to construct specific simulations of an entity or event in its absence". These are similar to dispositions, schemata and mental models. The difference between simulators and mental models is that the former not only address "simulations of specific entities and events" but also "underlying generative mechanisms that produce a family of related simulations" (Mental models are specific simulations). A simulator is equivalent to a concept*

(including knowledge but also accompanying processes) and sets up categories and allows categorical inference (predictions). Simulators are based on the same representational mechanisms that are used in implicit memory, filling in, anticipation and interpretation. The generic simulation scheme (see also 2.2.3) strongly coheres with Barsalou's notion. He uses the term 'simulation' in a very similar manner for the actual (real-time) process of staged elements, activities (he terms it 'entities and events') and causal relations. He adds the term 'simulator' for explicitly denoting the case in which a simulation is *not* parameterized and activated. This stresses the generative character of simulation. The feature that distinguishes simulators from mental models is exactly that generative character that can be compared to "operational rules" in the generic simulation scheme. Mental models can be understood as single states of mental simulations. The "family of related simulations" can be understood as models with different parameterizations and the "generative mechanism" as the operational rules that define how the parameterizations have to be performed.

5. *Perceptual symbols can be integrated in frames. Frames and simulations constitute a simulator. Schemata (Minsky M. 1975; Rumelhart & Ortony 1977) and scripts (Schank R. & Abelson R. 1977) are similar to frames. A frame (schema) comprises predicates, attribute-value bindings ('parameterization'), constraints and recursion (Barsalou 1993). Specific simulations (mental models) are generated via a constraint satisfaction process and, hence, represent the strongest (actual) attractor in a frame's state space. Event concepts are simulated by recursing this process. Simulations allow not only for retrieval of entities and events but also for transformations on them.* This property is particularly interesting because it provides a simple approach for dynamics: temporal order is conceived as a recursion of a constraint satisfaction process. An open question is, however, how the rule according to which the simulation is processed can be conceived, i.e. what temporal logic is.

It seems that Barsalou uses frames for denoting the representation of elements and activities as predicates + attribute-value bindings. (The generic simulation scheme used mental models for this purpose.) A frame already contains most specifications necessary for setting up a simulation, i.e. constraints + recursion. It is left unclear what is actually added to a frame to obtain a simulator, i.e. what the simulation itself is?

Barsalou's notion can be summarized as:

> frame = predicates + attribute-value bindings + constraints + recursion
> simulator = simulation + frame
> simulation = mental model = ?

It is actually unclear whether a simulation is less specific than a mental model in Barsalou's view. It is clear that a simulation is more specific in the generic simulation scheme:

> mental model = elements + activities + causal relations
> simulator = mental model + operational rules - activation (e.g. as memory structure)
> simulation = mental model + operational rules + activation

However, the notion of a simulation as an attractor in a frame's (mental model's) state space reconciles both notions. The slight differences are to be further discussed in other contexts.

6. *Perceptual symbols of speech and audition can help to develop simulators for words in memory. These can control simulations, if they are linked to simulators of concepts. In this manner linguistic indexing and control is realized.* This view corresponds nicely to the here-proposed role of lingual aspects as specific cases of a more general cognitive mechanism.

The discussion of these six core properties of a conceptual system already shows that the generic simulation scheme can be connected to Barsalou's account. Beyond these six core properties, Barsalou demonstrates derived properties that characterize a fully functional conceptual system. As before, paraphrases will be shown *in italics* and will be related to the generic simulation scheme as plain text added afterwards.

A. *'Productivity' refers to the Chomskian notion of systems' capability to generate, principally, an infinite number of conceptual structures. Perceptual symbol systems implement productivity through combinatorial and recursive processing of simulators. Production can be conceived as the reversal of the symbol formation process. It allows for transcending experience, especially in imagination (e.g. the "Cheshire cat"), although productive, perceptual symbol systems are constrained (as opposed to amodal symbols). For example, a schematic perceptual symbol (e.g. running) can hardly be applied* *to a simulated entity (e.g. a watermelon), because it lacks critical characteristics (e.g. leg). Humans can control productivity by linguistics that enables them to communicate, i.e. share simulations of non-experienced entities and events.* Productivity is a prerequisite of many processes in the generic simulation scheme (see also 2.2.3): particularly, the realization of the cognitive and, consequently, artificial momentum necessitates such generative powers.

B. *Propositions refer to the ability of a conceptual system to construct a given situation in an infinite number of ways by an infinite number of propositions.* The consideration of propositions provides means to establish consistency between perceptual symbol systems and other theoretic approaches such as schema theory, causal mental models. For the generic simulation scheme, this was already demonstrated in the context of causal mental models (see also 3.1.2).

C. *'Variable embodiment' is the idea that a symbol's meaning reflects the physical system in which it is represented. This property is not present in amodal symbol systems because it is assumed that they are represented independently from the concrete implementation. This 'disembodiment' is typical for functionalism. This concept provides a source for explaining both intra- and inter-individual conceptual variability and conceptual stability.* Variable embodiment is reflected, for example, in the generic simulation scheme by the clear differentiation between the natural, artificial or cognitive domain in which the models are operated. Disembodiment is prevented in that it is assumed that there is no representation 'outside' of these domains.

D. '*Abstract Concepts' can be represented by the three mechanisms of (i) 'framing' as representation in the context of a larger body of temporarily extended knowledge, (ii) 'selectivity' as highlighting the core content against the event background and (iii) 'introspective symbols' as extraction from internal experience.* Abstraction is applied in the generic simulation scheme for realizing the complexity reduction, i.e. freeing 'rich' representations (for example a natural situation with a free flying fly) from unnecessary background activity (for example by designing a reduced experimental situation).

The six core properties and the four derived properties of perceptual symbol systems show that all premises for being regarded as a conceptual system are given. The attempt of relating the generic simulation scheme to perceptual symbol systems reveals no major inconsistencies and illustrates how simulation can be conceived as cognition. Next, the actual operation of a mental model will be described in cognitive terms. As before, this shall be done along the lines of perceptual symbol systems (*italics* for paraphrases of Barsalou's account and added plain text for comments).

3.1.4. Simulation in action

In the generic simulation scheme (see also 2.2.3), simulation refers to an operated model. For example, mental simulation is given when a cognitive system has a mental model of a situation, e.g. a moving twig perceived during a walk adds possible causes for motion (e.g. wind, cat and monster) and applies operational rules to the model (e.g. wind with a given velocity). The result is a mental simulation of a moved twig. The simulation serves the function of generating an explanation. In the moving twig example, this could be an answer to the question what actually moved the twig? Along the lines of Barsalou's perceptual symbol systems it will be probed whether the generic simulation scheme can be described in cognitive terms.

Working memory is the system that runs perceptual simulations. Attention extracts schematic perceptual symbols. Automatic processing is the running of a highly compiled simulation, whereas strategic processing is the construction of a novel simulation using productive mechanisms. Skill results from compiling simulations for most of the plans in a domain through extensive experience (see Anderson 1993; Logan 1988; Newell 1990). Working memory corresponds to short-term memory in the information processing approach (see also 1.3.6). In the generic simulation scheme, short-term memory was also assumed to be the stage where simulations are operated (see also 2.2.3). Attention was seen as the selective mechanism that provides elements and activities from the ongoing perceptual activity. The poles of automaticity (automatic vs. strategic) can be regarded as the state of a learning process. Automatic processing applies learnt mental models, while strategic processing is a generative mechanism. The role of computer simulations is to mediate between these poles by doing some of the automatic processing, but leaving enough space for performing strategic processing to be performed in order to learn the underlying model. Skill refers to an automatic processing of simulations rather than models. It implies operational rules.

Long-term memory as well as categorization is best conceived as propositional construal. Memory retrieval is another form of perceptual simulation. It was pointed out here that elements, activities or causal relations could enter the ongoing simulation from memory (see also 2.2.2.1). A kind of propositional construal was also used in the account of causal mental simulation in the sense of Gentner and Gentner (1983) that was identified as the most appropriate cognitive theory for the notion of models applied in the generic simulation scheme (see also 2.2.3).

Concepts arise from the ability to simulate an entity or event perceptually. In the generic simulation scheme, concepts are important to complexity reduction because they enable the organism 'not to bother' about the details of a system. In this sense, a concept has no perceptual presence, but is characterized by an implicit character and automatic processing. A neuron, for example can be 'unfolded' from a simple binary element to a highly complex system (see also 1.2.4). An expert, however, can reduce a complex system and process it as a singular entity, simultaneously being able to detect references from other aspects of the situation (e.g. a presynaptic structure) to details of the model (e.g. a specific membrane channel).

Language comprehension can be viewed as the construction of a perceptual simulation to represent the meaning of an utterance or text. Perceptual simulation offers a natural account of how people construct the meanings of texts, or what other researchers have called situation models and mental models (see e.g. Johnson-Laird 1983). This function of simulation was already discussed in more detail (see also 3.1.2). Particularly, the account of Hesslow (Hesslow 1994; 2002) provides an idea of the lingual aspects of mental simulation.

Problem solving is the process of constructing a perceptual simulation that leads from an initial state to a goal state. Decision making can be viewed as specializing a simulated plan in different ways to see which specialization produces the best outcome (cf. the simulation heuristic of Kahneman & Tversky 1982). Similarly to Barsalou's description of problem solving, the generic simulation scheme (see also 2.2.3) used simulation as leading from question to answer. Decision-making corresponds to the evaluation stage. However, decision making often has an explicit character – it might better be conceived as a post-hoc construction of an inner perspective that results from recurrent networks with attractor dynamics achieving relaxation states (cf. Cruse 2003).

Formal symbol manipulation in logic and mathematics becomes possible through the simulation of arbitrary symbols. From perceptual experience with external symbols and operations, the ability to construct analogous simulations internally develops. In the example of the neuroscientist applying a logistic function as a characteristic of a neuron, formal symbol manipulation was described in terms of the generic simulation scheme (see also 2.2.3). It is worth adding that people often construct non-formal simulations to solve formal problems. For example, mathematicians, logicians, and scientists often construct visual simulations to discover and understand formalisms (see e.g. Thagard 1992). Non-academics similarly use non-formal simulations to process formalisms (see e.g. Bassok 1997).

Even though perception, imagery and cognition are neither identical behaviorally nor neurally, they "share representational mechanisms to a considerable extent." In terms of the simulation scheme, the difference between perception on the one hand and cognition as well as imagery on the other hand is simply its source, i.e. externally driven or internally driven. All lead to representations of elements, activities and causal relations that are operated in a simulation.

The implementation of artificial intelligence is possible, but peripheral devices of computers are very different from natural peripheries ('bodies') at the moment. Therefore, it should not be expected that a technical perceptual symbol system would be similar to humans. Similarly, it was concluded from the generic simulation scheme that there is no principal difference between natural and artificial models. Since both are mentally simulated to be 'realized', successful implementation is not an ontological but an epistemological question.

Whereas photos and videos only capture information holistically, a perceptual symbol system extracts particular parts of images schematically and integrates them into simulators. Once simulators exist, they can be combined *to construct simulations productively. Such abilities go far beyond the recording abilities of photos and videos.* This view corresponds well with the differences between computer simulations and other media (see also 2.1). As opposed to other media, simulations, no matter whether artificial or natural, have inherent functionality and can be operated with various configurations.

Cognition is continuous with perception in that bottom-up and top-down information can be merged. (Fodor (1975) claimed that perception is impenetrable by cognition because of its modular nature.) Sensorimotor systems are penetrable but not always. When bottom-up information conflicts with top-down information, the former usually dominates. When bottom-up information is absent, however, top-down information penetrates, as in mental imagery. Perhaps most critically, when bottom-up and top-down information are compatible, top-down processing again penetrates, but in subtle manners that complement bottom-up processing. This statement is interesting with respect to possible formulations of the function of computer simulations as carriers of an artificial momentum ('inherent functionality'). Consider first a pure thought experiment as a special case of imagery. In this case, the bottom-up input, i.e. perception of external entities and activities is 'missing' and, following Barsalou's notion, the probability of top-down information that interferes in the simulation is increased. Top-down information can overtake control if the user does not "get a grip" on the simulation. At this instance, computer simulations aid the learning process by providing a perception-driven precompiled sensorimotor cycle that decreases the probability of interfering top-down information. Computer simulations can provide this grip as an external force imposed on the elements and activities in the ongoing mental simulation. On the other hand, if the computer simulation (sensory input) is not 'compatible' with cognition, computer simulation will fail because no stable connection between sensory input and cognition can be established. In terms of the generic simulation scheme, this is the case when the representation function between the artificial preparation and the mental model of the artificial preparation cannot be setup because the models do not commute.

In sum, no major contradictions between Barsalou's account and the generic simulation scheme could be detected. On the contrary, Barsalou's account provides a rich source of concepts concerning the cognitive theory of simulation. This indicates that the generic simulation scheme can be conceived as a genuine cognitive mechanism.

3.1.5. Summary and conclusion

General background of this section is a conception of 'explaining by simulation'. Natural explanation, i.e. the notion of an explainer actually performing an explanation is envisaged here (as opposed to logical explanation, i.e. the notion of an explanation being something independent of an explainer). Since natural explanation is primarily a cognitive phenomenon, the issue of 'explaining by simulation' is reduced to the task of describing simulation as cognition. It was probed whether the generic simulation scheme coheres with existing cognitive theories. The information processing approach (see also 1.3.6) served as a starting point. Its capability of bearing the notion of simulation scheme can be seen as a first sign of consilience with existing cognitive theories. A review of cognitive theories that account for the characteristic properties of simulations, namely representing dynamic and complex content and showing inherent functionality, showed that candidates sufficing these conditions can be found. Particularly causal mental models as characterized by Gentner and Stevens (Gentner & Stevens 1983) account for the notion of models used in the generic simulation scheme[24].

What is left somewhat unclear by causal mental models is how representations (i.e. elements, activities and causal relations) can be conceived as being part of a natural explanation, for example, which aspect of an explanation is natural, which is cognitive or artificial? Causal mental models lack a concrete specification of the 'nature' of tokens contained by the models because they are primarily propositional (see also 3.1.2). However, the approaches to naturalized representations as proposed by Bechtel (2001b) and Barsalou (1999) allow to integrate models in the generic simulation scheme in the natural, cognitive and artificial domain. Therewith, it becomes principally possible to see explanations as a natural phenomenon

[24] It should be noted, though, that most of the theories of knowledge and representation described above fail to put the issue of dynamics in concrete terms. Frequently, dynamics are conceived as state or event sequences (see e.g. Barsalou 1999). This is a good for the beginning. But more global ('higher order') differentiated temporal structures like velocity, acceleration (derivatives) or integrated structures like duration (integrals) are not explained by simply gluing together states. Thus, future work should provide a deeper notion about how they are represented (explicitly or implicitly): we know how long it takes to get to work; we can imagine a car slowing down etc. Detailed accounts should also explain different perspectives on time, e.g. time-moving vs. ego moving (Gentner 2001).

and – as a side effect – to account for simulations as brain processes (Jeannerod 1994; see 2001 for compelling evidences). Not only maintaining models, but also operating them in order to generate the inherent functionality of genuine simulations calls for a comprehensive theory of a fully functional conceptual (cognitive) system that goes beyond plain modeling. Barsalou's approach of perceptual symbol systems provides both a fully functional conceptual system and a treatment of the issue of simulation. Therefore, it served as a pattern for an exemplary analysis in which it was probed whether or not the generic simulation scheme can be described in cognitive terms. The analysis showed that perceptual symbol systems show large correspondences to the generic simulation scheme. In conclusion, simulation can be described as a cognitive mechanism that produces explanations.

