ATLANTIS THINKING MACHINES

VOLUME 1

SERIES EDITOR: **KAI-UWE KÜHNBERGER**

Atlantis Thinking Machines

Series Editor:

Kai-Uwe Kühnberger

Institute of Cognitive Science
University of Osnabrück, Germany

(ISSN: 1877-3273)

Aims and scope of the series

This series publishes books resulting from theoretical research on and reproductions of general Artificial Intelligence (AI). The book series focusses on the establishment of new theories and paradigms in AI. At the same time, the series aims at exploring multiple scientific angles and methodologies, including results from research in cognitive science, neuroscience, theoretical and experimental AI, biology and from innovative interdisciplinary methodologies.

For more information on this series and our other book series, please visit our website at:

www.atlantis-press.com/publications/books

AMSTERDAM – PARIS

Enaction, Embodiment, Evolutionary Robotics

Simulation Models for a Post-Cognitivist Science of Mind

Marieke Rohde

Multisensory Perception and Action Group
Max Planck Institute for Biological Cybernetics
Spemannstrasse 41
72076 Tübingen, Germany

With a preface by

Dr. Ezequiel A. Di Paolo
University of Sussex
Brighton, UK

ATLANTIS PRESS

AMSTERDAM – PARIS

World Scientific

Atlantis Press

29, avenue Laumière
75019 Paris, France

For information on all Atlantis Press publications, visit our website at: *www.atlantis-press.com*

Atlantis Thinking Machines

ISBN: 978-90-78677-23-9
ISSN: 1877-3273

For the people of Brighton and the CCNR

Preface

This is an unusual book. It launches a new style of research into the nature of the mind, a style that proficiently uncovers, explores and exploits the synergies between complex systems thinking, sophisticated theoretical critique, synthetic modeling technologies and experimental work. Rather than adopting a grandiose programmatic approach, Marieke Rohde presents us with a pragmatic conjunction of elements, each of them strongly feeding off the others and making it impossible to shelf her work strictly under any one rubric such as psychology, robotics, artificial intelligence or philosophy of mind. Perhaps the least unjust choice is to call this a work of *new cognitive science*.

It is yesterday's news to remark on how our conceptual framework for understanding complex systems is changing. There is a recognized need to supplement the scientific categories of mechanistic, XIX century thought for new ways of thinking about non-linear forms of interaction and inter-relation between events and processes at multiple scales. Since the times of cybernetics and in parallel to the development of the computer as a scientific tool, we have witnessed several proposals for "revolutionary" ways of dealing with complexity: catastrophe theory, general systems theory, chaos, self-organized criticality, complex networks, etc. Despite not always fulfilling their stated potential, these ideas have helped us increase our capability to understand complex systems and have in general left us with new concepts, new tools and new ways of formulating questions.

This conceptual change, however, has not been homogeneous. Only very recently have some of these ideas begun to make some way into mainstream theoretical biology (even if they were contained *in nuce* in the work of many pioneers) for example, in models of protocells and minimal metabolic systems, genetic regulatory networks, embryogenesis, immune networks, to evolutionary and ecosystems dynamics.

It is often the case that the sophisticated theoretical developments necessary to tackle a specific problem (or to reformulate it in a workable manner) have already existed for some

time, but only become acceptable once they are seen to work in the form of conceptual models, synthetic machines, or novel data.

If there is a contemporary domain of enquiry where complexity reigns supreme and where mono-disciplinary linear thinking is bound to fail, this is the realm of cognition. The future history of science will write that today we still know close to nothing about the mind. It will point out that essential categories such as autonomy, agency, values, meaning, intentionality, and many others remain poorly defined at the start of the XXIst century, and that we are, in ways still too preliminary, only beginning to grasp the complexity not only of brains, but of bodily physiologies and mechanisms, of experience, of structured and structuring environments and of social interactions. A proper study of the mind, or for that matter cognition, requires us to get a handle on biological (evolutionary, psycho-physical, neuroscientific), psychological, technological, socio-cultural, linguistic and experiential constraints. Mind is the realm of the über-complex. How to begin to think about it?

According to Rohde, this is how:

We need new concepts

The enactive framework that serves as the basis for Rohde's investigations attempts to explore the relations between life and mind. This approach, with roots in the work of Francisco Varela, has recently become a wellspring of novel conceptual developments, many of which are described and put to work in this book. As examples one could mention a deeper sense of embodiment as rooted in the autonomous organisation of a cognitive being and offering a route towards a naturalisation of normativity and meaning as well as workable novel concepts such as adaptivity and sense-making. Workable, and improvable, as there is no pretension that the last word on the subject has been spoken.

These ideas in combination with the tools of synthetic modelling and dynamical systems theory are the ingredients of the results presented here. The emerging picture is by no means a simple one. It reveals, on the contrary, the subtle complexities and interweaving of factors that are not always easy to isolate. This is perhaps just as it should be – to study, as Rohde does, real cognition. We should suspect any story that renders the complexity of the mind too easily graspable, because that is likely a sign that we have not really made the effort to change our ways of thinking. This is not a self-defeating position – we can indeed understand the mind and develop scientific methods, like the one presented in this book, to this end, but chances are we shall have to radically change our own minds in the process. If this change has not taken place yet, then any understanding will be a façade

that works simply by covering the difficulties with its own blind-spots. Any semblance of an explanation is suspect if all it does is to sweep complexity under the carpet of concepts designed to stop us from asking questions. Such is the case with traditional distinctions between content and vehicles, use of representational language, and widespread functionalist assumptions about the mind – all of which are explicitly or implicitly rejected in this work.

We need new tools

The combined use of dynamical systems ideas and synthetic approaches such as evolutionary robotics allows Rohde to create a micro-loop of scientific enquiry running within her laptop. Pre-conceptions are put to the test by a process capable of generating (under certain constraints) exemplars of the behaviours of interest without having to be too specific about the underlying mechanisms. The result is precisely a process of exploration of mechanisms, which is rendered possible by the use of artificial evolution, a method that is often less biased than our own engineering-laden approach towards designing systems. Thus, evolutionary robotics can be put to the service of generating novel proofs of concept, to question intuitions and overall to exercise our scientific mind and train it in understanding complex embodied and embedded systems. Such models are often deliberately simple thus seeking maximum conceptual impact. In the phrase of Randall Beer, one of the pioneers of this methodology, working with these models is a form of mental gymnastics.
The versatility of this modeling methodology is clearly demonstrated in this book, where it is put to work in conceptual models (querying the logical consistency of the idea of value systems), empirically inspired problems (studying the role of linear synergies in human pointing) and in a direct dialogue with empirical studies where evolutionary robotics models serve the role of hypotheses-generators as well as providing guidance in the dynamical analysis of data. The subtlety with which the present work demonstrates this methodological versatility should not be missed and is likely to set the standard for important forthcoming developments.

We need not shy away from the grand challenges

A common remark about the enactive, dynamical approach to cognition and also about evolutionary robotics is that these ideas and tools sit well at the lower levels of intelligence, where sensorimotor constraints dominate. In other words, they serve us well to understand the intelligence of insects, but we shall find these concepts and methods lacking when dealing with the complexities of the human mind.

This is a real challenge, although at the same time, one could wonder how well traditional approaches have really helped us understand the complexities of the human mind (presumably if they had been successful we should by now have been able to build an artificial person?). Instead, this book takes a practical approach to this question and in doing so it undermines the commonly held assumption that, in order to be useful, a scientific model must match the complexity of the target phenomena that needs to be explained. Rohde provides several clear demonstrations that this is not the case. That it is possible to learn about important factors affecting human cognitive performance (bodily synergies, social perception and time perception) with models that do not nearly match humans in complexity and still can provide us with clear insights about the problem. These models allow us to enter into a concrete dialogue with empirical and theoretical efforts.

We need interdisciplinary dialogue and cross-fertilisation

Rohde does a remarkable job at doing justice to the different disciplines involved in her research moving easily between domains like a Renaissance mind. At no point are concepts trivialised or assumed to map into each other unproblematically. At no point does she render any of the disciplines or methods redundant or secondary to others. A sense of integration, not of unification, and dialogue comes through and hopefully this book will be followed by similar efforts and similar collaborations. Some of the methodological connections she draws, for instance between evolutionary robotics modeling and psychological experiments, are uncommon and probably presented here for the first time.

These are Rohde's proposals for the new sciences of cognition. If they sound idealistic in this preface, this impression will be corrected when the reader is confronted with the practical thrust of her work. This is what gives her book a genuine chance of changing the way we study the mind.

It would be useful to put Rohde's proposal in context. The computational-representational view of the mind, with traditional AI as its theoretical core, has been the target of multiple criticisms during its 6-decade long history. Some of these criticisms have been very insightful and sometimes apparently devastating. On the table of conceptual disputes one wonders how the traditional perspective managed to outlive such attacks. Probably for different reasons (the strong support of funding bodies since the 1960s playing no small part). But importantly due to a genuine, *bona fide* scientific advantage of this perspective: it has been able to provide enough friction to drive a scientific process. Thanks to the computational-representational view of the mind it is possible for psychologists to for-

mulate clear hypotheses about human and animal perception, decision-making, learning, memory and problem-solving and perform clean elegant experiments to probe these hypotheses. Thanks to the computational-representational view of the mind it is possible for neuroscientists to spell out neural function as information processing and turn this idea into the guiding heuristic for experiments and models both at the level of the cellular and biochemical processes as well as on the level of brain anatomical organisation. Thanks to the computational-representational view of the mind cognitive science can define itself as the scientific programme of specifying the functional architecture of intelligence, building new formalisms and deriving their specific implications to be ultimately tested against empirical evidence and by construction in the form of artificially intelligent machinery. In short, thanks to the computational-representational view of the mind, it has been possible for science to move on.

This is what marks the computational-representational view of the mind as a fruitful scientific paradigm, not its theoretical soundness or its in/ability to deal with problems that cognitive scientists choose to ignore (when they arguably should not). Those of us who sustain that this perspective is flawed must still recognise its fruitfulness and be realistic about this oft-neglected fact: the success of a scientific paradigm is not solely judged at the court of logical consistency and empirical evidence. Its maturity, by definition, also lies in the fact that it provides the right set of ideas and tools to tackle the problems that it sets itself, never mind the critical stance that points to its blind-spots as also being genuine and relevant problems.

Consequently, it should not surprise us that the biggest shocks to the system have come less from well argued conceptual attacks than from real practical evidence of its limitations. The connectionist revolution-come-reform clearly demonstrated the benefits of breaking with a rigid understanding of functionality as expressed only in logical rules. It expanded the paradigm without overturning its central tenets. And this was achieved with the help of workable tools leading to better models of brain function, better match with empirical data at least in some domains, and novel technologies for AI and robotics as well as for wider applications.

The Brooksian revolution in autonomous robotics ushered in a more radical break. It explicitly questioned the view of intelligence as complex information processing in the shape of world-modelling strategies. It showed how simple, loosely coupled systems not necessarily organised hierarchically but working in parallel, could achieve real world performance in ways that had been eluding traditional AI engineers. Conceptually, the field

of situated robotics also questioned the understanding of intelligence promoted by traditional cognitive science as human-level, cold reasoning. This achievement, accordingly, is the icing in the evolutionary cake that represents the history of natural intelligence, much of which is opportunistic, affective, embodied and dynamic. Natural intelligence (including much of human everyday intelligence) is fond of "cheap tricks" of local applicability and less prone to general problem solving. In a large part, the questioning of computational-representational views, which has emerged during the last decade, stemmed from this Brooksian revolution. It grew from the ensuing questioning of concepts such as computation and representation and the probing of otherwise abstract ideas like embodiment and situatedness, which were now beginning to yield concrete and measurable results in the fields of autonomous and evolutionary robotics.

Clearly, it is factual, observable changes that fuel conceptual shifts, and as a benefit these changes re-signify existing conceptual criticisms, giving them new tools and techniques to drive a research programme, in other words, to provide an alternative, non-computational framework with the chance to become a new paradigm in its own right.