Reading Advice

The previous section showed that the generic simulation scheme holds as a general explanatory mechanism because it can be described in terms of general cognitive phenomena. But the gap between a general account and a concrete case can be deep. Whether or not simulation is the adequate approach to explanation requires a case-to-case evaluation. The following "Explanation-Special" demonstrates how such an evaluation could be performed (empirically) for a single case of "Explaining Brains by Simulation".

3.2. Explaining brains by simulation: Design for a case study

The issue of ‘explaining by simulation’ was introduced in the second chapter and extended in the first section of this third chapter. Simulation is described as a cognitive mechanism that generates explanations. Simulations as explanatory tools are necessary to account for complex and dynamic systems such as brains. Generally, the issue of this study ‘Explaining Brains by Simulation’ is a cognitive issue. However, the approach to simulation proposed here is not to be understood as a complete subjectivistic or even solipsistic account of science and learning – it does not describe cognition as an isolated domain, but embeds simulation in a situation in which it is applied: the ‘natural’ explanation that is analyzed here has an explainer and, additionally, a concrete (situated) explanation most often has a recipient. Thus, besides describing simulation as a cognitive issue, it is also characterized as a concrete situation. There are many different concrete cases of explainer-recipient situations: a lecturer explains for students, students explain to other students during learning for an exam, students explain for lecturers in exams, an assistant explains to another assistant during five o’clock tea, a scientist explains to colleagues in a talk etc. The general role simulation plays in all these situations is described by the generic simulation scheme. But the specific role simulation plays in all these different explanatory situations can be very different. So far, a coarse framework of understanding the role simulation plays has been developed. The actual variety of specific situations is not yet covered.

How can the variety of specific situations be covered? One possibility is the discussion of various explanatory situations in which simulations are applied. Though a valuable alternative, this would be a task for years and could be realized only unsatisfactorily superficial in this study. What then? I will provide a design for a case study on an empirical approach to “Explaining by Simulation“. This can be presented with sufficient detail and can mediate a concrete impression of what it means to analyze and evaluate a single explanatory situation. This concrete impression can then be hypothetically transferred (extrapolated) to the variety of possible explanatory situations that rest on simulation. The cognitive theories reviewed above (see also 3.1.1) already demonstrate applicable empirical approaches to an evaluation of cognitive mechanisms. However, the underlying studies of these cognitive theories do not address directly the issue of brains. For being neatly bound to the actual theme of this study, and bringing together the aspects of the previous chapters and sections, the case provided here will apply a model that relates to brains and, thus, to the issue of "Explaining Brains by Simulation".

Which model that reflects brain issues can be chosen? Models in the Brain Sciences come from such different fields such as Neurophysiology, neural network theory or cognitive psychology (see also 1.3.2). A common scheme prevalent in the brain sciences was identified in the analysis of the state of theoretical integration in the Brain Sciences as being an information-processing creature showing adaptive behavior (see also 1.3.6). A Braitenberg vehicle (Braitenberg 1984) is a possible model for this common scheme. Thus, the exemplary case study on "Explaining Brains by Simulation" applies a simulation of the Braitenberg vehicle. The role simulation plays in brain science was identified as improving the handling of dynamics and complexity in order to gain control (see also 1.1.11). Therefore, the task in the case study is to control different Braitenberg vehicles of varying complexity and dynamics. The role attributed to computer simulations was to provide some of the system's momentum (inherent functionality) in order to show the user how the model can be mentally simulated. This role of computer simulations is tested in the case study by comparing an experimental group that receives an instructional computer simulation with a control group that receives no such computer simulation and must work exclusively with thought experiments.

3.2.1. Introduction

The world is full of things that we do not perfectly understand: refrigerators, flower buds, computers, brains and the like. Nevertheless, we can handle situations in which these things play a role. Thus, it is not necessarily a problem that we do not see through them. But certain situations, particularly malfunctions, force us to understand the underlying mechanisms in order to predict their behavior – at least this is true for certain people such as the fridge repair service, the orchid grower, the network admin or the scientist. Things that we do not perfectly understand are often complex and dynamic: we have problems arranging all the elements, activities and causal relations of the system under scrutiny. But such proper arrangements are necessary to comprehend the underlying mechanism. Experts who understand those systems must have undergone a learning process during which they acquired skills and knowledge: experts have learnt to accomplish these arrangements and are able to generate hypothesis about (mal-) function, predict and test. In an optimal case, experts are able to perform a mental simulation of the system under scrutiny and, therewith, are able to test every possible cause of malfunction 'in their head' to come up with diagnosis and decisions what to do next.

The theoretical question addressed here is how mental simulation can help to solve complex and dynamic problems. This shall be achieved by analyzing the learning process taking place when subjects interact with models on a computer. The specific objective is to design a prototypical experimental framework for testing hypotheses on processes involved in learning models showing considerable dynamics and complexity. Influences of specific presentations of the models on the learning process, e.g. different trainings or different instructional material shall be assessed. Thus, the practical objective readily implied is to assess the design of educational model simulations.

Since the experiments cannot be directly derived from a tried and tested paradigm for evaluating learning with model simulations, this work has the status of pilot tests. These pilot tests shall eventually lead to a conclusive experimental framework, which is approached in first steps by the studies presented here.

3.2.1.1. *General task*

For the sake of clarity (but at the risk of loosing suspense), a short overview of the task will be presented before the procedures are explained in detail in the theoretical and technical sections. One general goal in this work is to elicit situations in which comprehension of mechanisms (the ability of controlled mental simulation) makes a difference, i.e. is advantageous or hindering. How could such a difference look like? Obviously, experts (as compared to naive learners) should show superior performance in verbal description tasks referring to a given mechanism: they should be able to generate better explanations. But verbal description during the task would interfere with the cognitive processes involved and verbal description after the task is remote as a primarily subjectively reconstructed interpretation of the individual experience during the task? Consequently, a behavioral, primarily non-verbal task (a “skill”) was chosen as dependent variable, namely the skill of controlling a device.

The general task is to steer a vehicle in a given environment (a circular ‘arena’) from a specified starting position towards a specified target position. The vehicle can be guided towards its destination by placing a virtual stimulus in the arena. Subjects are told that the vehicle has to find nutrition and can be guided by a light source. Initially, the vehicle ‘waits’ in the middle of the arena (see fig. 24). The learners see the target (a blue circle). Then, they can place a stimulus by clicking on the arena’s rim. In the moment of the stimulus placement, the vehicle is activated and begins to move relative to the stimulus. The vehicle stops when it crosses the arena’s rim. Depending on the learners’ choice of stimulus position and the final architecture of the vehicle, the target can be hit or missed. Before the first condition of the experiment is presented, learners have two make two test trials.

Each condition is a set of twenty trials (‘runs’): during one condition twenty target positions of a fixed set are presented in a randomized order. When learners have finished the first condition, learners are informed that they might find another vehicle in the next set of trials. Five conditions are defined: the first two are conditions in which the vehicle moves directly towards the stimulus (positive phototactic), directly in opposite directions of the stimulus (negative phototactic). In the last two conditions, the vehicle shows skew trajectories that tend to miss the stimulus. Again the first is positive the second negative phototactic.

The third is the critical condition in which the control group makes a neutral task (tic-tac-toe) while the experimental group receives an instructional treatment on the internal organization of the vehicle (see also 3.2.1.1). The experimental group is informed that the vehicle is bilaterally symmetric and has two sensors and two motors and sensor-motor connections. The activity of the sensors depends on the stimulus position (bright vs. dark light) and determines the activation of the motors. The direction of activation (acceleration vs. deceleration) can cause positive or negative phototaxis and differential sensitivities of the sensor-motor connections on either side can cause skew trajectories. Learners are instructed to rebuild one of four predefined configurations of the vehicle and can observe the resulting trajectories. The sensor's activity during the run is dynamically visualized during the run. Learners of the experimental group, thus, have the opportunity to learn the vehicle's architecture and the influence of different configuration to different behaviors. They receive information necessary to conclude a vehicle's configuration from observations of the vehicle's runs in conditions '3' and '4' and predict trajectories for subsequent runs. The instructional section of the experimental group contains instructions of how a mental simulation of a specific problem can be performed.

There is a problem distinguishing between training effects resulting from steering the vehicle through the arena and effects of instructional unit. It is conceivable that the experimental group outperforms the control group because learners have more opportunities to learn about the problem of steering the vehicle, but not because they receive the instructional unit. This could be tested by having a second control group that has to steer another vehicle in the critical condition and compare it to the control group and the experimental group. However, for the sake of (conceptual and statistical) clarity, the preferred approach is to quantify the overall effect of the instructional unit first and dissect this expected overall effect in further studies.

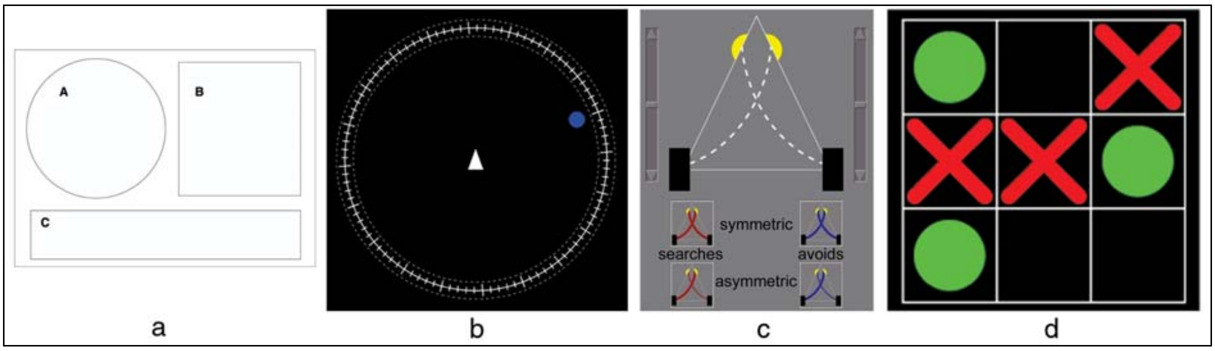


*fig. 24: **The task.** **(a)** Schematic presentation of the GUI. A is the 'arena' where the task of steering a vehicle is presented; C contains instructions and feedback as text. B contains instructional material on the vehicle for the experimental group and task-independent information (tic-tac-toe) for the control-group. **(b)** 'Arena'. Learners have to lure the vehicle (triangle), with a stimulus (circle) towards a target (square). **(c)** Internal organization of the vehicle. This is presented exclusively in the experimental condition for the experimental group. Learners can choose between four different vehicle configurations and test them in the arena. Additionally, learners can change weights of the configurations. **(d)** Tic-tac-toe game. This is presented exclusively in the experimental condition for the control group.*

3.2.1.2. *Learning strategies*

The rationale in the experimental design is that the ability of mentally simulating the run before the decision of stimulus positioning takes place, helps to generate more exact predictions of a specific trajectory. These should be measurable as smaller distances between target position and the final position of the vehicle, i.e. smaller errors. The experimental treatment with the instructional unit on mechanisms acting in the vehicle shall provoke a clustering of learners that mentally simulate successfully and those who do not. Consequently, a difference in the performance between groups is predicted. However, as it is typical for performances on complex tasks various strategies for solving the tasks might be developed and applied by learners. More detailed, some learners of the experimental group might not mentally simulate and some of the control group might do so.

This problem can be explained by the assumption that mental simulation corresponds to the handling of complex and dynamic systems. The more elements and activities a system has and the more time steps are to be considered in which elements and activities can influence the overall behavior of the system, the more possibilities the learner has to 'navigate' through this problem space. All elements and activities might give rise to different elementary cognitive skills that finally act together to predict the behavior of a system. In this sense, mental simulation is a concerted action of elementary cognitive skills that happens before the decision of stimulus positioning. It is obvious that it is not trivial to predict exactly each performance of each learner on a complex and dynamic task. This is a fundamental problem of the analysis of mental simulation and performance on complex and dynamic situations, respectively. It should be noted that this problem of multiple learning strategies is a question of finding the correct experimental design – it is put forward here that each 'interesting' design (i.e. one that creates situations of considerable dynamics and complexity) will suffer from this problem. It would be naïve to assume that complex and dynamic situations can be investigated with simple designs.

As a consequence of the multiple learning strategies, it must be considered that, even if a difference between experimental and control group was found, it couldn't necessarily be referred to mental simulation. The expected variance necessitates a more detailed analysis of the different strategies to carry out the task – different learner models have to be specified.

3.2.1.3. *The model*

The vehicle's motion and, therewith, the resulting trajectory of a given run is based on a model of Braitenberg (Braitenberg 1984) that is implemented as an experimental computer program.[25] The vehicle has sensors and effectors ('motors'). The effectors move the vehicle, the sensors 'react' to certain events in the environment (see fig. 25). Effectors and sensors may have various properties (e.g. linear or dynamic characteristics) and may be interconnected in several ways (e.g. feed-forward or recurrent). By exploiting the combinatorial possibilities, small but considerably complex and dynamic vehicles can be designed. However, for the sake of simplicity of the experiments, very reduced architectures are considered in the pilot phase. Sensors are designed as visual elements that can detect light stimuli – even though other principles of stimulus detection (e.g. distance dependent as in the case of odor detection) are possible. The receptive field of the 'eyes' is monocular (no overlap between both hemispheres) and has a sine wave sensitivity function: it is maximally sensitive for stimuli being exactly lateral to the respective eye (90° relative to the symmetry axis of the vehicle) and become less sensitive to the front (0°) and the back (180°).

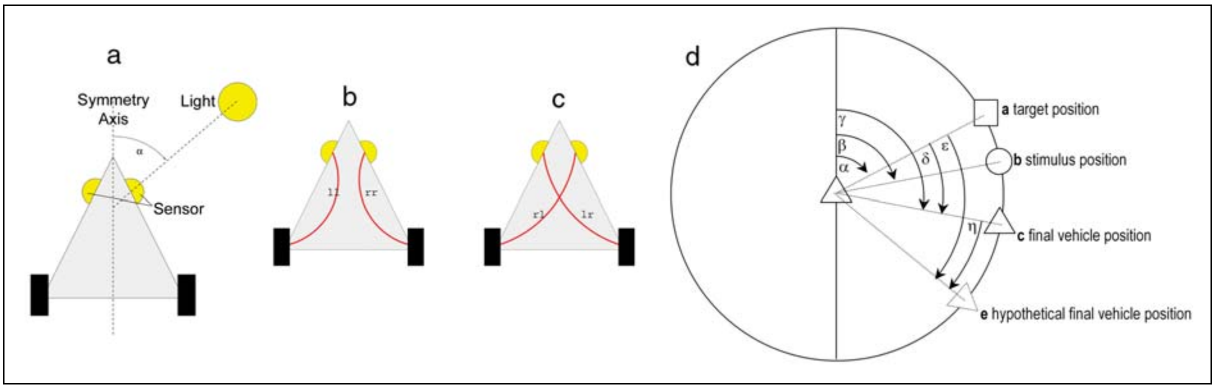


*fig. 25: **Model of the vehicle. (a)** The vehicle has no binocular vision but registers only information coming from one side at one instance of time. The sensitivity of the sensor varies with the A. **(b)** and **(c)** Types of connections inside the vehicle. In the experiment only the crossed type was applied (r: right, l: left). **(d)** Points and angles relevant for data analysis. All angles are measured from the symmetrical axis of the vehicle (0°). The hypothetical final vehicle position is that final position of a given vehicle measured when the stimulus is put on target position.*

[25] This program was developed by the student coworker Olaf Gerstung.