This is the process that we are witnessing today and which is likely to keep on developing over the next decade. The present book is an example of this process. Recently at a talk at the University of Sussex (in June, 2009), the philosopher of cognitive science Andy Clark was asked for his opinion about the future of cognitive science and philosophy of mind over the next 10 years. He said it would be something dominated very much like what goes on today under the name of enactivism, but "without the silly bits". What better way to get rid of the "silly bits" (assuming there are any!) – and so enter into a new phase of scientific development – than to put these ideas to work and see what they can do? And what better way to start this process than with interdisciplinary research spanning conceptual critique, experimental work, target oriented and abstract modeling in an exemplary methodological dialogue? This is what this book is about – a foray into what's to come in the sciences and technologies of the mind. This is what makes it an unusual book.

Ezequiel A. Di Paolo
University of Sussex
Brighton, UK

Acknowledgments

I would like to thank Miriam Kyselo and Kai-Uwe Kühnberger for comments on a draft of this book. Also, Ezequiel Di Paolo has been a great mentor and friend. Thanks to all my friends, family and colleagues for their support.

Marieke Rohde
Tübingen, October 2009

Contents

List of Figures

Chapter 1

Introduction

"Thinking machines" – the title of the series in which this book is published echoes a long gone optimism. It is the optimism of the cyberneticists, psychologists and mathematicians who invented Artificial Intelligence (AI) as a research program at Dartmouth in 1956, in the face of the powerful new digital technologies. Digital computers proved to be able to do things that previously only humans were able to do: logical deduction, mathematical calculation, syntactical composition of words. Computer programs could meet or even exceed our standards in all those activities that rely heavily on our strong symbolic capacities, that which distinguishes us from mere animals, the pinnacle of our intelligence. Is this what intelligence comes down to? The syntactic manipulation, storage and logical recombination of inputs, of symbols representing the state of the world as we know it through our senses? The idea of the brain as an organic thinking machine and the dream to recreate this machine *in silicio*, as the irrefutable proof of our scientific understanding, became the unifying vision for a new interdisciplinary and scientific study of mind: "cognitive science".

The spirits of cognitive scientists in the 21st century have sobered down substantially. The notion of the "failure of AI" is commonplace. What critics like (Dreyfus, 1972) have been pointing out long since, has now become impossible to deny: reason is not all there is to thinking, and, as far as other skills are required, computers cannot do them very well. By the time I, the author, studied cognitive science early on in the new millennium, the limitations of the 'cognition as computation'-metaphor had become obvious, but were not yet everywhere acknowledged explicitly. The 'embodied turn', a shift in emphasis away from abstract logical properties of thought and towards studying the influence of physics and physiology on mind was only just gaining impact. Today, the point of controversy is not so much anymore *whether* the body shapes the mind, but much rather *how* and *to what extent*.

This book is written from one of the most radical positions possible in favour of this embodied turn. It essentially promotes *enaction* (Varela *et al.*, 1991; Stewart *et al.*, forthcoming) as a candidate for a new paradigm in cognitive science. This paradigm is to be described as a form of *non-reductive naturalism* where mental phenomena and functions *emerge* from a strictly physical substrate. This view is non-reductive in claiming that mental phenomena cannot be reduced to any particular material object or local process, as for instance neural processing. It is naturalistic in that it does not postulate magic or mystical forces to explain the non-reductive character of mind. There are two distinct levels of description – physical mechanism and emergent function – that constrain each other, but one cannot be reduced to or defined in terms of the other. This view relates to the idea of *self-organisation* in physics and complex system theory.

The enactive approach entails a *constructivist* epistemology. In a crude simplification, that means that knowledge is not about veridically representing (in the brain/mind) the objective world, but about the active construction of knowledge through our interaction with the environment and according to viability constraints.[1] In the absence of subjective observers, the environment is filled with an abundance of 'stuff', but not with meaningful objects and events. The things that we perceive, think about and act on are those we need to know about and we choose to know about because they matter. Our brains do not indiscriminately and passively crunch any structure that can be detected in a never ending stream of sensations, sent to us from the outside world. Cognition is a lot about *discarding irrelevant* information and *going out to get relevant information*. The actions we perform are based on our previous inputs and on our intentions and they partially determine our future inputs. This closure of the sensorimotor loop implies that situated cognition is a dynamical system, prone to nonlinear behaviour. Open-loop approaches, restricted to describing input-output mappings, are unable to capture this circular causality and the emergent phenomena it can bring about.

The enactive approach assumes that the physical processes underlying cognition and knowledge construction are self-regulatory with respect to inherent goals or values and that the cognitive processes themselves change depending on success or failure, not only the tokens that are being processed. Computational approaches, presuming that symbolic tokens are processed by a central cognitive program that blindly executes syntactic rules, are insufficient to capture such inherently meaningful self-regulation. They assume an external interpretation ('symbol grounding') of processes that are themselves meaningless and

[1] Viability here does not necessarily mean survival – chapters 2 and 3 explain this point in more detail.

independent of behavioural success. In computational approaches, meaning enters through dedicated channels as yet another symbolic input token. These two issues – i.e., inherent valence vs. external symbol grounding and closed-loop sensorimotor behaviour vs. passive information processing – are possibly the most important points of disagreement between the computationalist and the enactive approach.

The physical system that, as a model, best describes the processes underlying mind in this enactive perspective is the living organism, not the digital computer. The processes that characterise simple life forms (e.g., bacterial self-repair, self-production or gradient following for metabolic integrity, autopoiesis) come much closer to the kind of intelligent process that the enactivist is after: an intertwinement of behavioural and metabolic functions; a dynamical system that constantly changes and exchanges matter and energy with its environment, yet maintains its emergent organisation. No part of the system can explain the global behaviour if examined on its own, no part controls it or defines it, yet what emerges when the organisation is placed into an environment, is a system with a purpose and an identity. The point here is not that all we do as living organisms has to be defined in terms of survival, as in a bacterium. The point is that, if this kind of self-organisation of motion, sensation, behaviour and valence works on a small scale, why would it not work on a larger, more complex scale? Could a multi-cellular organism work according to similar principles? Could a brain self-organise in a similar way as a living organism? Could different such processes interact in complex organisms, such as animals or humans? A bacterial cell does not involve magic, yet it can do what computers still cannot do: it can act according to norms that are inherent in the process, not externally defined. Taking the living cell as a model for cognition, information processing structures do not *a priori* have a place in its explanation, not even as central modules to take care of more abstract tasks.

The discipline of 'Artificial Life' (ALife) took inspiration from this idea. In deliberate opposition to the term 'Artificial Intelligence', this approach aims to synthesise life-like structures to understand and recreate biological intelligence, but without central 'cognitive' computational control. What is the place of 'thinking machines' in this picture? The methods of ALife include real and artificial chemistries (origins of life, proto-cellular life), multi-robot systems (swarm robotics), merely mathematical dynamical models (e.g., Conway's game of life and other cellular automata) and the study of 'intelligent' morphology or materials for robots in order to 'outsource' tasks which intuitively appear to require reasoning to the periphery. If the systems we work with are chemicals, materials or parts of the body, is it appropriate to label the system a 'machine'? Even if intelligent behaviour

emerges from the interaction of simple local processes, if each of the individual units or machines we engineered is 'stupid', where is the thinking? And if our targets for modelling include bacteria and insects, is it appropriate to talk about 'thinking' or 'cognition' in the first place, even if no compositional structure, abstract syntax or consciousness is involved? In enactive cognitive science, everything comes in degrees, including thinking and cognition (life-mind-continuity). If ALife replaces AI in a post-cognitivist cognitive science, the notion of the 'thinking machine' is weakened and extended to the point that it is questionable whether the term 'thinking machine' is very useful at all. In order to stay tractable, ALife modelling approaches are frequently confined to simple life-forms and low-level reactive behaviour, which are not always thought of as cognitive. Followers of the enactive approach that work with higher level cognitive faculties usually work with empirical and conceptual, not with formal or synthetic methods (e.g., cognitive linguistics/anthropology (Núñez, forthcoming; Hutchins, forthcoming), cognitive neuroscience (Le Van Quyen, forthcoming) or phenomenology (Havelange, forthcoming)). Stewart even describes enactive cognitive science as a multidisciplinary project that involves a dialogue "at the very least between psychology, linguistics and neuroscience" (Stewart, forthcoming), a listing that is characterised by a remarkable lack of the disciplines of computer science or AI.

This book carves out a space for computational methods in enactive cognitive science. Based on the author's doctorate dissertation (Rohde, 2008), it presents case studies of how Evolutionary Robotics (ER) simulation models can be used to study cognition using computational methods in a *bona fide* enactive spirit. Furthermore, unlike most ALife models, the models in this book are applied specifically to problems of human cognition and behaviour. The book alternates between concrete examples and the overarching methodological main question: *how can simple ER simulations be used to explain human level cognition?*

Even though the argument is developed using ER simulation models, which is a typical ALife technique, in a larger context, the methodological conclusions drawn hold for synthetic and modelling techniques in general. Going through an iconoclastic crisis of rejecting the computational metaphor and discarding the dream of the 'thinking machine', the enactive computer scientist has to work with what is left, reconstructing her niche. If computational models can be useful for any other science, why should they not be useful for the science of mind? This book proposes to drop the concept of the 'thinking machine' in favour of the concept of computational models as 'machines for thinking' (or 'tools for thinking'), a status that models and simulations holds in other sciences, too. What makes

the case of cognitive modelling special is that here we are dealing with 'machines for thinking about thinking'. This can easily lead to the confusion of the *explanans*-thinking with the *explanandum*-thinking, whereby one can easily slip back into a computationalist stance, believing that the model of a cognitive faculty or phenomenon is indeed a 'thinking machine', rather than just a model, a machine for thinking about a system that thinks. Therefore, conceptual hygiene is one of the most important virtues for an enactive cognitive science.

This book starts off with two conceptual chapters. Chapter 2 is an introduction to the paradigmatic struggle in cognitive science. It begins by giving a historical account of the birth, rise and decline of cognitivist-computationalist cognitive science. It then introduces some of the main proposed alternatives and clarifies how they differ from each other and from the computationalist paradigm. Then, the enactive paradigm is outlined and advocated in more detail. In the context of criticism, the question of computational methods in post-cognitivist cognitive science, which is briefly sketched previously, is reposed.

The longest chapter in this book is the method(ological) chapter 3. It not only introduces the techniques used for the research presented. It also presents work on a number of science-theoretic questions, such as the consequences of a constructivist world-view for scientific practice and interpretation, the role of Dynamical Systems Theory in the enactive approach and the methodological difficulties associated with the scientific study of experience. It concludes with the outline of how minimal experimental and modelling (ER) approaches can be integrated to form an interdisciplinary framework to address questions of perceptual experience from the enactive perspective.

The following four chapters present concrete results on scientific problems of different kinds, to illustrate the use of Evolutionary Robotic simulations in enactive cognitive science:

Chapter 4 presents a simulation model studying linear synergies as a principle in motor control. The problem of redundant degrees of freedoms and the concept of motor synergies are introduced, as well as the experimental study that inspired the simulation model. The results are evaluated in terms of what they imply for the study of motor control and for the use of ER simulation models in cognitive science.

A simulation model caricaturing architectures that implement an internal value system to self-supervise learning is presented in chapter 5, in order to illustrate the implicit premises underlying this kind of approach. This chapter also discusses the problem of values, as it had been sketched in the previous outline (i.e., value as dedicated value signal vs. value as

intrinsic property of a physical process). Again, the model and its results are presented and evaluated with respect to the question the model addresses as well as with respect to the methodological theme of the book.

Chapters 6 and 7 present the results from two simulation models of two subsequent and very related experimental studies on human perceptual crossing in a one-dimensional (chapter 6) and a two-dimensional (chapter 7) minimal virtual environment. These chapters implement the combination of ER modelling and minimal behavioural experiments on human perception proposed in chapter 3.

A conceptual interlude on time cognition and time perception is given in chapter 8. It analyses a broad variety of literature on time and temporality, including Kant's epistemology, Husserl's and Merleau-Ponty's phenomenology, Lakoff and Núñez anthropology, Piaget's developmental psychology, Varela's neurophenomenology, Shanon's study of altered states of consciousness, Libet's neuroscientific work on neuro-behavioural latencies and work on the psychophysics of time perception. This chapter prepares for the following three chapters that investigate the phenomenon of sensorimotor recalibration of perceived simultaneity presenting experimental and modelling results.

Chapter 9 presents an experimental study on human adaptation to sensory delays. There is evidence that adaptation to increased sensorimotor latencies can lead to a recalibration of simultaneity in some situations, but not in others. The hypothesis that time pressure in the task is the crucial factor for this recalibration to take place is tested and not supported by the data. The experiment is the basis for the ER simulation model presented in the following chapter 10, which provides new insights in the sensorimotor processes involved in delay adaptation and how they may relate to recalibration of perceived simultaneity. The ideas of how simultaneity is constructed from regularities in our sensorimotor flow are presented in chapter 11, which also takes into consideration the theoretical analyses from chapter 8.

The conclusion from this collage of ER simulation models is that there is an abundance of areas of applicability for simple simulation models in an enactive cognitive science of human level cognition. These range from down-to-earth applications to motor control (chapter 4) to very high-level philosophical proofs of concepts (chapter 5). Furthermore, by taking an experimental-behavioural approach to human perceptual experience that is as minimal as the model itself, ER simulations can enter a fruitful dialogue with such empirical techniques. This approach has been applied to the perception of agency (chapters 6 and 7) and the perception of simultaneity (chapters 9-11). Chapter 12 draws an optimistic conclusion: even though abandoning the idea of the 'thinking machine' may be painful at

first, it opens up new possibilities to use computational techniques in the non-reductionist enactive study of cognition that is not blind to the fact that we are living organisms, too. Computational methods can bring formal rigour to our thinking about thinking, without the overconfident ambition to turn thinking into a formal business altogether.

Chapter 2

Enactive Cognitive Science

This opening chapter introduces the philosophical and paradigmatic context in which the research presented in this book has been generated. It forms the foundation for the description and development of the methods employed and developed (chapter 3) and their later application (just modelling: chapters 4-7; combined modelling and experimental work: chapters 8-11). The significance of the results of each of the models and experiments for the particular research question they address is discussed within the respective chapters. The paradigmatic and methodological implications of these studies, which are the unifying research theme for the present work, are identified and evaluated in chapter 12.

In many ways, the methodological research question underlying this book can be seen as yet another episode of the decade-old paradigmatic struggle between traditional computationalist cognitive science and more embodied and dynamic approaches. Therefore, this chapter starts (Sect. 2.1) with a summary of the key issues, persons and milestones that have determined this debate, which is as old as cognitive science itself. In cognitive science, there is a tendency to present the paradigm struggle as a black-and-white battle between the traditional 'GOFAI' (good-old-fashioned Artificial Intelligence; Haugeland, 1985) approach, on the one hand, and everything which is '$\neg$ GOFAI' (or 'New AI'), on the other hand. Various alternative proposals (Connectionism, Dynamicism, Behaviour-Based Robotics, ...) have originated from the observation of similar shortcomings of the traditional paradigm and often have significant methodological and ideological overlap. However, they cannot be seen as a single alternative that comes in different flavours: significant tensions exist between them. Section 2.2 summarises a number of alternative paradigms, identifies their maxims and core assumptions and points out in how far they are prone to the same criticisms as GOFAI. Section 2.3 presents the enactive approach as a candidate for a new paradigm in cognitive science and that underlies the research presented in this book. Finally, Sect. 2.4 reflects on the main challenges this new paradigm faces and on the

role computational models can play in it. Special attention is paid to a criticism that dynamical modelling approaches frequently face, i.e., that such models serve well to address low-level behavioural issues but not high-level cognitive issues. This last section finishes by outlining the scientific challenge that has driven the research presented in this book, i.e., to identify ways to use simple Evolutionary Robotics (ER) simulation models in cognitive science in general and, in particular, for the scientific study of human cognition.

2.1 The Rise and Fall of Traditional Cognitive Science

To my knowledge, it is not clear when the term 'cognitive science' was first employed. Its birth is, however, frequently associated with the birth of a more traceable term, i.e., 'Artificial Intelligence' (AI; e.g., Eysenck and Keane, 2000; Haugeland, 1981; Russell and Norvig, 1995), a label that has first been used in the call for the Dartmouth Conference in 1956 (McCarthy *et al.*, 1955). This conference brought together researchers that were employing the then newly emerging digital computer technology in disciplines as different as psychology, computing, linguistics, neurobiology and engineering.

At the time, Behaviorism was at its peak in psychology. Behaviorism had arisen out of a partially justified methodological scepticism towards introspectionism in psychology, whose data was not observable by anyone but the introspecting subject and thus did not meet the scientific standards of the natural sciences. Therefore, the behaviourists demanded to confine scientific inquiry to physically measurable behaviour. The most radical critics went as far as to claim that mind and mental phenomena "could not be shown to exist and were therefore not proper objects of scientific inquiry at all" (Stilling *et al.*, 1998, p. 335) and the very use of mentalistic language was, as a consequence, frowned upon.

The analogy between computing processes in digital computers (or formal Turing Machines, TMs) and the human mind drawn by the researchers in the newly founded discipline AI, therefore, fell on fertile grounds with scientists that were interested in studying mental phenomena. Digital computers perform intelligent tasks that previously only humans could do, such as logical reasoning, mathematical computation, syntactically correct chaining of words, *etc.* If we can physically explain and formally and functionally describe how the machine does it, why would the same not be possible for the human mind, the 'black box' of Behaviorism? Computer technology and AI provided the language and concepts that, in the oppressive scientific climate at the time, made it acceptable to use mentalistic terms without falling subject to accusations of lacking scientific rigour.

The science of cognition, rather than the science of 'just' behaviour, therefore, is frequently defined in terms of this metaphor of the digital computer for the human mind. This metaphor comes in different variants, e.g., physical symbol system hypothesis (Newell and Simon, 1963), computational theory of mind (Fodor, 2000) or cognition as computation or information processing (Stilling *et al.*, 1998, p. 1). It became the underlying dogma for the interdisciplinary study of the mind, in which cognitive psychologists and linguists empirically measure the behavioural data to be modelled; computer scientists and AI researchers generate the computational models of this data that map inputs to outputs and predict further not yet measured input-output mappings; brain scientists identify the neural circuits and brain areas that instantiate these formal models; philosophers take care of the mental side of things and relate the formal and scientific results to mind, which is scientifically not measurable. Had it worked, it would have been a great idea.

The problem with the mind-as-machine metaphor is that neither the human mind nor the human brain are very much like digital computers. A digital computer is a device that maps input symbols to output symbols following syntactic rules, and, even though humans are much better at performing such mappings than most animals, it is by far not everything they do. Computers can model those aspects of our behaviour that are syntactic *in their nature*, but such behaviours are but a very limited subset of the things we do. Consequently, over the last 50 years, cognitive scientists have repeatedly run up to the limits of this metaphor. This led to the identification of a whole catalogue of problems that can ultimately be traced to originate from the mind-as-machine metaphor. A non-exhaustive list features:

- The frame problem in AI: how to keep track of everything that does not change in response to my actions? (e.g., Russell and Norvig, 1995)
- The credit assignment problem in search and machine learning: in solving a complex problem, which of the many steps taken were relevant to obtain behavioural success? (e.g., Minsky, 1961)
- The symbol grounding problem in philosophy: how do symbols get their meaning? (Harnad, 1990)
- The binding problem in neuroscience: how are features that are processed in different channels or parts of the brain brought together to form one coherent perception of the world? (e.g., Revonsuo and Newman, 1999)
- The problem of context in formal semantics: how do I functionally derive word meanings that depend on the situation in which they are expressed? (e.g., Cole, 1981)

All these problems are a consequence of having separated the symbolic representational token from its meaning, a separation which characterises computational systems (cf. Haugeland, 1981). Local structures do their job, applying syntactic rules without knowing if they are playing chess or launching a nuclear bomb. This ignorance of the algorithm is beautifully illustrated in (Searle, 1980)'s famous 'Chinese room' thought experiment, which features a Chinese interpreter that applies the rules of Chinese language without knowing any Chinese.

This is just one, and perhaps the most drastic implicit premise contained in the computational metaphor. A number of assumptions about brain and mind that are not supported by empirical evidence piggyback on this premise – assumptions that have been vehemently criticised over and over in the 50 year old history of AI (e.g., Dreyfus, 1972; Pfeifer and Scheier, 1999; Harvey, 1996; Port and van Gelder, 1995). Those include the idea that exact timing does not matter, that the brain/mind is strictly functionally modularised, that inputs are passively parsed rather than actively sought, that there is an external world of objects, waiting to be represented and that explanatorily atomic homunculi provide meaning wherever it is lacking. Such problematic implications of computational views are discussed later on in this chapter, throughout this book and in many of the references provided.

However, the point of this section is not in the first place to convince the reader that there are problems with the computational metaphor. The 'failure of AI', as it is commonly called, is by now acknowledged even by some of the most central and radical defendants of the computationalist paradigm in cognitive science (e.g., Fodor, 2000). However, the methods, and with them the language, the concepts, the modelling assumptions and the rejection of other ways of doing cognitive science prevail. Having started as a rebellion against the constraints that Behaviourism imposed on language, thought and action, computationalist cognitive science has now itself become an intellectual straightjacket, an obstacle in the way of scientific progress and the understanding of mind. While the mind-as-machine metaphor provided the language to describe cognitive processes that are syntactic in their very nature, it did not provide the language to talk about semantics, about meaning. This is a problem, because, as the enactive approach argues, mind is an inherently meaningful phenomenon. Even worse maybe, the metaphor took away the language to talk about behaviour or anything external to the former black-box of Behaviourism, because it presumes that internal representation and symbol manipulation, the formal description of the mind-machine, is all there really is to know.

Possibly, both Behaviorism and computationalist GOFAI cognitive science have been so successful because they are based on seductively simple ideas. Is it time to replace computationalism with another seductively simple idea? Probably not. The world is complex, mind is complex, the brain is complex, the body is complex. Any simple theory will be doomed to follow the same destiny, i.e., to rise, to turn into dogma, and to fall, but not to explain cognition. Fortunately, the enactive approach is not simple. Section 2.3 tries to capture the essence of what this new and still dynamic and evolving approach takes from different predecessors, some ancient, some more recent, and how it aspires to explain mind. Before that, some of the main alternatives proposed as alternative paradigms are reviewed, to be able to argue in how far the enactive approach is similar or different.

2.2 Alternative Paradigms

Sceptics have pointed out the limitations that are summarised above over and over again. But does giving up on computationalism imply giving up on the idea to scientifically explain mind and cognition? Or are there ways to cut out the mind-as-machine metaphor but to keep cognitive science as such an interdisciplinary project? Many proposals have been made to substitute the mind-as-machine metaphor with a new and different paradigm to be programmatic for a new cognitive science.

There is a tendency to perceive such alternative proposals as a unified 'opposition', rather than as the diverse set of paradigms that it is. For instance, in *Connectionism, artificial life, and dynamical systems: New approaches to old questions*, (Elman, 1998) presents three alternative paradigms to the computationalism and describes how he believes they go hand in hand:

> "The three approaches share much in common. They all reflect an increased interest in the ways in which paying closer attention to natural systems (nervous systems; evolution; physics) might elucidate cognition. None of the approaches by itself is probably complete; but taken together, they complement one another in a way which we can only hope presages exciting discoveries yet to come" (Elman, 1998).

Earlier on, Elman writes:

> "While there are significant differences among these three approaches and some complementarity, they also share a great deal in common and there are many researchers who work simultaneously in all three" (Elman, 1998).