Generally, two feed forward connections are possible – one ipsilateral (connecting sensor to motor on the same side, i.e. right to right and left to left) and one contralateral (connecting sensor to motor on the other side, i.e. right to left and left to right). However, since the effective behavior does not principally change, only one connection type will be considered (i.e. the contralateral). The connection strength ('weights') between sensors and motors can be varied. By setting weights, differential propulsion on the motors on either side of the vehicle can be realized that results in different types of behavior. For example, positive contralateral connections result in vehicles moving towards the light source (positive phototactic), while negative contralateral connections cause the vehicle to steer clear of the light (negative phototactic). Asymmetric weights, e.g. one connection weak (0.1) and the other strong (0.9) will result in skew trajectories.

Note that the vehicle's behavior is not explainable as a simple stimulus-response relation. Information processing in the vehicle is continuous over the whole run so that the sensory input will change with each bit the vehicle moves. Thus the vehicle tends to produce curved trajectories instead of running directly towards or away from the light source. More precisely, motors cause constant propulsion from the onset of stimulation with the light source to the moment the vehicle reaches the arena's rim. If all connections were maximal and no sensory input were present the vehicle would move, say 5 pixels per instant of time of the model simulation. If the arena had 200 pixels, so the vehicle would receive 40 sensory inputs before it reaches the arena's rim and would stops after, say 4 seconds. If sensory inputs were continuously maximal, the vehicle would move, say 10 pixels per instant of time and reach the rim after 20 steps (2 seconds). If only one sensory input were maximally activated, the vehicle would move towards the side where the activation ('the light') is, because the propulsion of the contralateral side is increased. (Alternatively, the same behavior can be realized by decreasing propulsion on the ipsilateral side with negative connections and ipsilateral connectivity). A directional turn (180°) is performed after, say, 5 steps (approximately 0.5 second), which would correspond to an angular velocity of 360 deg/sec. The tuning of the vehicle's movement is a trade-off between smoothness (not too slow) and transparency of the movement: learners should be able to observe that the actual sensory input continuously determines the turning direction, i.e. the trajectory can change gradually during the run.

3.2.1.4. *Learning the vehicle*

Learners are not informed about the configuration of the vehicle in a given condition. By observing the vehicle's behavior, they have to learn to control the vehicle. Thus, they have to conclude from their observations of the vehicle's behavior to the 'mechanism' that controls the vehicle, e.g. "the vehicle is scared away by light". Learners are informed when they are confronted with a new vehicle (a new condition). The performance measure for the success of learning is the "goodness of steering". This is defined as the reversal of the error (angle), i.e. the deviance of the vehicle from the target (dependent variable). The prediction following from these assumptions is that the error will decrease over runs and a learning curve will be measurable.

More precisely, the arena's top (along the symmetry axis of the vehicle's starting position) is defined as zero degree, angles are given in radians [-π, π], clockwise angles are positive, counterclockwise are negative. α is the angle between top and target position *a*, β is the angle between zero and stimulus position *b* (see fig. 25a) and γ the angle between top and the vehicle's final position *c*, i.e. where it's tip touches the arena's rim. The normalized error δ is determined by the distance between α and γ (see fig. 25d), the inner angle between the points *a* and *c* is δ [0,1].[26]

3.2.1.5. *Error correction*

However, not all targets are equally easy to be hit. This poses a problem for computing the learning curve over several runs. If target positions are presented in randomized order, variance of observed errors averaged over learners might not be caused by the degree of comprehension of the mechanism that controls the vehicle (as intended), but by the varying difficulty of the target position. On the other hand, if target positions are presented in a fixed order, specific learning strategies for a given run interfere in the results. Therefore, the error for a single target position should be weighted with its difficulty. But how can error correction be achieved? Assume that the circle is divided into *100* tics and a given target position counts as 'hit' if the vehicle's final position is within the tic. If only

[26] As will be described later, some targets are not reachable. Then, the nearest reachable point – depending on the end position of the vehicle – in left or right vicinity of the target is taken as point *a*.

one stimulus position caused the vehicle to hit the target, the probability would be one out of hundred ($p = 0.01$) if every possible stimulus position led to a hit, the probability would be *100%* ($p = 1$). Thus, the amount of possible stimulus positions causing hits indicates the chance level for hitting a given target position. Consequently, each target position in each vehicle configuration (each run '*i*') has its specific chance level $p_i = n_{hit} / n_{tics}$ (n_{hit} is the number of hitting stimulus positions, n_{tics} the number of tics). The chance level in a given condition is determined by the mean individual chance level over the number of trials n_i: $p_{mean} = \Sigma(p_i) / n_i$. This individual chance value can be applied as a weight for each stimulus position when assessing the error that learners make for a single run.

$$E = \delta / (1 - p)$$

It should be noted that this conception of the error implies no assumption of learning strategies – it just reflects the statistical prerequisites. This normalized and weighted error allows for the consideration of learning curves during the runs averaged over learners for a specific vehicle when target positions are presented in a randomized order.

3.2.1.6. *Distinguishing strategies*

Different vehicles are used to introduce different conditions (each 20 runs). Some conditions might very well be performed without understanding the control mechanism of the vehicle, while others might be better performed when comprehension has taken place. For instance, some vehicles might be successfully controlled with simple strategies, such as: "The vehicle hits the target when I place the stimulus on the target position." Other vehicles require more sophisticated strategies, such as: "The vehicle hits the target when the stimulus is placed slightly on the right (seen from the perspective of the approaching vehicle)." and/or "...the lower the target the greater the required shift of the stimulus to the right."

Which strategies can be distinguished? An initial question might be that the vehicles' behavior is either positive or negative phototactic. But a learner having experienced both types of vehicles has to take into account both possibilities. Compared to a naive learner, the problem space, i.e. the number of possible positions n_{tics} is doubled. For predicting the performance of a learner who does not know which vehicle is present the chance level consequently also has to be doubled. In the actual experiment, learners are instructed that one vehicle is present and will only be changed if they are informed. Thus, they will assume that they have the same vehicle during successive runs. Nonetheless, when they are confronted with a new vehicle (and informed about that), they initially have an enlarged problem space. After a certain amount of trials, learners might be able to assume a given mechanism for the vehicle, i.e. if it is positive or negative phototactic. This assumption bisects problem space again. This learning process should be accompanied by a drop of the error production from the level determined by the possible occurrence of positive and negative phototactic vehicles to the level determined by the possible occurrence of only one vehicle type. This should be visible in learning curves as an initial drop of error (boost of performance).

If the learner experiences quantitative deviations ('skew vehicles'), much more trajectories are possible and the problem space becomes larger. One approach to predicting learner performance is to compute problem space with all possible positions. But it seems implausible that actual learners will take into account all possible positions, but rather bisect problem space for positive and negative phototactic vehicles and then try to figure out the direction of skewness before they finally try to determine skewness quantitatively. Consequently, a different approach of predicting performance by learner modeling is applied for skew vehicles: If the learners' error is corrected with a *difficulty factor* for each single target position the error levels predicted for straight vehicles should not change.

How can this difficulty factor be conceived? Consider the simple case that the vehicle runs directly to the light source and compare it to the situation in which the vehicle runs away from the stimulus and tends to hit the arena's rim slightly right to the stimulus. It can be assumed that the operation of predicting the shift is more difficult than predicting no shift and that difficulty is increased with increasing angular distance of the shift. Thus, the difficulty of a given target position is assumed to be proportional to the angle ϵ (normalized by $1/\pi$) between the target position a and the point e indicating the Stimulus-On-Target (SOT) error. The SOT error ϵ is the resulting error angle when the stimulus was set directly on the target position ($a = b$).

What can this SOT error ϵ tell about difficulty as subjects perceive it? It expresses how far a given vehicle would deviate from a given target position if learners assumed that the vehicle moved straight towards the target position when the stimulus is set on the target. If difficulties ranged between zero and one, the “straight vehicle” would be defined as having zero difficulty. This would imply that the difficulty is greatest if the vehicle lands straight on the other side. Of course, this measure of difficulty is consistent only for the assumption that the vehicle is positive phototactic. Thus, this measure of difficulty is sensitive for (and at the same time dependent on) a learning strategy. Negative phototactic vehicles have to be treated alternatively: learners have to assume that the stimulus position is directly opposite to the target position $a = | b - \pi |$. (A simple practical solution is to base all computational operations not on the target position a but on the mirrored target position $a' = | b - \pi |$, *ceteris paribus*.)

The SOT error can be used as a difficulty measure that allows for an assessment of learners’ performance in tasks with the same target position but different vehicles: Given are the vehicles V_1 and V_2 of different difficulty and learners that produce the errors V_1: $\delta_1 = y_1(V_1)$ and V_2: $\delta_2 = y_2(V_2)$. Now, it does not make much sense to compare these errors directly because it is clear that learners produce greater errors for more difficult vehicles. Now, the error angles ϵ_{V1} and ϵ_{V2} that result from the assumption of straight vehicles (steered by placing stimulus on target position $a = b$) indicate the difficulty of the vehicle. By correcting the observed error with the corresponding SOT error, the performance of the learner can be separated from the theoretical error. For example, if learners consistently performed on the level predicted by correcting the observed error δ by the SOT error ϵ, it would be indicated that they have learned how to control the vehicle. If they performed worse, on the other hand, it might be due to “additional confusion“. Put another way, if the performance (e.g. learning curve) on straight vehicles and the corrected performance on skew vehicles is not distinguishable, the learner successfully compensates the SOT error.

It should be noted that this measure primarily refers to single trials and not to the whole condition (vehicle) because the SOT error ϵ is not uniformly distributed over the arena, but varies with target position. Since the vehicle’s starting orientation is always ‘up’ (0 degree), targets in the back pose different problems than targets in the front. For example, the vehicle must turn to reach targets in the back. Depending on the actual configuration of the vehicle, the turn might increase the SOT error ϵ in back positions relative to front positions. Thus, it is not precise to compute an overall error for the vehicle simply by averaging ϵ over all possible positions. However, this is what learners might do – the mean vehicle skewness helps determining another level of learning: Given that the probability of target positions is homogenously distributed (e.g. each target position appears once or multiple integer times), it can be conceived simply as the mean SOT error ϵ_{mean} over all possible target positions.

$$\epsilon_{mean} = \Sigma\ \epsilon_i\ /\ n_i$$

The mean *vehicle shift* provides a means for assessing the learning process: the more the learners can 'move' away from the error level predicted by vehicle shift over trials, the more they leave their hypothesis of a "straight vehicle". In case they minimize the distance, they learn to compensate the mean shift of the actual vehicle. Increases in distance, on the other hand, would indicate that they are confused. Thus, a possible learning strategy would be to minimize $\eta = \delta - \epsilon$ that can be seen as an indicator of the amount of learning that has taken place (e.g. if a level of a steady state in error production is reached).

As already explained above, SOT errors are not distributed uniformly over the arena, but vary with position. Learners performing on the level predicted by a mean vehicle shift still make considerable errors: They will sometimes overestimate and sometimes underestimate the shift. But learners who correctly anticipate not only the direction of the shift and the mean magnitude of the shift, but also the distribution of specific shift magnitudes produce an even smaller error than predicted by the mean vehicle shift – optimally zero. (A jitter resulting from the imprecise positioning of the stimulus and a residual error is expected, though). For a comparison of performances on straight and skew vehicles (e.g. a comparison of learning curves) the predicted performance is not simply determined by mean vehicle skewness, but demands a trial based correction with the SOT error ϵ. Then these learners should perform exactly as if they control a straight vehicle because they perfectly anticipate the shift.

In sum, the following problems for learners can be differentiated with the help of the error conceptions introduced above:

1. The task: “steer vehicle to target“.
2. The control principle: “use stimulus to direct vehicle to target“.

   >>*Prediction A* (for 1. and 2.): Learners who do not understand 1. and 2. perform on the chance level. A corresponding learner model positions stimuli randomly.

3. The stimulus quality / vehicle preference: “positive and negative effects of the stimulus must be considered“.

   >>*Prediction B*: Due to bisection of problem space the error should decrease instantaneously after vehicle preference is understood.

4. The shift direction: “vehicle misses the target on a specific side when stimulus is set on target“
5. The shift compensation: “shift can be compensated with a corresponding shift of stimulus position“
6. The shift magnitude: “the (compensatory) shift is about *x* tics“

   >>*Prediction C* (for 4., 5. and 6.): learners who compensate the mean vehicle shift perform worse than on a straight vehicle. Actually, individual learners with different error sources (corresponding to steps 4., 5. and 6., respectively) are summarized in this model.

7. The shift distribution: “the (compensatory) shift is about x_1 tics when target in front of the vehicle, x_2 tics when target on right side, x_3 tics when target on left side and x_4 tics when target is in the back“.

   >>*Prediction D*: learners who compensate the skewness of a given vehicle for each target position perform as if the vehicle was straight. (The difficulties introduced by shift distribution can be subdivided into several subclasses. For example, learners might or might not distinguish between different shift distributions on the left and on the right. However, these differences are assumed to be too fine-grained for the present state of the analysis.)

Generally, all problems (1-7) in the task have to be overcome by learners in order to control a difficult (e.g. 'skew') vehicle. Having analyzed these problems encountered by all learners, no matter whether experimental group or control group, we arrive at a stage where experimental and control group can be compared. Most problems encountered by learners might be overcome just by observing, analyzing and evaluating the vehicle's behavior - be it by trial-and-error or by goal directed hypothesis testing. Thus, it is not mandatory for learners to hypothesize about the mechanisms underlying the vehicle and, therewith, it is not mandatory to receive the instructional treatment of the experimental group: it is also possible to detect the correlation between stimulus and response by phenomenological analysis. The control group, hence, has a chance of competing the experimental group. But the global hypothesis behind the experimental design is that skew vehicles (presented after the experimental treatment) are hardly predictable without mentally simulating them. Consequently, learners of the experimental group who have received instructions for mental simulation are predicted to outperform other learners. This should be measurable as differential error niveaux produced by the two groups. Following the assumption that mental simulation helps particularly in difficult situations (i.e. situations showing dynamics and complexity) 'skew' vehicles should particularly be better steered by the experimental group.

3.2.2. Material and methods

3.2.2.1. *Experimental procedure*

A computer (Pentium II, 256 MB RAM or comparable), a mouse and a JAVA-program (JRE 1.3) are required. Learners sit in front of a monitor and use a mouse with at least one switch as exclusive steering (input) device. Monitored by the program are mouse clicks (onset, offset, duration, *x-y* position). It is possible to refer all monitored user data to events in the runtime of the application, i.e. data referring to the same 'master time' in the application.

Relevant information is presented via a graphical user interface (GUI) 1024x768 in size on an otherwise black screen. The GUI has the sections A, B, and C (see fig. 24).

A-Task field. This is the 'arena' in which the user has to place a stimulus in order to lure the vehicle towards the target (see fig. 24b)
B-Experimental field. Different tasks are presented for experimental and control group. The experimental group receives instructional material on the organization of the vehicle, the control group a tic-tac-toe game.
C-Text field. The textual information provided in this field may be instruction or feedback (i.e. if a condition is finished etc.)

The experiment is organized into four phases, (i) introduction, (ii) 'pre', (iii) critical (experimental vs. control group) and (iv) 'post' (see table 8 for a detailed description of the phases). The introduction contains a textual instruction and several trials for luring the vehicle towards a target by placing a stimulus on the arena's rim. The vehicle used in the introduction is the simplest possible: if the stimulus is placed on the target, the vehicle always hits the target, i.e. the Stimulus-On-Target (SOT-) strategy is successful. After the introduction, learners should have understood the general principle of the task. Only if the learner has performed two successful trials, the program enters the next mode.

The phases 'pre' and 'post' show the same organization: each has two conditions (i.e. 'pre1', 'pre2', 'post1', 'post2'). Each condition comprises 20 trials. In each trial, one of a fixed set of 20 target positions that are uniformly distributed over the circle is presented. The program randomizes the order of presentation in each experiment to prevent the influence of a specific sequence of target positions. The conditions vary in the vehicle's configuration that, in turn, determines the vehicle's behavior. (Between conditions learners are informed that a new vehicle is presented). Even though vehicles show curved trajectories, the vehicles were tuned to hit the arena's rim before they turned 360°. Vehicle 'pre1' moves towards the light ('positive phototactic') and shows the same (mirror-symmetrical) behavior on its left and right side ('symmetrical'). Vehicle 'pre2' moves away from the light (negative phototactic), but is also symmetrical. Vehicle 'post1' is positive phototactic but asymmetrical: when users place the stimulus on the target, the vehicle will miss the target with different absolute error angles on the left and on the right side ('skew vehicle'). Vehicle 'post2' is skew and negative phototactic.

***table 8: Schematic overview of the task.** First, instructions for experimenters before the task are shown, then the overview of the actual experiment. Column 1 shows the name of the phase (prog.=program), 2 the vehicle type (see text for details) presented on screen A, 3 the experimental/control treatment on screen B, 4 the instructions (translated from German), 5 the way of progression in the program and 6 the number of iterations. Finally, the questionnaire applied after the task is shown.*

Preparation

Experimenter opens program, chooses 'E' for experimental group or 'C' for control group enters Learner-ID and confirms by pressing enter.
Experimenter welcomes learner and asks learner to sit down.
Experimenter informs learner that anything else will be explained on the computer screen.

Experiment

Prog. Part	Screen A	Screen B	Screen C	Progression	#
Intro	"intro"	–	*You see an artificial being in the middle of an arena. Without help, the being (triangle) is not able to get to its feed (blue circle). But by placing a light on the arena's rim, you can lure it towards its feed. Just click where you think the light should be!*	forward only when target was hit	1
	–	–	*Great! Now comes the next feed!*	auto-forward 3 sec.	
	"intro"	–	–	forward only when target was hit	1
Pause	–	–	–	auto-forward 3 sec.	1
Part 1	"pre1"	–	*Let's get started! Notice, however that the being is exchanged for the following 20 runs. It might behave different from the one of the test run.*	forward after first stimulus positioning	1
	"pre1"	–	–		19
	"pre1"	–	*Next run.*	forward on click	
Pause	–	–	–	auto-forward 3 sec.	1
Part 2	"pre2"	–	*For the next 20 runs, you get another being again. It might behave different again.*	forward after first stimulus positioning	1
	"pre2"	–	–		19
	"pre2"	–	*Next run.*	forward on click	
Pause	–	–	–	auto-forward 3 sec.	1
Exp.	user settings	inter-active model	*Here, you can learn something about the functioning of the beings for better control of the beings presented later. Beings register light signals with their eyes and pass it over cables to the motors. Beings only differ in their cables: Light signals passed by red cables boost the motor, light signals passed by blue cables slow the motor down. Thick cables boost or slow down stronger than thin cables. In the model, you can exchange cable types with the slider. When your being is ready, place a light and observe the resulting behavior. Try it!*	forward after first stimulus positioning	1

	user settings	interactive model	*Now, you have 20 test runs (you can place a light 20 times). Build the beings shown in boxes 1-4 and make yourself clear the following characteristics:* *- Being 1 does not reach lights in the rear parts of the arena because it turns too slowly.* *- Being 2 runs away from light because light on the right side, for example, slows down the left motor and the now stronger right motor turns the being to the left.* *- Being 3 reaches almost any light in the right half but not in the left half because the cable from the right eye is too thin for boosting the turn sufficiently.* *- Being 4 runs away from light, but makes better turns to the left because light coming from the right slows down the left motor too much for a sufficient turn.*	forward on click	19
Contr.	-	Tic Tac Toe	*The game TicTacToe is about placing 3 tokens in straight or crosswise line. The computer program places its token promptly after yours. Then it's your turn again. In the next game the computer program begins. You play 20 games.*	forward on click	1
	-	Tic Tac Toe	-	After game over	20
	-	Tic Tac Toe	*[Game number]: [won / draw / lost]*	forward on click	
Pause	-	-	-	Auto-forward 3 sec.	1
Part 3	"post1"	-	*For the next 20 runs, you get another being again. It might behave different again.*	forward after first stimulus	1
	"post1"	-	-		19
	"post1"	-	*Next run.*	forward on click	
Pause	-	-	-	Auto-forward 3 sec.	1
Part 4	"post2"	-	*For the next 20 runs, you get another being again. It might behave different again.*	forward after first stimulus	1
	"post2"	-	-		19
	"post2"	-	*Next run.*	forward on click	
End	-	-	*Thanks.*		

Questionnaire

Item	Options	to be filled by …
Name	first name, last name	learner
Age	integer	learner
Gender	[male | female]	learner
Profession	…	learner
Schooling	(for students, subject of study)	learner
Braitenberg Vehicles prior known	[yes | no]	learner
Other prior knowledge on task	…	learner
Tic Tac Toe	[yes | no]	experimenter
ID	…	experimenter
Date	…	experimenter

A single trial comprises the following sequential states: in the 'inactivated' state, it is not possible to place a stimulus. The activated state is entered automatically after a program internal downtime of the system and is characterized by 'waiting' for learner's choice of the stimulus position. The active state is entered when the learner places the stimulus: the vehicle begins to move relative to the stimulus position and stops on the arena's rim. The target can be hit or not. The program provides feedback to the learner in several forms. The target is 'hit' if the vehicle exceeds the arena's rim within a corridor of 5 % of the circumference of the arena (5 out of 100 tics). In successful cases the target changes its color from white to red. Moreover, learners see a counter that indicates how many targets were hit. This feedback has exclusively motivational reasons. By instructing learners to hit as many targets as possible, it is ensured that they focus on minimizing their error. The mere number of hits is, however, a measure too coarse (and statistically problematic) to evaluate performance of learners - the error angle is taken as measure instead.

The critical condition comprises a short introduction in textual form and the actual task. The control group is provided with information how to play Tic-Tac-Toe and plays 20 games as task-irrelevant treatment. The experimental group is provided with a short introduction on the organizational principles of the vehicle, the respective mechanisms and the information on their opportunity to perform 20 tests on a given vehicle configuration. In the actual task, they see a model of the vehicle with its most important components (sensors, motors, connections). They can vary the connection strengths of the left and right connection with sliders that are on the left and right, respectively. The upper end of the slider is 'full positive', the middle is zero and the lower end 'full negative'. The connection strength is visualized directly as a color code in the vehicle model: red for positive and blue for negative values, thick for strong connections and thin for weak connections (zero is visualized as a dotted line). Below the model are four switches with symbols of specific configurations. Characteristic are the differential colors of the connections. The symbols have a text label indicating the behavior of the vehicle (e.g. "positive phototactic, asymmetric"). Learners are advised in the instructional text to rebuild these models by changing the connection strengths. These pre-configured vehicles help learners to match a given configuration with a given behavior. The pre-configured vehicles correspond to the four vehicles presented in the different conditions during the experiment (e.g. ‘pre1’ is upper left symbol), but learners are not explicitly informed about that. The learner can start a test anytime. It does not matter whether configurations were changed by sliders or symbols or both or none. The number of already performed tests is indicated by a counter.

3.2.2.2. *Model implementation*[27]

The activation of the vehicle’s sensors depends on stimulus position. Only one sensor at a time can be activated because there is no geometrical overlap between the sensitivity functions of both sensors that could build a binocular visual field. The activation maximum for stimuli being directly on the left or on the right (90°). That means: a stimulus of the amplitude 1 causes the activation 1 when the stimulus is directly on the right or on the left relative to the middle of the vehicle. The sensitivity decreases to the front and to the back following a sine function and becomes zero directly on the symmetric axis. The stimulus amplitude I_{in} is constantly 1 and does not depend on distance. Thus the activation of a sensor at a given instant of time is defined by $I_S = \sin(I_{in})$.

3.2.2.3. *Propulsion of the vehicle*

Position and alignment of the vehicle is computed with a 4x4 transformation matrix. (The 3D approach was chosen because of easier implementation in JAVA.)

$$M_T = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{23} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix}$$

The upper 3x3 matrix with the constituent elements m_{ij}; $i,j \leq 3$ the rotation matrix:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} = \begin{pmatrix} \cos(\gamma) & \sin(\gamma) & 0 \\ -\sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

[27] *This was carried out by the student coworker Olaf Gerstung.*

The angle γ indicates the rotation relative to the z-axis. The values $m_{14} =: t_x$, $m_{24} =: t_x$ *and* $m_{34} =: t_x$ determine translation along x-, y- and z-axis, respectively. Since the vehicle moves on an x-y plane, m_{34} as well as m_{41}, m_{42} *and* m_{43} is zero. Thus M_T has the concrete form:

$$M_T = \begin{pmatrix} \cos(\gamma) & \sin(\gamma) & 0 & t_x \\ -\sin(\gamma) & \cos(\gamma) & 0 & t_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Global reference point is the starting position of the vehicle $p_0 = (x_0,y_0,z_0)^T$; $z_0 = 0$. The initial alignment of the vehicle is $\gamma_0 = 0°$ relative to symmetric axis of the vehicle. By combining p_0 and γ_0 to a vector $V = (p_0 \gamma_0)^T$ the actual position can be computed by multiplication of the matrices.

$$M_T = \begin{pmatrix} \cos(\gamma) & \sin(\gamma) & 0 & t_x \\ -\sin(\gamma) & \cos(\gamma) & 0 & t_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_0 \\ y_0 \\ 0 \\ \gamma_0 \end{pmatrix}$$

For computing a step of the vehicle, the angle $\Delta\gamma$ has to be determined by multiplying the maximal possible rotation γ_{max} with the given activation of the sensor weighted with the connection strength (weights) *rr*, *ll*, *rl* or *lr* (see fig. 25b, 25c): $\Delta\gamma = \gamma_{\max} (-ll \cdot S_L - lr \cdot S_R + rl \cdot S_L + rr \cdot S_R)$

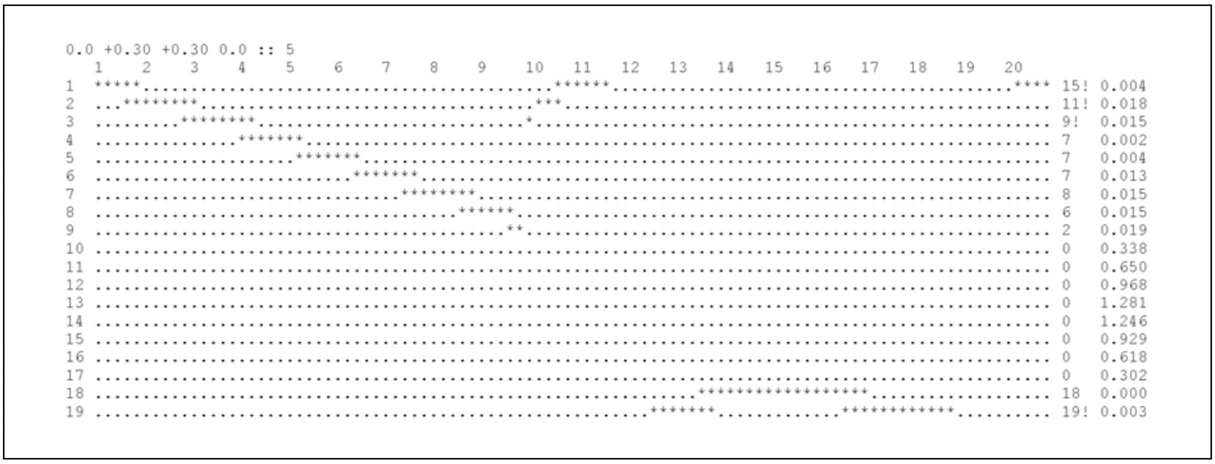
Translation t_x and t_y is first computed as a distance: $s = 1 - |\Delta\gamma|$ and then separated in the corresponding angles by:

$\Delta t_x = -s \cdot \sin(\gamma)$, $\Delta t_y = -s \cdot \cos(\gamma)$

The program registers and stores the values α (target position), β (stimulus position), γ (end position) and the trajectory length. The reference line for the angles is a vertical line through the middle of the arena. Clockwise angles are positive, counterclockwise negative and are in the interval $[-\pi, \pi]$. Trajectory length is defined relative to the shortest possible trajectory. The connection strengths *ll, rr, lr*, and *lr* (*r*: right, *l*: left, see also fig. 25b, 25c) between sensors and motors are those parameters that determine the different conditions (see also 3.2.1.1). Only the connections *rl* and *lr* are actually applied in the experiment. If both sides are equal, the configuration is symmetric, if they are unequal it is called asymmetric.