The proposal here is that Elman's take on the situation is misconceived. However, his misconception is common, which is rooted in two facts that Elman also observes: (1) Alter-

native approaches tend to be driven by a common demand for more biological plausibility and (2) there is methodological overlap between alternative paradigms. However, using the same methods and coming from the same origin does not imply compatibility. Identifying the largest common denominator between different paradigms bears the danger of watering down the original radical and new proposals and dilute them "into a background essentially indistinguishable from that which they initially intended to reject" (Di Paolo *et al.*, forthcoming). Therefore, the ideological commitments associated with some alternative paradigms that are all subsumed under the umbrella term *new AI* have to be clarified.

2.2.1 *Connectionism*

Connectionism is frequently seen as the most important alternative proposal to GOFAI. This perception probably relates to the fact that Connectionism had been posed as an explicit challenge to logic-based approaches quite early on (McClelland *et al.*, 1986) and that Artificial Neural Network (ANN) theory had been developed alongside logic-based AI. Connectionism (or parallel distributed processing, PDP) proposes to replace GOFAI's digital computer with "a large number of simple processing elements called units, each sending excitatory and inhibitory signals to the other units" (McClelland *et al.*, 1986, p. 55). Benefits of this approach are its "physiological flavour" (McClelland *et al.*, 1986, p. 55) because ANNs are inspired by neuroscience, drawing the analogy between processing units and biological neurons. A lot of the paradigmatic debate in cognitive science has focused on identifying the differences between Turing Machine/logic based approaches and ANNs and their implications.[1]

From an enactive perspective, ANNs are only interesting as one among many formal tools, not as a modelling paradigm that is intrinsically more biologically plausible. In their non-dynamic form (i.e., feed-forward networks), they only represent input-output-mappings just like computationalist models. In their dynamic form (i.e., recurrent networks), they can represent dynamical systems – however, that a dynamical system takes the form of an ANN rather than just any differential equation is not of explanatory importance either. As argued extensively elsewhere (e.g., Cliff, 1991; Harvey, 1996), Connectionism suffers from most of the problems associated with the computationalist paradigm. Indeed, it is just a variant of the computational paradigm, not presuming 'cognition as digital information processing' but rather 'cognition as parallel distributed processing'.

[1]Noticeably: (Fodor and Pylyshyn, 1988)'s conceptual criticism and the responses it triggered; (Minsky and Papert, 1969)'s formal proof of limited computational capacities of perceptrons.

ANN theory has produced some very useful formal tools, learning algorithms and representations for dynamical systems and mathematical functions. At its interface to theoretical neuroscience, it has also generated models that contribute to the understanding of brain physiology and dynamics. However, in order to understand mind, cognition and behaviour, it is necessary to investigate not just what comes in and what goes out, but much rather what happens in closed loop interaction with the world and how such physical agent-environment interactions relate to experience. ANN theory is not at the heart of such a project, it is not even an essential component.

2.2.2 *Dynamicism*

The dynamical hypothesis in cognitive science (van Gelder, 1998; Port and van Gelder, 1995) is a more recent alternative proposal, based on the claim "that cognitive agents are dynamical systems" (van Gelder, 1998, p. 615). The problem with this approach is, again, that a mathematical formalism to substitute GOFAI's Turing Machine is proposed, rather than to part with the idea that a formal tool has to be at the core of cognitive science in the first place. This idea is at tension with the non-reductive nature of cognition proposed in the enactive approach, with the idea of emergence and with the emphasis on lived experience and inherent meaning (cf. Sect. 2.3).

Dynamical Systems Theory (DST) does play an important role in the enactive approach, and this methodological importance is elaborated in the following methodological chapter (Sect. 3.2). However, there are models that are not enactive but fall within the realm of Dynamicism. (Elman, 1998) presents, as an example of a DST model, a recurrent neural network that is trained to recognise the context-sensitive formal language $a^n b^n$, which he sees as an example of a dynamical model of "realms of higher cognition" because it is "applied to the case of language" (Elman, 1998, p. 30). In the light of the previously identified problems with the computational paradigm, it is mysterious what this completely disembodied model (which basically represents a pushdown automaton) can explain about cognition: why is this model superior to a TM recognising the same formal language, or how does it not fall victim to the same criticisms?

This example illustrates where the methods of Dynamicism and Enactivism part, despite the overlap. The answer to the problems identified with the computationalist paradigm cannot be in the appropriate choice of formalism alone, and this point does not just concern the explanatory power of any particular model, but also the status of simulation models within cognitive science as a principal concern: a formal model in cognitive science cannot

explain but an aspect of the *explanans*, it cannot itself be the phenomenon (cf. Sect. 3.3). Even though DST is of crucial importance for an embodied and enactive cognitive science, it is not in itself a satisfactory new paradigm.

2.2.3 *Cybernetics, ALife, Behaviour Based Robotics*

While Connectionism and Dynamicism focus their criticism of the computationalist approach on the properties of the formalism used for modelling, both Behaviour Based Robotics (BBR, e.g., Brooks, 1991) and Artificial Life (ALife, e.g., Langton, 1997) emphasise the importance of embodiment and situatedness of cognition. The computationalist paradigm focuses on what comes in and what goes out but fails to account for how what goes out impacts in turn on what comes in (i.e., the *closure of the sensorimotor loop*) and its relevance for explaining cognition.

Associated with these approaches is a strong scepticism of the objectivist assumption implicit in computationalism, i.e., that the brain builds an internal representation of the external world which justifies to exclude the world itself from the explanation of cognition in favour of a Cartesian theatre. As (Brooks, 1991) puts it: "the world is its own best model". This sceptical position is frequently called *anti-representationalism*, even though I am not aware of anyone adopting this label for themselves. (Harvey, 1996), however, appropriately remarks that from being the 'billiard balls' of explanation in computationalism (i.e., part of the *explanans*), human capacity to represent becomes an *explanandum* in non-computationalist paradigms and, though intriguing, loses its central role in explanation. He also points out that there are very different and partially contradictory meanings associated with the term 'representation' in cognitive science and everyday life (e.g., correlation, stand-in, re-presentation, something mental, something in the brain, a computational token, ...) and that computationalists are frequently reluctant to define their usage of the term. Therefore, the term is problematic and ambiguous and bears potential for misinterpretations. Followers of embodied and situated approaches, therefore, are frequently reluctant to use the term as part of their explanations of how the mind works in the closed-loop.

Both BBR and ALife emphasise the fact that living organisms differ in that respect from digital computers, i.e., they exploit the dynamics of closed-loop interactions with the environment. ALife can be seen as a direct counter-proposal to GOFAI that focuses on explaining "life as it is and how it could be" (Langton, 1997) rather than 'intelligence' which is associated with logic, rationality and the kinds of things that computers are good at. These synthetic approaches clearly have their predecessors in the cybernetics move-

ment that started during the first half of the last century (e.g., Ashby, 1954; Braitenberg, 1984; von Holst and Mittelstaedt, 1950; Holland, 2002, (Holland on Walter's work from the 1940s/1950s)), whose aim can maybe be described as explaining living organisms as machines (not as Turing Machines (!)) using the formal language of control theory. Brooks sees his BBR approach in direct succession to the cybernetics movement, whose limitations he diagnoses to be due to the limited technologies and formal tools available at the time (Brooks, 1991). The early work in cybernetics is a major source of inspiration for behaviour-based and ALife approaches.

Naturally, there are also a multitude of opinions and disputes within the BBR and ALife community, e.g., about whether simulation models really count as embodied and situated, whether energy constraints are essential, *etc.* As common denominator, BBR and ALife, in continuation of early cybernetics ideas, presume that behaviour has to be studied, modelled and synthesised in closed loop agent-environment interaction. As argued in Sect. 2.2.5, there is no direct contradiction between this paradigm and the enactive approach.

2.2.4 *Minimal Representationalism and Extended Mind*

There have been a number of proposals that explicitly aim at reconciling the old computationalist paradigm with the growing group of critics becoming aware of the need to take embodiment, situatedness and real-time interaction dynamics seriously. As we assess:

> "In the opinion of many, the usefulness of enactive ideas is confined to the 'lower levels' of human cognition. This is the 'reform-not-revolution' interpretation. For instance, embodied and situated engagement with the environment may well be sufficient to describe insect navigation, but it will not tell us how we can plan a trip from Brighton to La Rochelle. [...] For some researchers enactive ideas are useful but confined to the understanding of sensorimotor engagements. As soon as anything more complex is needed, we must somehow recover newly clothed versions of representationalism and computationalism" (Di Paolo *et al.*, forthcoming).

Main proponents of this kind of approach include (Clark, 1997), (Clark and Grush, 1999) and (Wheeler, 2005). These approaches aim at incorporating syntactic symbol manipulation processes into an embodied and situated story in order to account for high-level human reasoning. The proposal is thus to abstain from the chauvinism associated with traditional computationalism (i.e., that a TM description will give you the whole story). However, such approaches extend the computationalist program, rather than to fundamentally change it: there will be some need to refer to dynamical, bodily and environmental variables, but at some level, cognition is and has to be still a homuncular symbol manipu-

lation process working on internal representations. Those cognitive capacities presumed to be thus implemented are called "representation hungry problems" (Clark, 1997).

The model of value system architectures presented in chapter 5 illustrates some of the conceptual problems associated with such hybrid architectures and homuncular modules. Problematic though these proposals may sound, they have to be taken seriously because they point towards the main challenges for an enactive cognitive science. There are, at present, not many enactive accounts of cognitive activities that involve the use of symbols (such as language, mathematics or planning). Dynamical systems accounts frequently focus on cognitive capacities that are strongly rooted in the here-and-now, which leads cognitivists to believe that this is all these accounts can offer. Anthropological work on language (e.g., Núñez and Sweetser, 2006; Lakoff and Johnson, 2003) or mathematics (e.g., Lakoff and Núñez, 2000) takes first steps to fill this gap. However, from the domain of embodied computational modelling, there have been little contributions towards explaining such symbolic cognitive phenomena.

As outlined in Sect. 2.4 below, for the enactive approach, this gap is not a failure but a challenge. There are no principal limitation, no catalogue of theoretical problems as those listed for the computational approach earlier on. Instead, there are horizons towards which this young paradigm can venture out next. For the present purpose, it is only important to point out that, in suggesting that human symbolic reasoning has to be a minimal form of symbolic digital computation, hybrid or 'on the fence' positions are not variants of the enactive paradigm but, if at all, variants of the computationalist paradigm.

2.2.5 *Methodological Overlap, Ideology Worlds Apart*

From the previous summary, it is easy to understand how alternative paradigms can get shuffled up: the shortcomings they aim to mitigate and the methods they propose overlap remarkably. However, as concerns the science-theoretic side of things, there are important differences and even contradictions between all these paradigms.[2] Most of them put too much emphasis on the descriptive formalism, just as computationalism does.

Within the described landscape, the approach proposed in this book acknowledges a substantial methodological overlap, but rejects most of the labels just mentioned. For instance, even though the work presented uses both ANNs and DST as formal tools, the work should not be labelled Dynamicist, and even less Connectionist, because descriptive formalisms

[2] At least as they are phrased by their most radical proponents; many researchers applying the mentioned methods and labelling themselves accordingly are highly respectable, produce great contributions and are usually more modest or less chauvinistic about their choice of method.

are not central to the underlying enactive paradigm, which goes far beyond formal issues, whereas they are at the core of both Connectionism and Dynamicism. As concerns ALife as a paradigm *for AI*, the label is appropriate: ALife's closed-loop modelling approach is the way forward for modelling the kinds of phenomena addressed. The disclaimer to be added is that ALife as synthetic paradigm is not the same as ALife as a paradigm *for cognitive science*. Even though the enactive paradigm in cognitive science has a space for synthetic methods in which ALife simulation modelling fits, modelling or synthetic recreation are not central to Enactivism. The following methodological chapter (Sect. 3.3) elaborates on the status of formal tools and methods within enactive cognitive science, which is introduced in the following section.

2.3 The Enactive Approach

The term 'enaction' in the context of cognition is usually associated with the publication of *The Embodied Mind* (Varela *et al.*, 1991) and the editors, i.e., Francisco Varela, Evan Thompson and Eleonor Rosch, as key proponents, even though the term has been used in related contexts before (cf. Di Paolo *et al.*, forthcoming, section 2). The research and method proposed in this book stands very much in the tradition of the interdisciplinary research program put forward by (Varela *et al.*, 1991), which may be construed as a kind of non-reductive naturalism, emphasising the role of embodied experience, the autonomy of the cogniser and its relation of co-determination with its world. In this section, the interpretation of the enactive approach underlying this book is outlined. This outline is in large parts a recapitulation of the positions we put forward in (Di Paolo *et al.*, forthcoming). As dissatisfaction with the classical computationalist paradigm grows, the term 'enactive' gains in popularity. In the light of the paradigmatic confusion sketched in the previous Sect. 2.2, there is a clear danger that the enactive approach as a paradigm is watered down, becomes a meaningless umbrella term or falls victim to self-contradiction. Therefore, the ideological commitments characterising this approach have to be made explicit. However, as the enactive approach is still emerging and developing, it is also important to avoid simplification, reduction and rushed exclusion of promising routes towards an open future. We write

> "[. . .] in trying to answer the question 'What is enactivism?' it is important not to straight-jacket concepts that may still be partly in development. Some gaps may not yet be satisfactorily closed; some contradictions may or may not be only apparent. We should resist the temptation to decree solutions to these problems simply because we are dealing with definitional matters. The usefulness of a research programme also lies with its capability to

> grow and improve itself. It can only do so if problems and contradictions are brought to the centre and we let them do their work. For this, it is important to be engendering rather than conclusive, to indicate horizons rather than boundaries" (Di Paolo *et al.*, forthcoming).

The collection (Stewart *et al.*, forthcoming) in which the cited contribution appears is an important step towards such an 'emancipation without dogmatisation'. We identify five central and conceptually intertwined concepts that constitute the core of the theory of enaction (Varela *et al.*, 1991; Thompson, 2005), i.e., autonomy, sense-making, emergence, embodiment and experience, five ideas that partially imply each other and that are outlined in the following.

2.3.1 *Autonomy*

Being autonomous means to live by your own rules, as the etymology of the term already suggests ('auto' means self and 'nomos' means law in Greek). The theory of autopoiesis (Maturana and Varela, 1980) argues that living organisms are autonomous because they constitute and keep building themselves and maintain their identity in a variable environment. This means that, at some level of description, the conditions that sustain any given process in a network of processes are provided by the operation of the other processes in the network, and that the result of their global activity is an identifiable unity, as it is best exemplified by the autonomy of the living cell.

Three things are important to realise about this idea of biological autonomy.

(1) The recognition of the agent as constructing, organising, maintaining, and regulating sensorimotor interaction with the world is in direct opposition to a representationalist perspective in which agents mechanically represent and react to a world with a pre-given ontology of meaningful objects.

(2) The constraints imposed on self-maintaining processes of identity generation are of *mechanical* nature. Living organisms are bound by the laws of physics but the possibilities to re-organise themselves and, with them, the world of meaningful interactions they bring forth, are open-ended. This open-endedness contrasts with explicit design of adaptive circuits in computationalist approaches, e.g., in the discipline of machine learning. Even if machine learning is a blossoming field as part of software engineering, such algorithms are *functionally* constrained by in-built rules.

(3) Against a common prejudice, autonomy does not equate to maximal moment to moment independence from environmental constraints (e.g., Bertschinger *et al.*, 2008; Seth, 2007). It means, contrariwise, "being able to set up new ways of constraining

one's own actions" (Di Paolo *et al.*, forthcoming), an idea we elaborated in (Barandiaran *et al.*, 2009).

The living cell may be the best example for biological autonomy, but, arguably, it is not the best example for the importance of autonomy in the scientific study of cognition. The cognitive capacities of cells, if you want to call them cognitive at all, are very limited. How autonomous identity preservation can happen at many possible levels, not only on the metabolic level, is elaborated in Sect. 5.5 of this book, which draws on some of Varela's conceptual work along similar lines (cf. Varela, 1991, 1997). Against another common prejudice, the enactive approach is not obsessed with bacteria cognition; Varela's late work was much more centred on the investigation of neuro-cognitive autonomy and human conscious cognition (e.g., Varela, 1999; Rodriguez *et al.*, 1999), and there are recent and interesting proposals that self-sustaining metabolism is altogether insufficient to give rise to mind or intentionality, which instead is postulated to result from self-sustaining closure at the behavioural or neural level ('Mental Life'; cf. Barandiaran, 2007). Such contemplations of neuro-cognitive identity and autonomy are contingent on the question whether or not such 'Mental Life' could exist without an organismic metabolic substrate, which is an open research question.

2.3.2 *Sense-Making*

The concept of sense-making is closely related to the concept of autonomy – it emphasises the constructivist and epistemological component in the enactive approach. In so far, "Enactivism thus differs from other non-representational views such as Gibsonian ecological psychology on this point (Varela *et al.*, 1991, p. 203-4). For the enactivist, sense is not an invariant present in the environment that must be retrieved by direct (or indirect) means. Invariants are instead the outcome of the dialogue between the active principle of organisms in action and the dynamics of the environment" (Di Paolo *et al.*, forthcoming).

As John Stewart remarked (in a plenary discussion at ARCo2006 in Bordeaux): the problem with information is not that there is not enough out there; the problem is that there is too much of it. There are infinite, countless invariances that could be detected and represented. Those *relevant* to the cogniser are those that are perceived, and what is relevant depends on the cogniser's organisation. The formation and perception of concepts, in turn, can alter the autonomous organisation of the cogniser, which can lead to the construction of

new meanings or the destruction of existing meanings. Cognition therefore is a *formative* activity, not the extraction of meaning as if this was already present.[3]

To realise this constructive role of the cogniser helps to disarm another common accusation, which is that the enactive approach is non-naturalistic, solipsistic or denies the existence of an external world. In the first place, the only thing denied is the observer-independent existence of meanings and secondary qualities – not the existence of a universe of meaningless matter, physical constraints and external forces outside our control. The notion of constructivism here adopted is a pragmatic one: how can a constructivist perspective be put to work to scientifically understand cognition? Further reaching question of the ontological implications of enactivism are not at the centre of this book, nor are they directly relevant to how the work presented is to be interpreted.

2.3.3 *Emergence*

In order to illuminate the concept of emergence, the example of the living cell is recalled. How do we know the cell is alive? And what exactly is alive? "The property of continuous self-production, renewal and regeneration of a physically bounded network of molecular transformations (autopoiesis) is not to be found at any level below that of the living cell itself" (Di Paolo *et al.*, forthcoming). It seems ill-conceived to call any of the component parts (a protein, the DNA strands, *etc.*) alive: these are just physical structures that can be isolated, the material substrate of the living cell that is constantly changed and renewed. It is undeniable, however, that the phenomenon of life is as real as it could be.[4]

We can very well scientifically investigate the material substrate of the living, and how it brings about relational properties such as 'life', 'death' or 'survival', without ever being able to (or wishing to) reduce them to the physical substrate. In the same sense, we can scientifically study the physical processes from which mind and meaning emerge. The latter are then not to be reduced to physical components of either the agent or its environment, but belong to the relational domain established between the two.

Thies central concept of emergence is at the root of enactive scepticism towards functional localisation as it is practised in traditional cognitivist psychology, AI and neuroscience. The problem is not that there would be sufficient evidence for a correlation or not. It is

[3]Some approaches that assume the label 'enactive' (e.g., Noë, 2004) seem to downplay/neglect this inherent meaningfulness of cognition and behavioural processes and focus instead on the issue of closed-loop sensorimotor dynamics. The position put forward in this book, however, sees this aspect as crucial and follows, in this sense, the original proposal of the enactive approach in (Varela *et al.*, 1991).

[4]Even if this seems to be forgotten by some modern biologists, as (Stewart, 2004) argues.

much rather that this kind of reductionist assignment is a category mistake. This question is explored further in chapter 5.

2.3.4 *Embodiment*

Embodiment is a concept widely discussed and valued in cognitive science. Therefore, the argument will not dwell too much on the dated idea that cognition was the meat in the 'classical sandwich' (Hurley, 1998), squashed between the negligible bread of peripheral sensor and motor systems that generate symbolic representations and execute symbolic motor outputs.

Instead, the important difference between embodiment and mere physical existence is brought to the reader's attention. "[A] cognitive system is embodied to the extent to which its activity depends non-trivially on the body. However, the widespread use of the term has led in some cases to the loss of the original contrast with computationalism and even to the serious consideration of trivial senses of embodiment as mere physical presence – in this view a word-processor running on a computer would be embodied, (cf. Chrisley, 2003)" (Di Paolo *et al.*, forthcoming). Embodiment is not 'symbol grounding' (Harnad, 1990) through implementation, an idea that keeps up the Cartesian separation between cognition and 'reality'. Much rather, embodiment means that cognition *is* embodied action, in that the sensorimotor invariances our body affords in interaction with this world constrain and shape the space of meanings constructed.

2.3.5 *Experience*

Steve Torrance (personal communication) remarked that experience is an 'embarrassment' for the computationalist approach: a full blown cognitivist architecture, which supposedly explains cognition, fails to account for one of the most central feats of the mental, i.e., what it feels like. With decades having passed since Behaviourism, the 'c-word' (consciousness) has become less and less of a taboo even in mainstream cognitive science. What it feels like has become one of the most important topics of debate and controversy in the philosophy of mind, where arguments about the 'explanatory gap' (Levine, 1983) and the qualia debate manifest as the cognitivist variant of the mind-body-problem. It is important to realise that the way this debate is led from within the computationalist paradigm is Cartesian (or closet Cartesian), in that the mental is considered a different kind of thing from anything else (objects, the world, meaning, the brain, representation, symbol manipulation; anything

'real' and physically explainable) and we are therefore left with the impossible and artificial task to re-unite these two things that we tore apart.

The enactive approach does not deny that experience does not manifest itself as physical objects. But in not being matter, experience is in good company, with other non-material, non-reducible, but nevertheless real phenomena such as life, meaning, emotions or intentionality. "[E]xperience in the enactive approach is intertwined with being alive and immersed in a world of significance" (Di Paolo *et al.*, forthcoming), not just as data to be explained, but as a guiding force in research methodology. This is not to say that the study of experience (through scientific or non-scientific means) is not methodologically problematic. Experience is the most difficult factor to incorporate into a paradigm for the cognitive sciences. But it surely does not help to pretend experience does not exist.[5] Section 3.5 discusses in more detail how experience can be methodologically incorporated in cognitive science, discussing the distinction between first, second and third person approaches.

A last issue to be clarified is the apparent contradiction between the centrality of the concept of experience in the enactive approach, on the one hand, and, on the other hand, its strong interest in non-human life and cognition and the phylogeny of cognition. Through experience, we know what things mean to us, to our socio-linguistic selves. How can we say anything meaningful about the meaning space of a different species, with a different or more primitive organisation, who cannot even linguistically express themselves? We can find the answer in (Jonas, 1966)'s work and (Weber, 2003)'s extension of it: the 'ecstatic' character of the living allows us to understand, from organism to organism, what something means to another subject, not as 'what it feels like', from the inside, but as 'what it means', reading the signs.

> "[. . .] the patient who is not anymore able to articulate himself, animals, even a paramecium that cramps before it is killed by the picric acid dribbled under the cover slip, the saddening look of a limb plant, the foetus that defends itself with hands and feet against the doctor's instruments – they all *present* the meaning of what is happening to them" (Weber, 2003, p. 118).[6]

[5] To some people, this is not as bizarre a suggestion as it seems. When stating that what I research is the mind, I encountered several fellow researchers with a strong ideological scientism who have, in response, claimed that the mind does not exist.

[6] My translation: "[. . .] der nicht mehr artikulationsfähige Kranke, Tiere, ja sogar das Pantoffeltierchen, das sich zusammenkrampft, bevor es von der unter das Deckglas geträufelten Pikrinsäure getötet wird, der trauig stimmende Anblick einer welken Pflanze, der Fötus, der sich gegen die Instrumente des Arztes mit Händen und Füßen wehrt – alle *zeigen* die Bedeutung dessen, was ihnen widerfährt" (Weber, 2003, p. 118).