3.2.2.4. *Model simulation*[28]

Assessing the performance of learners for different vehicles is only possible if the 'behavior' of a given vehicle is known. Therefore simulations of different vehicles were performed in the run-up of the experiments. The simulations recurred a specific procedure: "For a given parameter set ('configuration' *ll, lr, rl, rr*), set target to position *a* (*a* = 1,...,20), and test for all 100 possible stimulus positions *b* what the final position *c* results and register the error angle *δ* between *a* and *c*." The simulations yield a comprehensive description of the individual behavior resulting from a given vehicle configuration. Several parameters were evaluated from the data. The simulation was applied for nemerous vehicles configurations with parameter sampling of 40 per parameter (0,05 steps in range [-1,1]). Stored was the distance between target position and end position in each run. Several aspects of the vehicle configurations were evaluated (see fig. 26 and fig. 27).



```
0.0 +0.30 +0.30 0.0 :: 5
   1    2    3    4    5    6    7    8    9    10   11   12   13   14   15   16   17   18   19   20
1  *****...........................................******..........................................**** 15! 0.004
2  ...********...................................***................................................... 11! 0.018
3  .........********............................*...................................................... 9!  0.015
4  ...............*******.............................................................................. 7   0.002
5  .....................*******........................................................................ 7   0.004
6  ...........................*******.................................................................. 7   0.013
7  ................................********............................................................ 8   0.015
8  ......................................******........................................................ 6   0.015
9  ...........................................**....................................................... 2   0.019
10 .................................................................................................... 0   0.338
11 .................................................................................................... 0   0.650
12 .................................................................................................... 0   0.968
13 .................................................................................................... 0   1.281
14 .................................................................................................... 0   1.246
15 .................................................................................................... 0   0.929
16 .................................................................................................... 0   0.618
17 .................................................................................................... 0   0.302
18 ...............................................................******************................... 18  0.000
19 ..........................................................*******.............************.......... 19! 0.003
```

fig. 26: ***Behavior of a specific vehicle configuration (overview).*** *Parameters of the vehicle configuration (ll,lt,rl,rr) are shown in the upper left followed by the number of SOT hits (shown after '::'), i.e. those cases in which stimulus on target (or directly opposite for negative phototactic) leads to a hit (ϵ < 0.2 RAD). Columns 1-20 of the table indicate final positions of the vehicle in a given run. Each comprises five tics, i.e. the circle is sampled by 100 tics. Rows 1-20 indicate the possible target positions. The asterisks represent hits (ϵ <0.2 RAD), the points in a row represent misses (ϵ >̲0,2 RAD). In the end of each row, the number of hits is shown, followed by the information on discontinuous hit (asterisk) series (exclamation mark) followed by the nearest possible approach to the target [RAD].*

[28] *This was carried out by the student coworker Olaf Gerstung.*

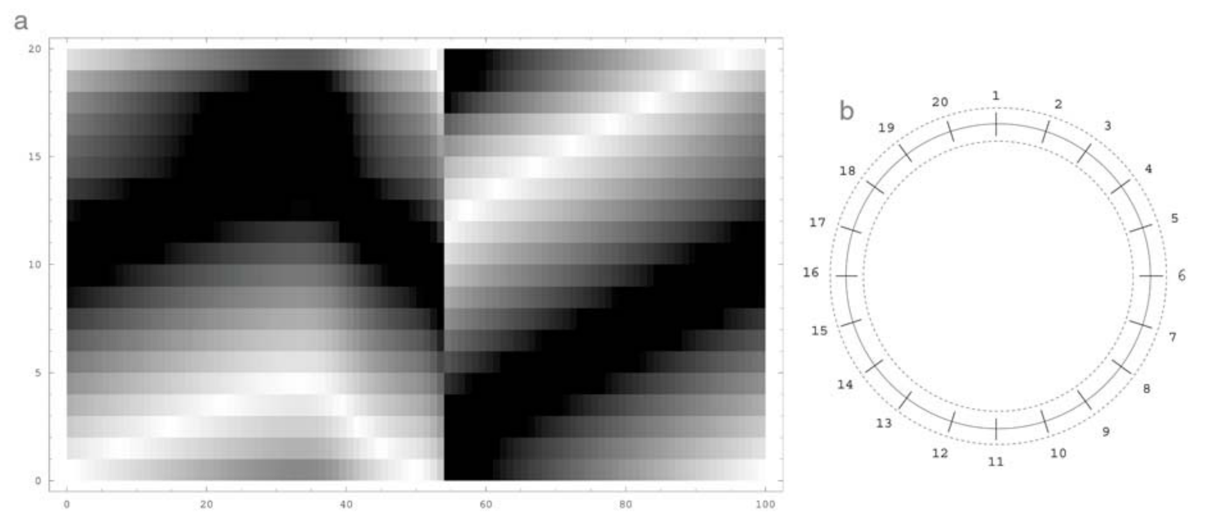


fig. 27: Distribution of errors for a given vehicle configuration. ***(a)*** *Columns 1-20 of the table indicate final positions of the vehicle in a given run. Each comprises five tics, i.e. the circle is sampled by 100 tics. Rows 1-20 indicate the possible target positions. Grey levels represent the error ϵ in a vehicle run: white corresponds to zero and black to maximum.* ***(b)*** *Possible target positions. Number 1-20 and large tics indicate the target position. Between each large tic are five small tics.*

3.2.2.5. *Selection of the vehicle configurations*

In order to decide, which concrete vehicle configuration should be applied in the experiment, the difficulty of a given vehicle configuration was approximated: first, the cases in which the strategy of placing stimulus on target (SOT) should be counted. Since the four vehicles in the conditions ‘pre’ and ‘post’ should not be too easy, only vehicles with less than 5 SOT hits (within a tolerance of >.2 rad) were taken as candidates for the four configurations. In some cases it is principally impossible to hit all targets. In order to keep learners motivated to try to get as near as possible, the number of valid targets should be maximized. (The learner should not get the impression that the target actually cannot be hit.) However, total exclusion of unreachable targets did not allow for ‘interesting’ vehicles, i.e. vehicles in which the number of SOT hits is minimized. (This dilemma shows that the process of finding the appropriate difficulty of a vehicle is always a trade-off between being interesting and being overtaxing.)

Split hit ranges (see discontinuous asterisk series indicated by exclamation marks in fig. 26) should also be minimized. These cases typically occur when the vehicle performs a 180° turn in order to move towards a target “in its’ back”, but is stopped before the turn could be completed because it exceeds

table 9: Overview of the chosen vehicles. *The rows show the different vehicles used and the corresponding condition. In columns 2 and 3 is shown if they are positive or negative phototactic and their symmetry, columns 4-7 contain the concrete configurations (connection strengths: l=left and r=right). The remaining columns contain statistics on behavior: SOT-number of SOT hits; splits-number of split hit ranges; invalid-number of unreachable targets.*

condition	*phototaxis*	*symmetric*	*ll*	*lr*	*rl*	*rr*	*SOT*	*splits*	*invalid*
intro	positive	yes	-1.00	0.00	0.00	-1.00	20	0	0
pre1	positive	yes	0.00	+0.30	+0.30	0.00	5	5	5
pre2	negative	yes	+1.00	0.00	0.00	+1.00	5	2	3
post1	positive	no	+1.00	0.00	0.00	+0.55	4	3	6
post2	negative	no	0.00	-0.80	-0.55	0.00	3	5	7

the arena's rim. In this way, it is possible to hit a target by placing a stimulus on a supposedly 'senseless' position. Analogously to the dilemma described above, split hit ranges could not be totally avoided without excluding all 'interesting' vehicles: these cases become increasingly probable with increasing asymmetry.

Five vehicle configurations were needed for the experiment (see also 3.2.1.1). One very simple for the intro phase, a positive and one negative phototactic, symmetric with less SOT hits for the pre-experimental phase, and one positive and one negative phototactic asymmetric for the post-experimental phase (see table 9).

3.2.2.6. *Data analysis*

A variety of interesting approaches to data analysis result from the conceptions introduced before (see also 3.2.1.2). But the central question posed in the experimental design concerns the differences between control group and experimental group. The influence of the instructional model simulation that was presented exclusively to the experimental group shall be quantified. The hypothesis is that the experimental group performs better in the conditions *post1* and *post2*, whereas the both groups should perform the same in the conditions *pre1* and *pre2*. Thus, a group-specific performance measure has to be determined for a single condition ('vehicle '). How can this be achieved? The performance in a single trial is measured by the error *δ*, i.e. the deviance of the vehicle of the target position in radians (see fig. 28). For determining the mean error for a single person, all errors could be averaged. But, as explained before (see also 3.2.1.2), the problem space is far too large at the beginning of a condition because the subject does not even know if the vehicle is positive or negative phototactic. Since it is not assumed that experimental and control group differ substantially in their ability to distinguish between vehicle preference (stimulus quality), this first phase (initial drop in the learning curve in fig. 28) can be ignored. Therefore, only the last 15 trials are averaged for determining the mean performance of a single subject. As already explained (see also 3.1.2.1), the correction for the differential chance levels has to be applied to the error of each single trial. The corrected values can be averaged for the experimental groups *exp* and control group *con,* which yields m_{exp} and m_{con} for each condition (*pre1, pre2, post1* and *post2*). Additionally, the inter-group variances s_{exp} and s_{con} for each condition has to be determined. A t-test for different variances (and possibly different sample sizes n_{exp} and n_{con}) $t = m_{exp} - m_{con} \, / \, (s_{exp} \, / \, n_{exp} + s_{con} \, / \, n_{con})^{-2}$ show the statistical certainty of the statement to be made (cf. Sachs 2003), i.e. whether or not there is a difference between the two groups in the conditions before the experimental treatment and after it. The obtained result will show the adequate way to carry on with the experiments.

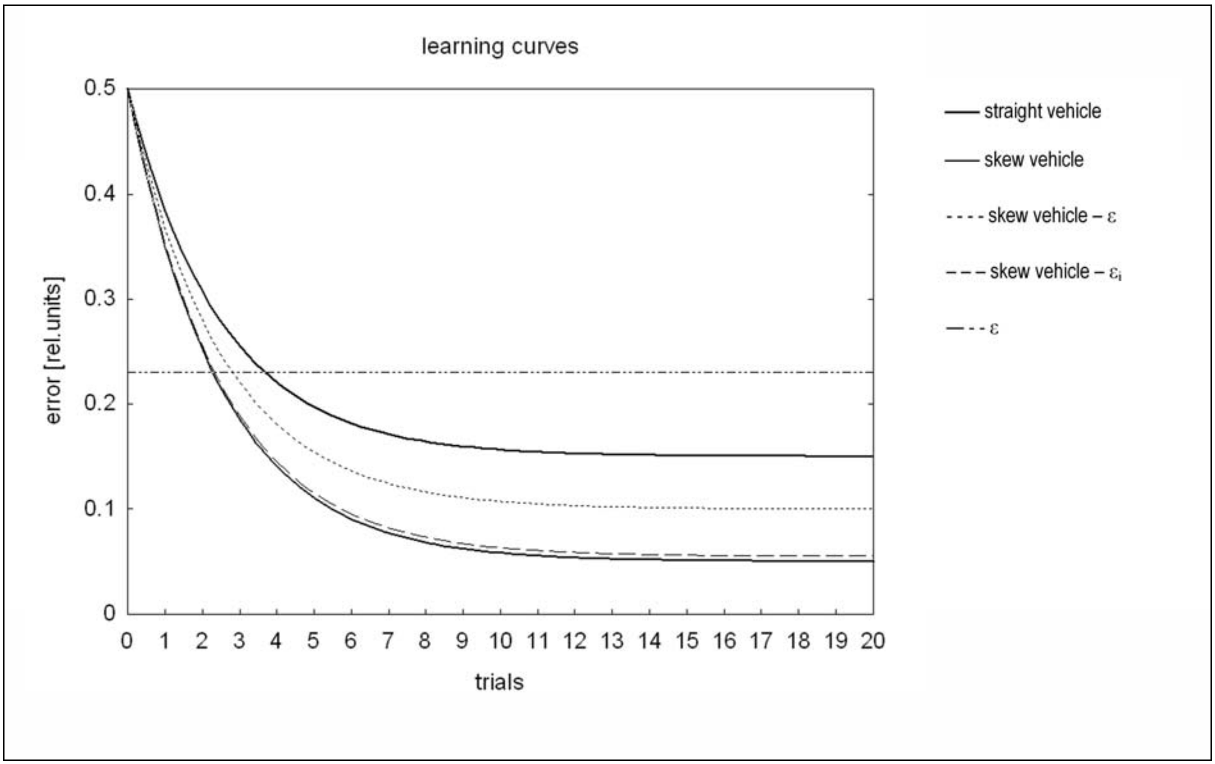


fig. 28: Learning curves. *Predictions for performance of modeled learners (ordinate) over trials (abscissa) on different tasks. Least errors are expected for straight vehicles. If a stimulus on target strategy is applied to skew vehicles larger errors will be produced that can however, be reduced if a generalized correction factor* ϵ *or a trial specific correction factor* ϵ_i *is applied.*

3.2.3. Discussion

3.2.3.1. *From performance to mental processes*

What exactly is the designated role of mental simulation in the experimental design proposed here? It is assumed here that mental simulation helps to overcome problems introduced by dynamics and complexity. In order to predict the final position of the vehicle in a simple task, e.g. a 'straight' vehicle, no hypotheses on intermediate positions of the vehicle are necessary. An interpolation between starting and final position is sufficient for producing correct results. But skew trajectories require interpolation with intermediate points. Thus, the *dynamic component in the task* is a change of direction during the trial that is introduced by the continuous sensorimotor processing: the internal states of sensors in each time step determine, via the motors, the turning angle in the next instance of time that determines the orientation to the stimulus and, therewith, the internal state of the next time step… It is assumed here that a mental simulation of the run (before the actual run) yields clues for determining the coordinates of intermediate positions. (This hypothesis might be nicely tested by studying eye movements). The intermediate positions of the vehicle are particularly helpful for illustrating to oneself that the differential activation of sensors can be very different in a late phase of a run as opposed to the starting position. This can lead to an understanding of the course of activation over the run and, finally, a prediction of asymmetric weighting of sensor activation that causes asymmetric shift magnitude distributions.

The *complex component in the task* is introduced by the interdependence of variables and, therewith, the amount of case differentiations necessary to predict the vehicle's behavior and minimize errors. For example, learners controlling a 'straight' vehicle just have to differentiate between positive and negative phototaxis (shift direction). But skewness of vehicles introduces an enormous amount of sources of error. One example is the mistaken direction of the compensatory shift: having a vehicle that ends left of target when stimulus on target, for example, the stimulus must be set more to the right of the target, but can erroneously be put more to the left. Another example is the failure of distinguishing between different shift distributions on the left and on the right or the assumption of a constant shift over a given range in the arena (whereas the shift magnitude increases with increasing angle).

These errors can be led back to a confusion of elements, mistaken mappings between two factors, misinterpretation of correlations etc. and are therefore summarized under complexity errors.