2.3.6 *The Roots*

This brief outline of the enactive approach and its central concepts and ideas has made little reference to the numerous predecessors from many scientific disciplines or related contemporary currents of research. It is important to acknowledge these sources of inspiration and explain where the enactive approach comes from.

Maybe the most important predecessor is Maturana and Varela's own theory of autopoiesis (e.g., Maturana and Varela, 1980, 1987). The idea of autopoiesis as the organisation of the living still plays an important role in the enactive approach (previous sections). However, autopoietic theory is more concerned with theoretical and epistemological questions, whereas the enactive approach focuses on scientific practice and explanation. Also, with the idea of 'enaction as embodied action', the enactive approach emphasises the active and engaging side of knowledge construction, whereas the original formulation of autopoietic theory has sometimes (unjustly) been criticised to endorse solipsism or non-naturalism.

There are, of course, also numerous predecessors and contemporary researchers with large ideological and methodological overlap among the countless participants in the universal and millennia old pursuit to explain mind. In section 2 of (Di Paolo *et al.*, forthcoming), we provide a non-exhaustive listing of scientific currents that relate to the enactive approach, featuring, e.g., Piaget's theory of cognitive development through sensorimotor equilibration (e.g., Piaget, 1936), the philosophical strands of existential phenomenology, continental biophilosophy and American pragmatism, holistic dynamical systems approaches in neuroscience, cybernetics, ALife researchers in AI and Robotics, *etc.* It is important to realise the cognation between these predecessors and related approaches and the enactive approach, not just to get a better impression of what enaction is all about, but also because the insights and findings resulting from such approaches can be used to enrich and advance an enactive understanding of the mind. Throughout this book, such related work is referred to as a complement or source of inspiration.

2.4 Challenges, Criticisms and Simulation Models

As already pointed out in Sect. 2.2.4, there are parts of enactionism that are still underdeveloped, areas in which the enactive approach does not have a lot of contributions yet. In particular, those are the GOFAI strongholds in which proponents of minimal representationalist views postulate "representation hungry" (Clark, 1997) problems that require explicit symbol manipulation processes for their explanation. Most of these involve higher

levels of cognitive performance: thinking, imagining, engaging in complex interactions with others, and so on. As already stated in chapter 1, the research described in this book results from frustration with the apparent incapacity of ALife methods to address questions of human level cognition, frustration with the existence of underdeveloped areas and computationalist strongholds. There is no reason to believe that the enactive approach is not able to explain these kinds of phenomena, but as long as it fails to do so, sceptics cannot be hushed. This section outlines how the enactive approach has to grow in order to invade such underdeveloped areas.

In (Di Paolo *et al.*, forthcoming), we argue that "[w]e must not underestimate the value of a new framework in allowing us to *formulate the questions in a different vocabulary*, even if satisfactory answers are not yet forthcoming" (Di Paolo *et al.*, forthcoming). To illustrate this point, we give examples from different areas and from our own modelling work, including the models here presented in chapters 5 and 6. The importance of a shift in perspective and how simulation models can be a technical aid in reformulating old questions is a central issue in this book.

A convenient property of the computationalist paradigm is that, as a consequence of the presumed localisation of function, increasingly sophisticated cognitive processes can be modelled by linearly adding more functional modules and computational complexity to an ever growing AI model of cognition. This is not the same in enactive cognitive science. Global complexity of embodied behaviour sometimes leads to unexpected effects of local changes, which sometimes seem impossible to understand or capture. Simple simulation models can help to make nonlinear interactions intelligible. The models presented in this book address five problems in different disciplines and with different levels of sophistication. Yet all of the models strive for minimalism and at capturing the essence of the behavioural dynamics. This work shows that a complex *explanandum* does not require a complex model to form part of the *explanans*.

The choice of problems addressed with the different simulation experiments reflects a personal journey towards identifying the kinds of questions of human level cognition that the enactive approach is likely to be able to address next. This journey produced the combination of methods proposed for the study of perceptual behaviour and experience that this book proposes. A key component in this set of methods is the kind of experimental work in Sensory Substitution/Perceptual Supplementation that the CRED in Compiègne use, for instance, to explain the sensorimotor basis of space cognition (Lenay, 2003). Space cognition and perceptual experience of space are rather abstract cognitive capacities, unlike the

kind of low-level processes that sceptics see the enactive approach confined to. The group explains plausibly the origin of certain spatial concepts and percepts through their minimal experimental and phenomenological approach. The later ER simulation models in this book model this kind of experimental work (agency detection in chapters 6 and 7 and adaptation to sensory delays in chapter 10). For the study of the sensorimotor basis of simultaneity detection and adaptation to sensory delays, the simulation was implemented alongside the experimental work during a placement in the group (chapter 9), so the modelling could guide the experimental design and data analysis.

This way of pursuing enactive cognitive science is just one in an infinite space of future possibilities. "A proper extension to the enactive approach into a solid and mainstream framework for understanding cognition in all its manifestations will be a job of many and lasting for many years. [...] The strength of any scientific proposal will eventually be in how it advances our understanding, be that in the form of predictability and control, or in the form of synthetic constructions, models, and technologies for coping and interacting with complex systems, such as education policies, methods for diagnosis, novel therapies, *etc.*" (Di Paolo *et al.*, forthcoming). There are other challenges for the enactive approach that will require different methods. This book explores and evaluates the usefulness and scope of applicability of ER simulation modelling to different kinds of such challenges. It concludes with an interdisciplinary research framework for studying the sensorimotor basis of human perception as a promising route to tackle problems of human level cognition, which is also a step towards invading computationalist strongholds.

Chapter 3

Methods and Methodology

As the topic of this book is methodological, this chapter is its core piece and contains a lot of novel material. It presents a methodological framework for the enactive study of human perception. Rather than to just iterate proven methods, large parts of this chapter are dedicated to science-theoretic and methodological argument, to explain and justify the methods proposed and to identify their scopes and limits (hence the title: 'Methods and Methodology', rather than just 'Methods').

The first Sect. 3.1 ties in with issues already raised in chapter 2, about the implications of a constructivist-enactivist world view that denies the existence of an observer-independent reality for scientific explanation. In a similarly general style, Sect. 3.2 assesses the importance and position of the mathematical language of dynamical systems theory for the enactive approach. Section 3.3 introduces Evolutionary Robotics (ER) simulation models. It presents technical details of the ER models used for the modelling parts of this book (chapters 4-7 and 10) and discusses their role in scientific explanation in general. Section 3.4 introduces minimalist experimental approaches to human perception, sensorimotor integration and sensorimotor adaptation, which is in large parts based on ideas developed by the CRED group in Compiègne. The experimental parts of the study of perceived simultaneity (chapter 9) were realised in collaboration with the CRED group. Also, three of the simulation models presented (chapters 6, 7 and 10) are applied to work conducted in their laboratory. Subjective experience is an absolutely essential but methodologically very difficult factor in the study of human cognition and perception. In Sect. 3.5, first, second and third person approaches to the study of experience are discussed. Finally, Sect. 3.6 brings together the methods presented and outlines how they can be applied in mutual benefit, in particular ER simulations, behavioural experiments with humans and perceptual judgements as crude indicators of experience.

3.1 The Scientist as Observing Subject

In the enactive view, knowledge is not represented, knowledge is constructed: it is constructed by an agent through its sensorimotor interactions with its environment, co-constructed between and within living species through their meaningful interaction with each other. In its most abstract and symbolic form, knowledge is co-constructed between human individuals in socio-linguistic interactions.

Science is a particular form of social knowledge construction, characterised by certain rules, dogmas, procedures, objectives and *the use of formal languages and techniques of measurement*, which, if applied correctly, give scientific knowledge properties that make it somewhat special. Most important for modern human society, scientific knowledge can be taken beyond our imagination, following the rules of logic and mathematical deduction and, thereby, allows us to build powerful tools, machines and medicines, to perceive and predict events beyond our immediate cognitive grasp of regularities in the environment, and also to construct further, even more powerful scientific knowledge. This practical power of (some) scientific knowledge, should, however, not seduce us to subscribe to some form of scientism, assigning scientific knowledge *ontological* privileges and a universality which it does not deserve. The significance of scientific knowledge always derives from the context of its generation and from what it means for an individual or a group of individuals (e.g., a society), just like any other form of knowledge, and the methods of science are not applicable to just any problem or phenomenon in the world (a science of love, for instance, will always miss something out, something which music, literature or folk psychology may be better able to capture).

In their early work on autopoiesis, cognition and the principles of life, Maturana and Varela (Maturana and Varela, 1987, 1980) have crucially identified and discussed this status of the scientist as observer and what it implies for scientific practice in biology and cognitive science. Maturana's statement that "everything said is said by an observer" (Maturana, 1978) has become programmatic for the epistemological strand of radical constructivism in the 80s and 90s, and their writings have crucially influenced many pioneers of enactive and proto-enactive approaches in the cognitive science and biology (e.g., contributors to Varela *et al.*, 1991). The importance of the scientist as subject and observer has been recognised by many other thinkers inside and outside the enactive community (e.g., Bitbol, 2001; Kurthen, 1994).

For cognitive science, the inclusion of the scientist as an observing subject leads to a situation where the snake bites its own tail: it applies the rules and methods of science to

explain processes of meaning construction, an *explanandum* that subsumes the application of the rules and methods of science itself. The observing subject in a scientific story is part of both *explanans* and *explanandum* at the same time. Therefore, a constructivist and non-objectivist science makes references to the specific processes of scientific knowledge construction where necessary, which, for the research presented in this book, becomes particularly relevant in the study of perceived simultaneity (chapters 9-11).

An additional problem in cognitive science is that mind and cognition are neither directly observable nor measurable nor quantifiable, where science is an activity that is largely about measurements, observations and quantification. Whilst our scientific measurement of external objects and events is mediated through technology and our sensorimotor interaction with the environment, our knowledge of mental phenomena is direct and subjective, cognition manifests as *experience*, a category not usually considered part of the scientific program. This is what led Descartes to his dualistic world view, distinguishing mind, the *res cogitans*, from basically anything else in the world which can directly take measurable causal effects in the environment and thus manifest in space, i.e., the *res extensa*. Cognitive science thus has the thankless task to explain (amongst other things) the *qualitative* dimension of cognition, including the experience of emotions, intentions, colours, numbers, memories, insights, competencies, communication, *etc*., without actually having the scientific words to express the *explanandum* in the first place.

The way traditional cognitive science deals with this problem is, typically, to *define* unmeasurable mental phenomena in terms of physically measurable variables and to *reduce* them to physical and quantifiable processes. Prominent examples of this practice include:

- The reduction of mind states to physical brain states on the basis of correlated occurrence, a practice that is popular with some philosophers of mind working in the qualia debate and on the neural bases of consciousness. In its most consequent and extreme form, this reductionism results in eliminativism (e.g., Churchland and Churchland, 1998).
- The functional reduction of cognitive phenomena to physically measurable processes that convincingly appear to bring about that cognitive phenomenon in an entity that is not oneself (Turing-test approaches, after Turing, 1950), a technique that is more commonly adopted in the areas of artificial intelligence and cognitive modelling and underlies (Dennett, 1989)'s ideas on the 'intentional stance'.

The problem with these reductionist approaches is, in a nutshell, that by picking an isolated physical phenomenon and explaining it, you explain the isolated physical phenomena you pick, but not cognition, mind or subjective experience.

Instead of indulging in ideological quarrels, in the remainder of this section, it shall be argued how cognition can be studied scientifically in a *bona fide* enactive way, avoiding the reductionist practices just mentioned. Firstly, it is necessary to establish as part of the scientific explanation how measured empirical findings relate to experiential phenomena (Sect. 3.5 discusses the methodological difficulties associated with the study of experience in detail). Instead, reductionists of the localist type identify a local correlation and presume that it 'does' the mental capacity we are after, taking it out of its physiological, physical and semantic context (e.g., reducing mental states to brain states). Secondly, in order to be able to say something meaningful about functional aspects of cognitive faculties, it is important to explain the mechanisms that generate it, rather than just to explain some mechanism that successfully imitates particular aspects of the cognitive faculty under investigation (reductionism of the functionalist style, Turing-test approaches). We have developed and discussed this point, focusing on the example of the scientific study of autonomy in (Rohde and Stewart, 2008) and the remainder of this section reproduces our argument.