Complexity is inferable from the vehicle's behavior, but has its roots in the elements and activities that determine the vehicle's internal mechanisms – particularly from its combinatorial powers. Therefore, complexity errors can be successfully prevented if the internal mechanisms are known and can be mentally simulated. For example, differentiating shift directions might be better performed if learners assume that the vehicle has two motors. Then learners can hypothesize, e.g.: "The right motor is stronger..." A correspondence between a left-right distinction in the internal organization and a left-right distinction in the behavior can be established more easily: "... so the vehicle shows a spin to the left." Learners who do not have a clue on the internal mechanisms might even be unable to detect the systematic in the erroneous behavior (shift direction): "Why does that thing always run beside the target?“

As already noted above, learners might also successfully predict the trajectories without a concrete hypothesis on the internal mechanisms – just by detecting the 'spin' of the vehicle in phenomenological analysis: "It tends to move to the left, so I have to lure it a bit more to the right!" It might even be that these naive learners actually produce fewer errors than learners thinking in mechanisms because they are not loaded with the elements and activities that make up the mechanism. Then, naïve learners have more capacity for processing and are not prone complexity errors, e.g. confounding elements or inferring an erroneous hypothesis such as "the right sensor is stronger so the vehicle tends to go to the right“. Naive learners cannot have that hypothesis since they do not know about a right sensor. Thus, the introduction of a mechanism in the task might imply not only the potential to prevent complexity errors, but also to provoke complexity errors that otherwise would not be there. Even a performance drop caused by instruction has to be taken into account. On the other hand, the self-evident benefit of applying mechanistic thinking is to generate concrete hypothesis that can be verified and falsified. The final level of error production should therefore (depending on the application situation) be lower for learners that mentally simulate internal mechanisms of the vehicle than for learners that apply phenomenological control strategies: in this sense, mental simulation is a cognitive investment strategy. This line of arguments predicts a U-type learning curve for mental simulators, i.e. performance decreases first, but increases later. Eventually, mental simulation before every run might not be necessary anymore or is performed implicitly. (This should also correspond to a decrease in reaction time.) In this case, the mechanism is condensed to a (more or less) simple relation between stimulus position and target position. Then, the controlled and active mental simulation is to be interpreted as an intermediate phase – a mental tool – for detecting and learning a relation.

3.2.3.2. *Efficiency of mental simulation*

Controlled and active mental simulation can be viewed as a mental tool for detecting and learning a relation. But would such a result be sufficing for justifying the expense of mental simulation? In order to assess this question, it must be clarified whether there is an alternative way of solving the task. The investments of mental simulation would be clearly worth it if mental simulation were the only strategy that leads to superior performance. But if there is a simpler way that leads to the same performance, the benefit of mental simulation seems less clear. In the concrete case of the vehicle task, a critical question is: isn't there a much simpler way to correctly learn the shift magnitude distribution? Seen from the design stance, it seems much more efficient to directly aim at the characteristic – and not take the deviation with mental simulation. But is it actually possible to detect difficult relations such as shift magnitude distributions in the vehicle task without mental simulation? One strategy that might be used to detect even asymmetric shift magnitude distributions just by testing some stimulus positions could assign the respective shifts ('stimulus-on-target errors' as ordinate) to the arena's coordinates (target positions as abscissa) and interpolate between the samples. The result would be a 'mental characteristic' from which errors can be predicted and correct stimulus positions can be inferred. This can equally well be termed a 'mental model' of the vehicle as it reduces the vehicle to a single input-output relation. It does not imply any hypothesis about the internal organization of the vehicle and it does not necessitate dynamic activation of the model, e.g. as a mental film. Only a minimal mental simulation of a diagram (or a formula) is necessary to generate the correct prediction. These learners can be termed – due to their minimal dependence on mental simulation during and after the learning process – 'direct mental modelers'. But is it actually probable that such perfectly rational learners exist? One can only speculate on this question: a skilled experimenter or mathematician might detect and perform this strategy, simply due to prior knowledge on the analysis of formal relations or experimental situations. But also learners using a form of verbal self-instruction might describe the same characteristic, e.g. "The deviation of the vehicle increases with increasing distance from top and decreases in the middle of the lower left section."

However probable direct mental modelers are, they could exist. Therefore they must be discussed as opponents to mental simulators in the question of assessing the efficiency of mental simulation: direct mental modeling represents the case in which the same result can be obtained with an alternative (more parsimonious) learning strategy. What could still justify the expense of mental simulation? One obvious argument is that learning direct mental modeling is left to learners with specific prior knowledge (e.g. mathematicians, experimenters etc.). The exclusivity of simulations is reduced to a specific group of learners with (or without) a specific profile of prior knowledge. But are there also aspects that are not contained in direct mental modeling but in mental simulation – aspects that do not affect performance, but are learnt? Put another way, might the 'load' that mental simulators additionally bear, pay off in other situations not tested here? Both learners know that the vehicle can be lured to a given target by choosing a given stimulus position. But an obvious difference is that direct mental modelers do not consider *how* this happens, i.e. what the mechanism is that makes this possible. If a distinction between declarative and procedural components in the tasks is applied, it is the procedural component of the task that is much more pronounced in mental simulators.

Any task that refers to mechanistic properties of the vehicle should elicit the specific skills of mental simulators. For example, the task of designing a vehicle that produces a given behavior should be better performed by mental simulators because they have 'loaded' constituent elements and contributing activities that direct mental modelers do not have. Also the prediction of a completely new behavior, such as a trajectory influenced by another (distracting) light source or an obstacle (casting a shadow), should be better performed by mental simulators because they can define intermediate points (e.g. the vehicle being in front of the obstacle) and ask themselves, which states the sensors will have, which activation of motors result etc. (These tasks can be additionally presented for separating mental simulators from direct mental modelers in future work.) In sum, mental simulators should have the advantage over direct mental modelers in performing on new modifications of the task because they have the opportunity to generalize and to autonomously solve problems relating to the task.

Beside the skills of mental simulators to solve generalized tasks, some meta-skills (generic skills) are also trained and have to be taken into account for assessing the effort of mental simulation. One part of these meta-skills can be classified as "complexity competence": for example, a complexity reduction in the sense of differentiating elements (sensors, motors) and activities (activations) has to be carried out. Systematic thinking in the form of defining and testing hypothesis is necessary: specific states have to be assigned to elements and their influence to behavior of the system has to be predicted. Repetitions of this procedure help to sample a given parameter space and might yield clues for detecting general rules in the system. Another part of meta-skills can be summarized as "dynamics competence": the system can best be analyzed when starting and final states are defined. Sometimes defining intermediate states helps to find crucial moments in the systems. Serialization, i.e. 'chopping' the (temporal) course in phases and analyze these one after the other helps to handle dynamics.

In conclusion, the comparison of direct mental modeler and mental simulator yields important conclusions for assessing the efficiency of mental simulation: simulation is expensive in terms of "cognitive load", but should pay off when they are generalized or new versions of the problem are encountered. Thus, simulations are recommended in the following cases:

1. Simulation is exclusive: there is no alternative learning strategy. This will apply for specific complex and dynamic problems. (The trivial case of using simulations for costly or dangerous devices such as airplanes and power plants is not discussed here.)

2. Simulation is non-exclusive: there is a more parsimonious alternative that produces the same (or superior) performance in a given task, but …
(a) it produces less stable or less general results or
(b) the training of generic skills (meta-skills), e.g. complexity competence or dynamics competence (or even computer handling) is a goal of application.

Evidently, the benefits of mental simulation will only become present when the appropriate problems are posed. If simulation can be called exclusive or not is task-dependent: a learner using mental simulation to solve a problem that could also be solved by a simpler strategy wastes time and effort if the results of mental simulation can never again be applied. Thus, another conclusion is that, if only the task-specific skill is the goal in learning, it might be more efficient to refrain from simulation and aim at direct learning processes[29].

3.2.3.3. *Outlook*

The programmatic experimental approach proposed here is designed to answer questions on mechanisms and efficiency of mental simulation empirically. In a first step, the experiments as described in the section on methods will prove the existence or non-existence of a general effect of the applied design. More precisely, the first phase will show if the instructional computer simulation provided for the experimental group will make them outperform the control group in difficult vehicle tasks. Subsequent variations of the task, in which the instructional computer simulation will be decomposed in the constituent elements, will be necessary to answer questions on the specific cause of differences found between the groups or will elicit the causes why no differences can be found.

The experimental design makes clear that the task of assessing the value of a specific mental or artificial simulation (not to speak of a general assessment) is not trivial and bound to considerable effort. Even a simple system such as the vehicle used here offers a very large problem space as it becomes directly evident if all target positions of a single vehicle configuration are envisaged (see fig. 26 and fig. 27). If all possible vehicle configurations are taken into account, the problem space enlarges by orders of magnitude (the model simulations for determining the correct vehicle configurations took several

[29] Whether or not mental simulation is efficient is presumably not interesting for a single learner because the choice of problem solving strategies is unlikely to be a controlled, voluntary process but rather the (spontaneous or automatic) result of situational and individual factors. But educators, for example, are in the situation to decide whether or not a problem solving strategy is recommended. For instance, when applying a computer simulation for triggering mental simulation - beside didactical and technical efforts etc. that have to be taken into account - educators should ask themselves if one of the above mentioned benefits apply. However, they should allow for a performance drop! Give learners time to test the new mental simulation. It is the educator’s responsibility to make sure that the initial confusion will come down to higher performance level.

days). But skilled users – whatever strategy they might apply –nimbly control almost any vehicle and lead it through it's target. This shows what powers our cognitive systems offer – if we understand how to use it.

Reading Advice

The previous section demonstrated for a single case how the role of simulation in explaining (brains) could be assessed empirically. It illustrates the effort necessary to make simulation an objectively founded explanatory tool – and no longer a subjective choice of a researcher or an educator. Of course, in most of the cases it is not necessary to show the value of simulation explicitly because success of research and learning is justification enough. But it is desirable that there were more well known cases in order to make clear to critics why simulation is needed and why it is often time consuming and complicated. Even simple systems reveal considerable dynamics and complexity as the example of the Braitenberg vehicle revealed. If we wish to have another than monocausal understanding for dynamic and complex systems and a 'feeling' for their handling simulation is the way. If there were a simpler way, it would have prevailed long ago. But, essentially, there is no simple way for explaining non-simple phenomena.

EXPLAINING BRAINS BY SIMULATION – A SUMMARY

The task of explaining brains is hampered by the dynamics and complexity of brains. Complexity results from the extraordinary processing power of brains that is reached by distributing the continuously incoming stream of sensory signals over myriads of nerve cells that all compute "bits and pieces" in parallel. Moreover, explanatory approaches to these computations are to be found on multiple levels of organizations, e.g. electric, molecular, synaptic, networks etc. Thus, the comprehension problem caused by complexity is to keep track *where* the relevant processes happen. Dynamics occur as the incoming sensory signals are not only processed straight "downstream" towards behavioral response, but are rather recurrently fed into the stream to be compared and computed with sensory signals that come in later. This mix forms an "ongoing brain activity" that cannot be understood as a simple stimulus-response behavior. Thus, the comprehension problem caused by dynamics is to keep track *when* the relevant processes happen. Explainers of brain phenomena (and their recipients) are permanently challenged to push the limits of their mental capacities in order to handle the dynamics and complexity they find in brains.

Simulation is a key to dynamics and complexity. Scientists use it as a standard method to explain even highly dynamic and complex brain phenomena. But the question is: How exactly does simulation help to unlock hardly explainable brain phenomena from dynamics and complexity? For answering this question I propose to focus on mental simulation as a mechanism that generates explanations. A detailed account of mental simulation can reveal the specific problems caused by dynamics and complexity. Solutions to these problems are simultaneously the specific roles of simulation in explaining brains.

Explanation as mental simulation

A phenomenon is something that happens in a situation being observed by an explainer. An explanation is a possible answer to a question concerning a phenomenon. The situation is constituted by elements and activities. Phenomena show the characteristic that elements and activities change by themselves ('autonomously') from an initial state to a goal state – the situation has a 'momentum'. (A typical example of a natural momentum would be the generation of an action potential, which is reaching its goal state autonomously after being triggered.) The explainer observes changes of elements and activities, but the explainer does not directly observe the causal relations that determine the changes. The explainer has to hypothesize causal relations in order to obtain an explanation. Perceptual elements, activities and causal relations build a mental model of the situation.

A mental model can be actively operated to analyze a situation ‘offline’, even without actual observation, i.e. in absence of a situation and, thus, in absence of an external perceptual input. The explainer has to apply operational rules to perceptual elements, activities and causal relations (“thinking” or “reasoning“) to change the mental model from an initial state to a goal state. Operating a mental model enables the explainer to mentally simulate a situation. When the explainer observes a situation, the momentum that causes elements and activities to change form an initial state to a goal state is provided externally by the situation. If the explainer mentally simulates the situation, the momentum is generated internally. The process that generates an internal momentum and causes elements and activities to change from an initial state to a goal state is the genuine simulation scheme. If an internal momentum occurs, a simulation is functional. A functional mental simulation can be an explanation of a phenomenon. Thus, an explanation is a specific configuration of elements, activities, causal relations and operational rules that yields a possible answer to a question. An evaluation of the answer confirms or falsifies the hypothesis. Different types of explanation can be distinguished. A mechanistic explanation is obtained if the question posed calls for specifying the causal relations within the participating elements and activities, a constitutive explanation yields an answer to questions concerning the overall function of the phenomenon and a reductive explanation breaks down elements and activities to constituent elements and activities and underlying mechanisms that determine their behavior.

Since the mental capacity of explainers is limited, not every mental simulation is functional – it can also fail. If the mental model of a given brain phenomenon is complex and dynamic, the mass of elements and the massive activity result in a mass of possible causal relations to be hypothesized. The operational rules needed to ‘drive’ the mental model in order to make a simulation functional cannot be defined: no specific configuration of elements, activities and causal relations is obtained and the mental model cannot be operated. The mental simulation obstructs and no momentum is generated that causes perceptual elements and activities to change from an initial state to a goal state. The causal relations within and between elements and activities cannot be extracted. The phenomenon stays unexplained.

Scientific work and simulation

Scientific work is a systematic approach to find explanations even for complex and dynamic phenomena. The general applicable strategy to handle dynamics and complexity is serialization and decomposition. In experimental work, for instance, the natural situation in which a phenomenon is originally observable is reduced to a natural preparation, i.e. decomposed until the minimum number of elements and activities is reached that is necessary to preserve the phenomenon. Additionally, operational rules are specified that cause the elements and activities to change from an initial state to a goal state thereby revealing the phenomenon. Decomposing a natural situation to a natural preparation and serializing it with appropriate operational rules yields an experiment. The procedure of experimental work is analogous to the procedure applied in mental simulation. Therefore, an experiment involving a natural preparation can be termed natural simulation. An experiment can answer a question, confirm or falsify a hypothesis and may provide an explanation of a phenomenon. An experimental procedure is an instruction of how to configure elements and activities in order to reveal causal relations and to obtain an explanation. By transferring the experimental procedure applied in the (external) experiment to the (internal) mental model, explainers can learn how to succeed in mentally simulating a phenomenon. Experiments are recipes for mental simulations.

When decomposition or serialization of the natural situation is difficult or impossible and mental simulation fails, scientists can strive for the construction of an artificial preparation (e.g. a computer simulation or robot): elements, activities and causal relations that are supposed to be crucial for the phenomenon are defined as parts of a mental model and externalized as an artificial preparation. Artificial operational rules can cause the artificial preparation to change from an initial state to a goal state. A functional artificial simulation yields an artificial momentum that can be fed to the mental simulation to make it functional. Thus, experiments with artificial preparations are performed in exactly the same way as experiments are performed with natural preparations. The principal difference between natural and artificial lies in the origin of the momentum – the former is designed by nature, while the latter is designed by cognition. Since cognitive design provides more degrees of freedom for artificial simulations, decomposition and serialization of the natural situation and therewith explanation of the phenomenon can be carried out more easily. Factors that cannot be controlled in the natural preparation can be eliminated and factors that cannot be applied in the natural preparation can be added. Dynamics and complexity can thereby be reduced or even increased but in any case better controlled. In this manner, artificial simulation can offer ways to handle dynamics and complexity that are locked up for experiments with natural preparations. However, with respect to their explanatory value there is no difference at all between natural and artificial preparations. Both support mental simulation in that the perceptual elements, activities and the momentum is caused externally and, consequently, the effort necessary to mentally simulate these does not have to be invested: cognitive capacity is saved and more capacity can be allocated to the analysis of possible causal relations that explain the phenomenon. Just as an experiment with a natural preparation, the artificial simulation results in an instruction of how to think about a phenomenon – how to explain it. Considering mental simulation clarifies why simulation is particularly important for explaining brains: dynamics and complexity of the brain prevent successful explanations and simulation offers remedies. Mental simulation is the most potent explanatory mechanism for complex and dynamic phenomena.

The view proposed here differs substantially from the notion of simulation commonly found in the public and even in science. One common conception is that simulation is reduced to computer simulation. In the view proposed here, computer simulation is merely a specific extension of mental simulation in which external models are designed and operated. A second common conception is that simulation is a weak form of scientific work because it does not deal with the "real" system. True is that artificial simulation does not deal with the natural. But wrong is that dealing with the natural means to deal with reality, whereas dealing with the artificial means to deal with something unreal. The natural preparation is not the slightest bit more real than the artificial preparation. (The stone in the woods and the coffee cup on my table are both perfectly real.) The difference is that the natural preparation shows a momentum that is designed by nature and the artificial preparation shows a momentum that was designed by cognition. Of course, the artificial preparation is a part of nature and builds a special case of natural phenomena, namely to be designed for feeding a mental simulation so that an explanation can be generated. Artificial simulations are explanatory tools in exactly the same sense experiments with natural preparations are explanatory tools. A third common conception about simulation comes from the pro-simulation fraction: they sometime hold the view that formal, mathematical descriptions of a phenomenon are the most elegant explanation of a natural phenomenon and a functional artificial simulation is the ultimate proof of its truth. True is that formal models are an optimal means for designing artificial simulations and therewith to generate explanations (of complex and dynamic phenomena). Wrong is that these explanations are in any way superior to explanations based on experiments with natural preparations. Both serve the purpose of mediating between natural situation and mental simulation. Both are inferior to mental simulation in that they depend on external preparations. (And the most versatile externalization of the mental simulation, however, still is a lingual explanation.)

The view proposed here does not only re-evaluate common conceptions about simulation by placing simulation where it belongs – away from the computer back in the head of the explainer – it also provides a means for assessing the value of simulation. The question of whether or not it is sensible to use simulation can be clearly answered: if it helps to explain a phenomenon, it is justified (ethical issues left aside). The specific value of a simulation is determined by its explanatory power. This value can even be quantified empirically. Putting mental simulation to the center allows to naturalize explanation, i.e. to treat explanation as a cognitive phenomenon that can be traced experimentally, for instance by applying methodology of cognitive psychology. Mental simulation as an explanatory mechanism is closely related to the psychology of mental models, learning, memory and thinking. Thus, theoretical and experimental approaches to mental simulation already exist. When, additionally, education and instructional design are taken into account, prescriptions for designing ‘good’ explanations that can be understood by a wealth of people come within reach. This finally leads to a fourth common conception about simulation –

them being a very special and rare event that has a highly theoretical background. True is that simulation helps us to develop a theory of what is happening in the world. Wrong is that simulation is rare – it happens any time you think.

Post Scriptum: Throughout the study, brains served as exemplars giving rise to complex and dynamic phenomena that can best be explained by simulation. But beyond that, brains themselves provide the causal powers that make possible mental simulation, experiments on brains and the design of artificial models of brains. Furthermore, learning has a well-evidenced neural basis and can be regarded as the basis of scientific work. Considering the brain inevitably introduces such a causal density in the theme that I continuously felt tempted to construct references from any topic to any other topic. But, even though I have a neuroscientific background (or exactly because of that), I tried to resist this temptation for preventing things to become too self-referential, too complex and too dynamic – and in favor of an at least approximated serial and decomposed argumentation structure. But the reader is invited to autonomously extend the issue of explaining brains by simulation with further aspects such as "brains simulating explanation", "brains explaining simulation", "simulating brains with explanation" ...

REFERENCES

Anderson, J. R. (1980). Cognitive Psychology and Its Implications, 5 ed. Freeman, San Francisco.

Anderson, J. R. (1993). Rules of the Mind Erlbaum, Hillsdale, NJ.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000). Perspectives on Learning, Thinking, and Activity. Educational Researcher 29, 11-13.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. Educational Researcher 25, 5-11.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1997). Situative versus cognitive perspectives: Form versus substance. Educational Researcher 26, 18-21.

Arbib, M. A. & et al. The Human Brain Project. http://www-hbp.usc.edu/, Access Date 2001.

Arbib,M.A. (1998). The handbook of brain theory and neural networks. (Cambridge, Mass.: MIT Press).

Arbib,M.A., Érdi,P., and Szentágothai,J. (1997). Neural organization: structure, function, and dynamics. (Cambridge, Mass.: MIT Press).

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes., eds. Spence, K. W. & Spence, J. T., pp. 89-105. Academic Press, New York.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of acompositional system of perceptual symbols. In Theories of memories, eds. Collins, A. C., Gathercole, S. E., & Conway, M. A., pp. 29-101. Erlbaum, London.

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences 22, 577-609.

Bartlett, F. C. (1932). Remembering: An Experimental and Social Study Cambridge University Press, Cambridge.

Bassok, M. (1997). Object-based reasoning. In The psychology of learning andmotivation (37), ed. Medin, D. L., pp. 1-39. Academic Press., San Diego.

Baudrillard, J. (1988). Selected Writings Stanford University Press, Stanford.

Bechtel,W. and Abrahamson,A. (2002). Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in networks. (Oxford: Blackwell).

Bechtel, W. (2001a). Cognitive Neuroscience: Relating Neural Mechanisms and Cognition. In Theory and Method in the Neurosciences, eds. Machamer, P. K., Grush, R., & McLaughlin, P., pp. 81-111. University of Pittsburgh Press, Pittsburgh.

Bechtel, W. (2001b). Representations: From neural systems to cognitive systems. In Philosophy and the Neurosciences: A Reader, eds. Bechtel, W., Mandik, P., Mundale, J., & Stufflebeam, R. S., Basil Blackwell, Oxford.

Bechtel, W., Mandik, P., Mundale, J., & Stufflebeam, R. S. (2001). Philosophy and the Neurosciences: A reader Basil Blackwell.

Bechtel, W. & Richardson, R. C. (1993). Discovering complexity: Decomposition and localization as strategies in scientific research Princeton University Press, Princeton.

Bennett, M. V. (1997). Gap junctions as electrical synapses. Journal of Neurocytology 26, 349-366.

Berry, D. C. & Dienes, Z. (1993). Implicit Learning: Theoretical and Empirical Issues Erlbaum, Hillsdale, NJ.

Berry, D. C. & Broadbent, D. E. (1988). Interactive tasks and the implicit/explicit distinction. British Journal of Psychology 79, 251-272.

Best,J.B. (1999). Cognitive Psychology. West Wadsworth).

BIOSIS. Current Zoological Record Subject Hierarchy. http://www.biosis.org/zrdocs/zr_thes/subjvoc/index.html, Access Date 2002.

Blackwell Science Publishers. A Glossary of Psychological Terms. http://www.blackwellpublishers.co.uk/psychol/Glossary.htm, Access Date 2002.

Borst, A. & Egelhaaf, M. (1989). Principles of visual motion detection. Trends in Neurosciences 12, 297-306.

Braitenberg, V. (1984). Vehicles: Experiments in Synthetic Psychology MIT Press, Cambridge, Mass.

Brewer, W. (1987). Schemas versus mental models in human reasoning. In Modelling Cognition, ed. Morris, P., pp. 187-197. John Wiley, New York.

Broadbent, D. E. (1958). Perception and Communication Pergamon Press., London.

Broadbent, D. E. (1975). The magic number seven after fifteen years. In Studies in Long-Term Memory, eds. Kennedy, A. & Wilkes, A., pp. 3-18. Wiley, New York.

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. Quarterly Journal of Experimental Psychology 29, 181-201.

Buckner, R. L., Petersen, S. E., Ojemann, J. G., Miezin, F. M., Squire, L. R., & Raichle, M. E. (1995). Functional anatomical studies of explicit and implicit memory retrieval tasks. Journal of Neuroscience 15, 12-29.

Bush, G. (1990) “Decade of the Brain” Presidential Proclamation 6158. http://lcweb.loc.gov/loc/brain/proclaim.html, Access Date 2003.

Carnap, R. (1995). An introduction to the philosophy of science, pp. 6-16. Dover Publications Inc., New York.

Carruthers, P., Stich, S. P., & Siegal, M. (2002). The cognitive basis of science Chapman and Hall, London.

Center for Naval Analyses. Nonlinear Dynamics and Complex Systems Theory: Glossary of Terms. http://www.cna.org/isaac/Glossb.htm, Access Date 2002.

Chiel, H. D. & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system. Trends in Neurosciences 20, 553-557.

Chrucky, A. Meta-Encyclopedia of Philosophy. http://www.ditext.com/encyc/frame.html, Access Date 2002.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. Journal of Philosophy 78, 67-90.

Churchland, P. M. (1986). Neurophilosophy: Toward a Unified Science of Mind/Brain MIT Press, Cambridge, Mass.

Churchland, P. S. & Sejnowski, T. J. (1992). The computational brain MIT Press, Cambridge, Mass.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. Trends in Cognitive Sciences 2, 406–416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. Cognition 31, 187–276.

Craver, C. & Darden, L. (2001). Discovering Mechanisms in Neurobiology: The Case of Spatial Memory. In Theory and Method in the Neurosciences, eds. Machamer, P. K., Grush, R., & & McLaughlin, P., pp. 112–138. University of Pittsburgh Press, Pittsburgh.

Crick, F. H. C. (1989). The recent excitement about neural networks. Nature 337, 129–132.

Crookall, D. (2001). State of the art and science of simulation/gaming. Simulation & Gaming 32, 449–450.

Cruse, H. (1996). Neural networks as cybernetic systems Thieme, Stuttgart/New York.

Cruse, H. (2001). The Explanatory Power and Limits of Simulation Models in the Neurosciences. In Theory and Method in the Neurosciences, eds. Machamer, P. K., Grush, R., & McLaughlin, P., pp. 138–154. University of Pittsburgh Press, Pittsburgh.

Cruse, H. (2003). The evolution of cognition -- a hypothesis. Cognitive Science 27, 135–155.

Daugman, J. G. (1990). Brain Metaphor and Brain Theory. In Computational Neuroscience, ed. Schwartz, E. L., MIT Press, Cambridge, Mass.

Dawson, M. W. The University of Alberta's Cognitive Science Dictionary. http://web.psych.ualberta.ca/~mike/Pearl_Street/Dictionary/entries.html#top, Access Date 2002.

Dienes, Z. & Perner, J. (1999). A theory of implicit and explicit knowledge. Behavioral and Brain Sciences 22, 735–755.

Dörner, D. (1989). Die Logik des Mißlingens Rowohlt Verlag., Reinbek.

Dörner, D. (1998). Bauplan für eine Seele Rowohlt Verlag, Reinbek.

Dörner, D. & Wearing, A. (1995). Complex problem solving: Toward a (computer-simulated) theory. In Complex problem solving: The European Perspective, eds. Frensch, P. A. & Funke, J., pp. 65–69. Lawrence Erlbaum Associates, Hillsdale, NJ.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). Lohhausen: Vom Umgang mit Komplexität Huber, Bern.

Dretske, F. (1981). Knowledge and the Flow of Information MIT Press, Cambridge, Mass.

Dudel,J., Menzel,R., and Schmidt,R.F. (2001). Neurowissenschaft: vom Molekül zur Kognition. (Berlin: Springer).

Egelhaaf, M. & Borst, A. (1993). A look into the cockpit of the fly: Visual orientation, algorithms and identified neurons. Journal of Neuroscience 13, 4563-4574.

Eliasmith, C. Dictionary of Philosophy of Mind. http://artsci.wustl.edu/~philos/MindDict/index.html, Access Date 2002.

Elsevier Science Publishers. Thesaurus of neuroscientific terms. http://www.elsevier.com/.dejavu/Thesauri/NEUROSCI/show/, Access Date 2002.

Fausett,L.V. (1994). Fundamentals of Neural Networks. Prentice Hall).

Fishwick, P. A. (1995). A Taxonomy for Simulation Modelling Based on Programming Language Principles. IEEE-Transactions on IE Research 30, 811-820.

Flake, G. W. The Computational Beauty of Nature: glossary. http://mitpress.mit.edu/books/FLAOH/cbnhtml/glossary-C.html, Access Date 2002.

fMRI Data Center. fMRI Data Center. http://www.fmridc.org, Access Date 2003. 8-9-2003.

Fodor, J. (1975). The Language of Thought Thomas Y. Crowell.

Fodor, J. (1987). Psychosemantics: The problem of Meaning in the philosophy of mind MIT Press, Cambrige, Mass.

Forbus, K. D. (1983). Qualitative Reasoning about Space and Motion. In Mental Models, eds. Gentner, D. & Stevens, A., pp. 53-73. Erlbaum, Hillsdale.

Frensch, P. A. & Funke, J. (1995). Complex problem solving: The European perspective Erlbaum, Hillsdale, NJ.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In Complex problem solving: Principles and mechanisms, eds. Sternberg, R. J. & Frensch, P. A., pp. 185-222. Erlbaum., Hillsdale, NJ.

Funke, J. (1992). Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung Heidelberg: Springer, Heidelberg.

Galvani, L. (1791). De viribus electricitatis in moto musculari commentarius (German translation). In Ostwalds Klassiker der exakten Wissenschaften, ed. Oettingen, A. J. v., Engelmann, Leipzig.

Gazzaniga,M.S. (2000). The new cognitive neurosciences. (Cambridge, Mass.: MIT Press).

Gazzaniga, M. S. (2002). Cognitive neuroscience, 2 ed. WW Norton & Co.

Gell-Mann, M. (1995). What is Complexity? Complexity 1 19.

Gentner, D. (2001). Spatial metaphors in temporal reasoning. In Spatial schemas in abstract thought pp. 203-222. MIT Press, Cambridge, Mass.

Gentner, D. & Stevens, A. (1983). Mental Models Erlbaum, Hillsdale, NJ.

Gentner.D & Gentner.D.R. (1983). Flowing Waters for Teeming Crowds: Mental Models of Electricity. In Mental Models, eds. Gentner, D. & Stevens, A., pp. 99. Lawrence Earlbaum Associates.