The scenario developed by Alan Turing in his 1950 classic paper 'Computing machinery and intelligence' (Turing, 1950), which he called the 'imitation game' expresses a deep pessimism towards the possibility to properly scientifically account for intelligence or cognition. Via a language interface, what is tested is the capacity to trick a human being into thinking that it was interacting with another person, assuming that this capacity would presuppose some form of thinking in the machine. Turing's original formulation of the test was rather tame, i.e., that towards the end of the 20th century "an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning [a computer]" (Turing, 1950) and may even have approximately true: there are programs that use simple techniques (e.g., grammatical pattern matching, rules to generate standardised answers to the most commonly asked questions, ...) that are quite good at tricking humans into the belief that they are actually communicating with a cognitive system with linguistic capacities, even if only for a short while. The reality of how such systems are programmed and the kind of mistakes they make, however, quickly reveals that these agents do not actually think or have any grasp of the meaning of the symbol strings they produce. The cognitive achievement here is to be attributed to the programmer, not

the programs. This is what (Searle, 1980) illustrates in his famous 'Chinese room' thought experiment.

As we argue in (Rohde and Stewart, 2008), knowledge about the mechanisms that generate a phenomenon has a tendency to produce such reactions of disenchantment, the prime example being to know how a conjuring trick works. This knowledge clearly takes away the excitement about the seeming supernatural powers at work being profane slight of hand or visual illusions. But, the important point to realise is that acquaintance with the underlying mechanism does not necessarily lead to disenchantment. On the contrary, sometimes, knowing how something works can produce the opposite effect: for example, a glider in the game of life does not look any different from a first-generation computer game sprite if you just look at it moving around on a two-dimensional grid. Only if you learn about the local cellular automata rules that underlie the emergence of a glider, their simplicity and the fact that they do in no way directly specify any of the emergent behaviour and appearance of the glider, it turns into a fascinating phenomenon, and there is no ulterior knowledge to be acquired that could take this fascination away.

Applying these ideas to the study of cognition, our argument is that learning about the simple algorithms and rules of symbol manipulation that bring about seemingly intelligent or linguistic behaviour in GOFAI systems can leave behind a similar taste of charlatanry as the revelation of a conjurer's trick. I have personally experienced this disappointment many times with laymen, who have seen robots do impressive things (such as playing a violin or taking verbal orders and execute them) in a short TV clip, from which naïve spectators conclude that their capacities would generalise to other situations that are equally cognitively complex. When learning about the limitations of these machines, the reaction is typically disenchantment.[1] Figure 3.1 is a toy-illustration of this discrepancy in the case of ascription of autonomy to robots or living organisms: if autonomy (or any other cognitive capacity) is ascribed to a robotic agent using a kind of Turing-test that relies on superficial acquaintance in (A), knowledge about the generative mechanisms can lead to a revision of judgement in (B). In contrast, when studying the autopoietic organisation of a living organism, acquaintance with the mechanism does not usually have this disenchanting effect.

[1] A recent example of such typical disenchanting revelations can be seen in a demonstration of Honda's ASIMO robot in 2006, that has been captured in video and made available in the internet (http://www.youtube.com/watch?v=VTlV0Y5yAww; retrieved 21.06.2009). In the video, the robot falls down the stairs and remains lying on the floor, but keeps talking and moving as if it was still climbing. This clearly reveals that ASIMO does not understand the meaning of the movements or the words it produces.

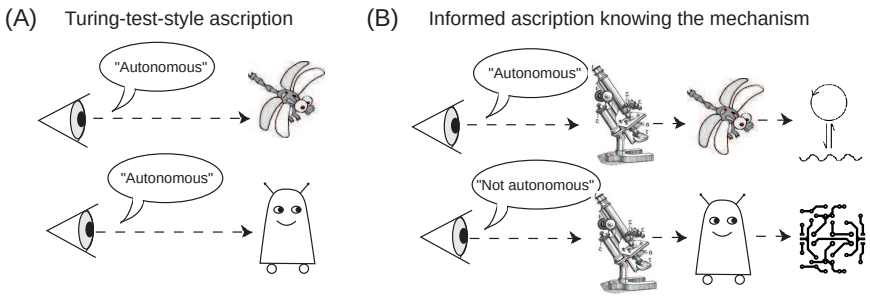

Fig. 3.1 Illustration of ascriptional judgements of autonomy based on naïve observation (A) and scientific study of the generative mechanisms (B).

That a mechanism be or be not convincing is by no means something inherent or restricted to living or ALife-style processes. Just as there are many genuinely fascinating machines (such as cars and computers), people can also get disappointed with processes generated by living organisms. For instances, there is a tendency to be disappointed by stigmergic processes in insects, as the example of the digger wasp (discussed, e.g., in Dennett, 1985) shows: the wasp appears to have an elaborate plan of clearing a tunnel it dug before putting a larvae in it. However, by dislocating its larvae while the wasp is inside the tunnel, the wasp can be trapped in an 'infinity loop' of repeatedly checking whether the tunnel is blocked. This reveals that it does not actually *know that* it is clearing the tunnel, in the sense of understanding the concept of tunnel clearing, but much rather *knows how* to clean the tunnel, following a sequence of behaviours that are triggered by changes in the environment. This behaviour is similar to a computer executing an algorithm and can lead to disenchantment in the same way – when realising that the apparently intelligent behaviour can be brought to break down so easily.

We therefore propose in (Rohde and Stewart, 2008) to substitute a Turing-test style statistical measure of intuitive ascriptional reaction with informed ascription based on the scientific knowledge about generative mechanisms. This is not to propose a project of defining cognitive or mental faculties in terms of the physical properties of the processes that generate it or to engage in any other form of reductionist activity. It is proposing to make use of the powerful characteristics that scientific knowledge has (as outlined above) in the larger endeavour to understand and explain mind and cognition, which is, in the end, what cognitive science is all about. Apart from being more robust and reliable than many other forms of knowledge, scientific knowledge has the advantage that it is subject to inter-subjective debate and agreement, which can resolve controversies about whether or not a mechanism

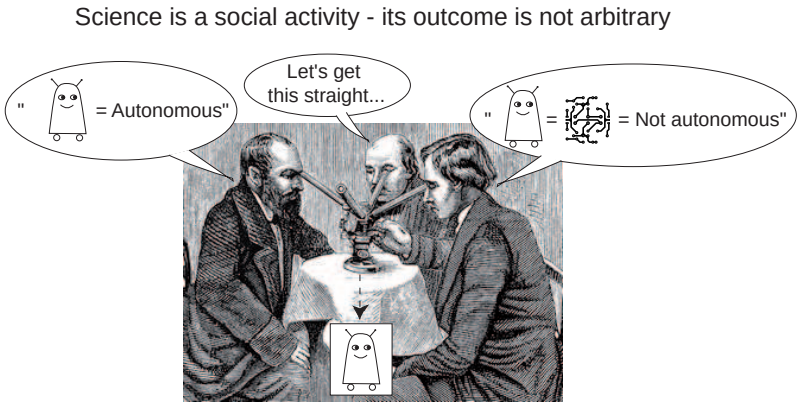

Fig. 3.2 Illustration of the social dimension of scientific knowledge construction.

'counts': "[if] the disagreement remains within the scope of a single paradigm, the normal process of Popperian refutation (or not) will lead to progress. If the disagreement occurs between incommensurable Kuhnian paradigms, then an element of subjective choice may remain" (Rohde and Stewart, 2008, see Fig. 3.2).

A criticism that this argument has stipulated repeatedly (in personal communication) is that the knowledge about generative mechanisms could equally well be substituted for by a perfect and complete description of the surface behaviour (i.e., how inputs and outputs relate over time), without any direct knowledge of the generative mechanisms. Supposedly, this 'LaPlacian Demon' type knowledge would be as powerful a basis for identifying autonomy as the scientific study of the generative mechanisms. Without even entering into a metaphysical quarrel whether or not this is strictly true in a principled way, this argument can be easily put to rest with epistemic arguments. Apart from the fact that for most real-life complex entities (and in particular living organisms), humans would be incapable of grasping the entirety of their sensorimotor couplings at once and confidently judge about their properties as a whole, the question to ask is one of parsimony: why bother with such an extensive project, if we can as well study the generative mechanisms?

3.2 Dynamical Systems Theory

3.2.1 *Definition*

In this section, some of the key terms and definitions in Dynamical Systems Theory (DST) are introduced that are referred to repeatedly throughout this book. Readers without train-

ing in formal languages who may find this section (or other formal/technical parts of this book) difficult to understand are encouraged to skim this section. From the natural language parts of this section, the core ideas and concepts of DST should become sufficiently clear to follow the main points made in this book. The definitions here used stem from the following sources: (Strogatz, 1994; Rohde, 2003; Ross, 1984).

A state x of a dynamical system is a set of system quantities that allows the complete description of the system's development across time. Formally, a state is a variable assignment to a set of variables (state variables) of a dynamical system. In a dynamical system that models a real world system, the state variables correspond to measurable quantities. Apart from state variables, a system can have control parameters, which can change on a slower time-scale than the state variables. Their change is not accounted for in the description of the system: control parameters define a parameterised set of different dynamical systems.

Dynamical systems can either be given as a set of differential equations (time-continuous) or as a set of difference equations (iterated maps; time-discrete). In the work presented in this book, the dynamical systems investigated are differential equations, even if they are investigated discretised in computer simulation (see below).

Details of different types of differential equations (ordinary, partial, stochastic, ...) and their formal properties are not relevant here (see (Strogatz, 1994) for an accessible introduction). The only important concepts to be briefly discussed are the distinction between *linear and nonlinear* dynamical systems and the notion of an *attractor*.

A linear dynamical system is basically a dynamical system in which the behaviour of the whole system is equal to the sum of the behaviours of its parts. This is in accordance with the general definition of a linear function in mathematics. In order for a differential equation to be linear, the terms that describe the change of the state variables must, therefore, not contain any nonlinear functions of state variables, such as power functions, products, trigonometric functions, *etc.* If they do, the differential equation is nonlinear. In a nonlinear dynamical system, the behaviour of the entire system cannot be understood from looking at the behaviour of its part in isolation because, once plugged together, their behaviour can be entirely different than we would expect it from a linear system. The claim underlying dynamical and situated approaches is that cognition and living organisms rely heavily on nonlinear dynamics (both inside the nervous system, in brain-body-interaction and in closed-loop interaction with the environment). Such nonlinear phenomena are methodologically difficult, as they have to be studied in the holistic context of embodied action.

Open-loop and localist approaches are unable to capture such nonlinear phenomena, as they investigate the behaviour of isolated structures, implicitly presuming that putting the parts together will explain the behaviour of the entire system as if it were linear.

The mathematical tools for the analytical computation and analysis of nonlinear differential equations are not yet very advanced and those that exist require strong formal skills. Therefore, computer simulations are important in the study of dynamical systems – even if we cannot formally solve a system of differential equations, we can investigate how it behaves in different settings by simulating it and looking at its behaviour. In order to simulate time-continuous dynamical systems in digital computer simulation, the differential equations have to be discretised using numerical methods. The only numerical method used for the work presented in this book is the forward Euler method which approximates the change in state of a differential equation $\dot{x}(t) = f(t, x(t))$ after a time step of length h as

$$x(t+h) = x(t) + hf(t, x(t)) \tag{3.1}$$

Among the interesting properties of dynamical systems are what is called *attractors*. According to Strogatz, "there is still disagreement about what the exact definition [of an attractor] should be" (Strogatz, 1994, p. 324). He defines an attractor as a closed set of states A that is *invariant, attracts an open set of neighbouring initial conditions* and is *minimal*.