Giere, R. N. (1992). Cognitive models of science University of Minnesota Press, Minneapolis.

Glenberg, A. M. (1997). What memory is for. Behavioral and Brain Sciences 20, 1-55.

Golgi, C. (1906). The neuron doctrine - theory and facts. In Nobel Lectures, Physiology or Medicine 1901-1921 Elsevier, Amsterdam.

Gordon, R. (1986). Folk Psychology as Simulation. Mind and Language 1, 158-171.

Greenberg, H. J. Mathematical Programming Glossary. http://carbon.cudenver.edu/~hgreenbe/glossary/glossary.html, Access Date 2002.

Greeno, J. G. (1989). Situations,mental models,and generative knowledge. In Complex information processing, eds. Klahr, D. & Kotovsky, K., pp. 285-318. Lawrence Erlbaum, Hillsdale, NJ.

Greeno, J. G. (1997). On claims that answer the wrong questions. Educational Researcher 26, 5-17.

Grethe,J.S. and Arbib,M.A. (2001). Computing the Brain: A Guide to Neuroinformatics. Academic Press Inc.).

Grush, R. (1997). The architecture of representation. Philosophical Psychology 10, 5-25.

Haag, J. & Vermeulen, A. B. A. (1999). The Intrinsic Electrophysiological Characteristics of Fly Lobula Plate Tangential Cells: III. Visual Response Properties. Journal of Computational Neuroscience 7, 213-234.

Haider, H. (1992). Implizites Wissen und Lernen. Ein Artefakt? Zeitschrift für experimentelle und angewandte Psychologie 39, 68-100.

Haider, H. & Frensch, P. A. (1996). The role of information reduction in skill acquisition. Cognitive Psychology 30, 304-337.

Heitzmann, W. R. (1973). The validity of social science simulations: a review of research findings. Education 94, 170-175.

Hertz,J.A., Krogh,A.S., and Palmer,R.G. (1991). Introduction to the Theory of Neural Computation. Addison Wesley Longman).

Hesslow, G. (1994). Will Neuroscience Explain Consciousness? Journal of Theoretical Biology 171, 29-39.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences 6, 242-247.

Heylighen, F. Principia Cybernetica Web: Dictionary of Cybernetics and Systems. http://pespmc1.vub.ac.be/ASC/indexASC.html, Access Date 2002.

Hirschfeld, L. A. & Gelman, S. A. (1994). Mapping the mind: Domain specificity in cognition and culture Cambridge University Press, New York, NY.

Hodgkin, A. L. & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology 117, 500-544.

Horstmann, W. (2002). Standardizing simulations -- uphill all the way. In Campus 2002, eds. Bachmann, G., Haefeli, O., & Kindt, M., pp. 218-230. Waxmann, Zürich.

Howe, D. Free On-line Dictionary of Computing (FOLDOC). http://wombat.doc.ic.ac.uk/foldoc/, Access Date 2002.

IEEE Consortium. Learning Object Metadata. http://ltsc.ieee.org, Access Date 2003. 2003.

ISI. Web of Knowledge. http://www.isinet.com/isi/, Access Date 2002.

ISO International Standards Organisation. Standard 5963: Methods for examining documents, determining their subjects, and selecting indexing terms. http://www.nlc-bnc.ca/iso/tc46sc9/standard/5963e.htm, Access Date 2002.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imager. Behavioral and Brain Sciences 17, 187-245.

Jeannerod, M. (2001). Neural Simulation of Action:A Unifying Mechanism for Motor Cognition. NeuroImage 14, 109.

Jeannerod, M. & Victor, F. (1999). Mental imaging of motor activity in humans. Current Opinion in Neurobiology 9, 735-739.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. Cognitive Science 4, 72-115.

Johnson-Laird, P. N. (1983). Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness Cambridge University Press, Cambridge.

Johnson-Laird, P. N. (1999). Deductive reasoning. Annual Review of Psychology 50, 109-135.

Joslyn, C. & Turchin, V. The Modeling Relation. Heylighen, F. Joslyn C. and Turchin, V. http://pespmc1.vub.ac.be/MODEL.html, Access Date 1993. Principia Cybernetica, Brussels. 11-9-2003.

Kahneman, D. & Tversky, A. (1982). The simulation heuristic. In Judgment under uncertainty: Heuristics and biases, eds. Kahneman, D., Slovic, P., & Tversky, A., pp. 201-210. Cambridge University Press, NewYork.

Kahng, J., Liao, W. K., & McLeod D. (1997). Mining generalized term associations: Count Propagation Algorithm. In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining pp. 203-206. Newport Beach, CA.

Kandel,E.R. and Schwartz,J.H. (1991). Principles of Neural Science. (New York: Elsevier).

Kandel,E.R., Schwartz,J.H., and Jessell T.M. (1996). Essentials of Neural Science and Behavior. McGraw-Hill Professional Publishing).

Kern, R., Lutterklas, M., Petereit, C., Lindemann, J. P., & Egelhaaf, M. (2001). Neuronal processing of behaviourally generated optic flow: experiments and model simulations. Network: Computational and Neural Systems 12, 351-369.

Kieras, D. E. & Bovair, S. (1984). The role of a mental model in learning to operate a device. Cognitive Science 8, 255-273.

Klahr, D. (2000). Exploring Science MIT Press, Cambridge, Mass.

Koch,C. (1999). Biophysics of computation : information processing in single neurons. (New York: Oxford University Press).

Koper, R. Educational Modeling Language. http://eml.ou.nl/, Access Date 2003.

Koslow, J. H. The Human Brain Project, National Institute of Mental Health. http://www.nimh.nih.gov/Neuroinformatics/index.cfm, Access Date 2001.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. Artificial Intelligence 33, 1-64.

Lambrinos, D., Moller, R., Labhart, T., Pfeifer, R., & Wehner, R. (2000). A mobile robot employing insect strategies for navigation. Robotics and Autonomous Systems 30, 64.

Leach, A. R. (2001). Molecular Modelling - Principles and Applications, 2 ed. Pearson Education Limited.

Lee, F. J. & Anderson, J. R. (2001). Does Learning a Complex Task Have to be Complex? A study in learning decomposition. Cognitive Psychology 42, 267-316.

Lemmon, E. J. (1987). Beginning Logic, 2 ed. Chapman and Hall, London.

Library of Congress. Library of Congress Subject Headings. http://www.loc.gov, Access Date 2002.

Logan, G. (1988). Toward an instance theory of automatization. Psychological Review 95, 492-527.

Machamer, P. K., Darden, L., & CraverC. (2001). Thinking about Mechanisms. Philosophy of Science 67, 1-25.

Magnani, L., Nercessian, N., & Thagard, P. (1999). Model-based reasoning in scientific discovery Plenum Press.

Malsburg, C. v. d. & Schneider, W. (1986). A neural cocktail-party processor. Biological Cybernetics 54, 29-40.

Mandler, J. (1984). Stories, Scripts, and Scenes: Aspects of Schema Theory Erlbaum, Hillsdale, NJ.

Markman, A. B. (1999). Knowledge Representation Erlbaum, Mahwah, NJ.

Markman, A. B. & Gentner, D. (2001). Thinking. Annual Review of Psychology 52, 223-247.

Marr, D. (1982). Vision: a computational investigation into the human representation and processing of visual information Freeman, San Francisco.

Maynard Smith, J. (2000). The concept of information in biology. Philosophy of Science 67, 177-194.

McClelland, J. L. & Rumelhart, D. E. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2 Psychological and biological models MIT Press, Cambridge, Mass.

Medin,D.L., Ross,B.H., and Markman,A.B. (2001). Cognitive Psychology. John Wiley & Sons, Inc.).

Medin, D. Cognitive Psychology: Glossary. http://www.psych.nwu.edu/psych/people/faculty/medin/book/glossary.html, Access Date 2002.

Merriam-Webster. Merriam-Webster Dictionary. http://www.m-w.com, Access Date 2003.

Millikan, R. G. (1984). Language, thought and other biological categories MIT Press, Cambridge, Mass.

Millikan, R. G. (1989). Biosemantics. Journal of Philosophy 86, 281-297.

Minsky M. (1975). A framework for representing knowledge. In The Psychology of Computer Vision, ed. Winston, P., pp. 211-277. McGraw-Hill, New York.

Mitchell, S. (1997). Pragmatic Laws. Philosophy of Science 64 Proceedings, 468-479.

National Institute of Health. Alcohol and Drug Abuse Thesaurus. http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm, Access Date 2002a.

National Institute of Health. Medical Subject Headings. http://www.nlm.nih.gov/mesh/2002/MeSHtree.G.html, Access Date 2002b.

Neisser, U. (1967). Cognitive psychology Appleton-Century-Crofts, New York.

Nernst, W. (1888). Zur Kinetik der in Loesung befindlichen Koerper, Theorie der Diffusion. Zeitschrift fuer physikalische Chemie 2, 613-637.

Newell, A. (1990). Unified theories of cognition Harvard University Press, Cambridge, Mass.

Nicholls,J.G., Martin,A.R., and Wallace B.G. (2001). From neuron to brain: a cellular and molecular approach to the function of the nervous system. (Sunderland, Mass.: Sinauer).

Nissen, M. J. & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology 19, 1-32.

Online Computer Library Center. Dewey's decimal classification. http://www.oclc.org/dewey/, Access Date 2002.

O'Reilly,R.C. and Munakata,Y. (2000). Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. (Cambridge, Mass.: MIT Press).

Pavlov, I. (1904). Physiology of Digestion -- Nobel Lecture. In Nobel Lectures, Physiology or Medicine 1901-1921 Elsevier Publishing Company, Amsterdam.

Payne, S. J. (1992). On Mental Models and Cognitive Artefacts. In Models in the Mind: Theory, Perspective & Application, eds. Rogers, Y., Rutherford, A., & Bibby, P., Academic Press, London.

Platt, M. (2002). Neural correlates of decisions. Current Opinion in Neurobiology 12, 141-148.

Ploetzner, R. & VanLehn, K. (1997). The acquisition of qualitative physics knowledge during textbook-based physics training. Cognition-and-Instruction 15, 169-205.

Principe,J.C., Euliano,N.R., and Lefebvre,W.C. (2000). Neural and adaptive systems: fundamentals through simulations. (New York: Wiley).

Putnam, H. (1975). The Meaning of 'Meaning'. In Mind, Language, and Reality pp. 215-271. Cambridge University Press, Cambridge.

Rall, W. (1989). Cable theory for dendritic neurons. In Methods in Neuronal Modelling, Computational Neuroscience, eds. Koch, C. & Segev, I., MIT Press, Cambridge, Mass.

Ramón y Cajal, S. (1906). The structure and connexions of neurons. In Nobel Lectures, Physiology or Medicine 1901-1921 pp. 189–217. Elsevier, Amsterdam.

Reber, A. S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 6, 855-863.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory, eds. Black, A. H. & Prokasy, W. F., pp. 64-99. Appleton-Century-Crofts, New York.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In Theoretical Issues in Reading and Comprehension, eds. Spiro, R. J., Bruce, B., & Brewer, W. F., pp. 33-58. Erlbaum, Hillsdale, NJ.

Rumelhart, D. E. & Ortony, A. (1977). The representation of knowledge in memory. In Schooling and the acquisition of knowledge, eds. Anderson, Spiro, & Montague.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature 323, 533-536.

Sachs, L. (2003). Angewandte Statistik, 11 ed. Springer, Berlin.

Schaffner, K. F. (2001). Extrapolation from Animal Models: Social Life, Sex and Super Models. In Theory and Method in the Neurosciences, eds. Machamer, P. K., Grush, R., & McLaughlin, P., pp. 200-230. University of Pittsburgh Press, Pittsburgh.

Schank R. & Abelson R. (1977). Scripts, Plans, Goals and Understanding Erlbaum, Hillsdale, NJ.

Schmucker, K. & Apple Computer Inc. (2000). A Taxonomy of Simulation Software. Learning Technology Review Fall 1999/Spring 2000.

Schutter, E. d. European Union Thematic network: "Computational Neuroscience". http://www.neuroinf.org, Access Date 2002.

Schwartz, D. L. & Black, T. (1999). Inferences through imagined actions: knowing by simulated doing. Journal of Experimental Psychology: Learning, Memory and Cognition 25, 116-136.

Schwartz, E. L. (1990). Computational Neuroscience MIT Press, Cambridge, Mass.

Searle, J. R. (1980). Minds, brains and programs. Behavioral and Brain Sciences 3, 417-457.

Shadlen, M. N. & Newsome, W. T. (1994). Noise, neural codes and cortical organization. Current Opinion in Neurobiology 4, 569-579.

Shepherd, G., Mirsky, J., Healy, M., Singer, M., Skoufos, E., Hines, M., Nadkarni, P., & Miller, P. (1998). The Human Brain Project: Neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. Trends in Neurosciences 21.

Sober, E. (1997). Two Outbreaks of Lawlessness in Recent Philosophy of Biology. Philosophy of Science 64 Proceedings, 458-467.

Thagard, P. (1992). Conceptual revolutions Princeton University Press., Princeton, NJ.

Tooby, J. & Cosmides, L. (1989). Evolutionary psychology and the generation of culture. Part I: Theoretical considerations. Ethology and Sociobiology 10, 29-49.

Walker, A. (1997). Thesaurus of psychological index terms, 8 ed. American Psychological Assoc., Washington, DC.

Walter, G. G. (1999). Compartmental Modeling With Networks Springer Verlag.

Webb, B. (2001). Can robots make good models of biological behaviour? Behavioural and Brain Sciences 24, 1033-1050.

Wehner, R. (1994). The polarization-vision project: championing organismic biology. In Neural Basis of Behavioural Adaptations, eds. Schildberger, K. & Elsner, N., pp. 103-143. Gustav Fischer Verlag, Stuttgart.

CV – CURRICULUM VITAE

13/02/1971 Born in Bielefeld, Germany

1977 – 1981	Elementary School: Grundschule Werther, Westf.
1981 – 1987	Intermediate Classes: Ev. Pro-Gymnasium Werther, Westf.
1987 – 1990	Senior Classes: Max-Planck-Gymnasium Bielefeld.
1990 – 1992	Community Service (environmental): Umweltzentrum Bielefeld, AKUT e.V.

10/92 – 09/98
Studies: Biology (Diploma), Psychology & Philosophy (Courses) at Bielefeld University

Since 06/96
Department of Neurobiology (Prof. M. Egelhaaf), Faculty of Biology, Bielefeld University:

06/96 – 09/98 Electrophysiological studies on motion vision in flies

10/98 – 12/00 Author and coordinator in the project RUBIN – educational simulations for computational neuroscience; funded by the UVM-NRW (state government)

01/01 – today Author and coordinator in the project MONIST – a platform for simulations in Brain Science education, developed by 9 Institutes at 6 Universities in Germany; funded by the BMB+F (federal government)

Declaration

I declare to be the only author of this work and have used only those sources and tools that are named.

Erklärung

Ich versichere, daß ich die vorliegende Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Bielefeld, 12/2003

Wolfram Horstmann