'Invariant' means here that any trajectory that starts in A ends in A. Invariant sets can be fixed points ($A = \{x^*\}$ with $f(x^*) = 0$), limit cycles (circular orbits in A), quasi-periodic (non-circular orbits on the surface A of a torus) or strange (chaotic, fractal) sets. The latter "exhibit sensitive dependence on initial conditions" (Strogatz, 1994, p. 235). This means that trajectories within a chaotic attractor A, even if they start at states that are very close, will describe very different orbits within A. Whilst fixed points can also exist in linear dynamical systems, limit cycles, quasi-periodic and strange (chaotic, fractal) attractors exclusively occur in nonlinear dynamical systems.

The set of initial states attracted to A is called the *basin of attraction* B of an attractor, where B contains A. The basin of attraction is characterised by the fact that the distance from $x(t)$ to A tends to 0 as $t \to \infty$. An invariant set without a neighbouring basin of attraction is not an attractor. Such invariant sets are *unstable* or – in rare cases – *semi-stable*.

Minimalism means here simply that there is not a subset of A for which the same properties (invariance, asymptotic stability) hold.

An orbit within the basin of attraction of an attractor that converges towards the invariant set is called a *transient*. A system is globally stable if all system states converge to a single

attractor, it is multi-stable if it has more than one attractor. A convergent (dynamically trivial) dynamical system is one that has only fixed point attractors.

A dynamical system is called an open system if it interacts with the environment; otherwise, it is called a closed system. Any particular dynamical system is characterised by a fixed *attractor landscape*. However, parameter changes can change attractor landscapes both quantitatively (i.e., location of attractor and basin in state space) and qualitatively (i.e., topology of attractor landscape). *Bifurcation theory* is the branch of mathematics that describes how attractor landscapes in dynamical systems change with gradual changes in control parameters. Such reshaping of attractor topology can be complex and nonlinear in itself.

3.2.2 *The Explanatory Role of DST*

Being based on the 'Mind as Machine' metaphor, traditional cognitive science centres around a mathematical formalism, i.e., the Turing machine/automata theory/formal logic as the fundament on which to build a unified interdisciplinary science of mind. Some approaches that are critical of classical computationalism and question the central role of this metaphor have tried to put other formal languages in its place, such as Connectionism proposing ANNs and Dynamicism proposing DST (cf. previous chapter, Sect. 2.2). (van Gelder, 1998)'s proposal of the 'dynamical hypothesis in cognitive science' distinguishes the *nature hypothesis* and the *knowledge hypothesis* (van Gelder, 1998) as two sides of the same coin. The nature hypothesis is the hypothesis that what is cognitive about a cognitive systems is fully captured by an abstract formal description of its behavioural and brain dynamics, i.e., it *is* this dynamical system, which can, in principle, be variably instantiated in material terms. The knowledge hypothesis is that a cognitive system is best studied with DST as formal tool.

The dynamical turn in cognitive science has gained in impact over the last years (e.g., Beer, 2000; Port and van Gelder, 1995; Thelen and Smith, 1994). Researchers identifying with Dynamicism work in areas as different as linguistics, physiology, cognitive psychology, developmental psychology, cognitive neuroscience, *etc.* Broadly speaking, the enactive approach can be seen as forming part of this dynamical turn, even though its core assumptions are not identical (cf. chapter 2). This difference does not entail a reservation: nearly all the work done under this label is thrilling, even from an enactive point of view. However, in contrast to van Gelder's dynamical hypothesis, for enactivism, DST is not seen as a

privileged formalism, but just a very suitable language for formalising the material aspects involved in cognition.

The reason why DST is so important for enactive cognitive science is the same reasons that assigns DST an important role in all natural sciences, and in particular in physics. As developed in chapter 2, the enactive approach investigates the mutual links between the material mechanistic level and the behavioural, cognitive and relational level. Enactivism is interested in the origins, adaptive changes and the maintenance of invariant emergent structures. Such self-organisation is an inherently dynamical phenomenon. DST, as the language of physics, serves to describe the evolution of a whole situation over time, including an agent, its body, its environment and its brain. In order to describe and model embodied and embedded agents in a way that minimises prior assumptions about how structure relates to function, DST as a descriptive formalism has a clear competitive edge because of its capacity to describe physical processes in general. For the description and study of the mechanistic or physical level without building in prejudices about functionality of structure, DST suggests itself. From this, it does not follow that other formalisms (such as automata theory, information theory, game theory, ...) cannot be equally useful for any particular research question.

3.3 Simulation Models, Evolutionary Robotics and CTRNN Controllers

3.3.1 *Evolutionary Robotics Simulations*

Evolutionary Robotics (ER) is a "technique for the automatic creation of autonomous robots [...] inspired by the [D]arwinian principle of selective reproduction of the fittest" (Nolfi and Floreano, 2000, preface). In this approach, some aspects of the robot's or simulated agent's architecture are specified, but others are under-specified. These are left to be determined in an automated way by an evolutionary search algorithm, according to the optimisation of an abstract performance measure called the 'fitness function' (see Fig. 3.3 for an illustration of the process).

There are studies in ER that test fitness in real-time real-world robot experiments. The ER models presented in this book, by contrast, have been evolved in simulation, which is the more common approach. The parameters evolved are the parameters of the neural network controller, but, in principle, many parameters, including morphology, sensory equipment or initial conditions can be evolved. This section describes the algorithm and techniques that are common to the different models presented in this book (control network, parameter

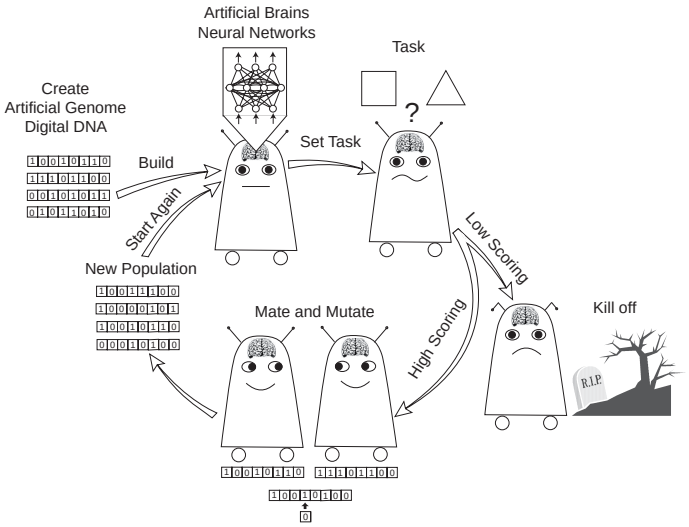

Fig. 3.3 Illustration of the evolutionary cycle in ER.

ranges, genetic algorithm, *etc.*). This section, again, is rather technical and may contain details that are not strictly relevant to a reader who is unconcerned with formal models or unfamiliar with technical jargon. Such readers are invited to move on to the next section, even though, in order to understand the research presented in this book, it is essential to get at least a rudimentary idea of the technique of ER (i.e., to understand Fig. 3.3).

In each of the modelling chapters 4, 5, 6, 7 and 10, more technical details are provided that are specific to the model. In some of the models, there are deviations from the general principles described here. These deviations are pointed out within the modelling section of the respective chapter.

3.3.1.1 *Continuous-Time Recurrent Neural Networks (CTRNNs)*

A method used and promoted by Beer is the use of a particular network type for ER neural control, i.e., Continuous Time Recurrent Neural Networks (CTRNNs, e.g., Beer, 1995). Even though the dynamical properties of CTRNNs can be seen as idealisations of real neural dynamics, CTRNNs are not used in direct analogies for the brain or brain areas here. Beer advocates this type of controller because "(1) they are arguably the simplest nonlinear, continuous dynamical neural network model; (2) despite their simplicity, they are universal dynamics approximators in the sense that, for any finite interval of time, CTRNNs can approximate the trajectories of any smooth dynamical system on a compact subset of $\mathbb{R}^n$

arbitrarily well" (Beer, 1995, p. 2f). Furthermore, they are very suitable for evolutionary approaches because of their interesting convergence properties – even very small networks can exhibit multi-stable, oscillatory or chaotic behaviour (Beer, 1995, 2006).

The network structure employed in most models in this book is a partially layered control network in which a layer of input neurons projects onto a layer of fully connected interneurons which, again, projects onto a layer of output neurons. However, in individual models this structure is modified, as indicated locally.

The dynamics of neurons in a CTRNN is governed by

$$\tau_i \frac{da_i(t)}{dt} = -a_i(t) + \sum_{j=1}^{N} c_{ij} w_{ij} \sigma(a_j(t) + \theta_j) + I_i(t) \tag{3.2}$$

where $\sigma(x)$ is the standard sigmoidal function:

$$\sigma(x) = \frac{1}{(1+e^{-x})} \tag{3.3}$$

Other variables are: $a_i(t)$ is the activation of unit i at time t, θ_i is a bias term, τ_i is the activity decay constant and w_{ij} is the strength of a connection from unit j to unit i. The $n \times n$ connectivity matrix C with $c_{ij} \in \{0,1\}$ specifies the existence of synaptic connections between neurons. In some simulations, the network structure is evolved, including the connectivity matrix C (see network structure specification in local method sections). In most models, however, a partial layering of the control CTRNN is implemented, where input neurons do not have incoming connections from within the network, input neurons cannot project directly to output neurons and output neurons do not have outgoing connections back into the network.

The biological analogy of CTRNNs frequently adopted is that a_i represents the membrane potential, τ the membrane time constant, θ the resting potential, $\sigma(x)$ the firing rate, w_{ij} the strength of synaptic connections between neurons and I_i network-external inputs impacting on membrane potential. As stated above, this biological interpretation of CTRNN dynamics is not relevant to the modelling approach taken here. In order to cache in on the biological plausibility, real neural structures and connectivity patterns would have to be modelled. Instead, the 'robot brains' modelled here have less than ten neurons as a whole. The CTRNN controllers represent neural dynamics in a more abstract sense: they link sensation and motion quickly and can transform patterns of stimulation nonlinearly in very diverse ways over time, which can lead to the emergence of interesting dynamical structures. Effectively, many of the evolved controllers discussed in this book rely on circuits that could even have implemented in linear systems, because the interesting dynamical phenomena emerge from the closed-loop interaction, not directly from complex

neural dynamics. The benefit of using CTRNNs is that, if necessary, it still is possible for more complex dynamical structures to evolve (such as the neural oscillator for arm control discussed in chapter 7). Therefore, the implementation is less biased with respect to the question whether a control system should be linear or nonlinear: both can evolve.

CTRNNs are actually continuous dynamical systems, but, as stated before, they are simulated using the Euler method (Eq. (3.1)). Applying the Euler method to the above Eq. (3.2), the following approximation yields:

$$a_i(t+h) = a_i(t) + \frac{h}{\tau_i}(-a_i(t) + \sum_{j=1}^{N} w_{ij}\sigma(a_j(t) + \theta_j) + I_i(t)) \tag{3.4}$$

In order for this equation to approximate CTRNN dynamics sufficiently closely, the τ_i have to be sufficiently large compared to the time-step h (in most models, $h = 1$). In the models here presented, the minimal ration $\frac{h}{\tau}$ set as parameter boundary is 10 but in most models, it is larger than that. In several models (chapters 6, 7 and 10), sensory delays d have been used, i.e., sensory inputs were held for d time units before they were fed into the network.

3.3.1.2 *Simulation*

CTRNNs are used to model the internal dynamics of the evolved agent controllers. The emphasis of ER is, however, on the *closed loop* modelling, i.e., a whole situation is modelled, not just input-output mappings or decoupled neural dynamics. In a diagram that Beer frequently employs to illustrate this idea (Fig. 3.4), the CTRNN dynamics can be seen as the dynamics in the innermost box (NS). In order to implement the external closure of the sensorimotor loop, i.e., how an agent's actions in the world impact dynamically on its sensations, the body (middle box) and the environment (outermost box) have to be modelled as well.

In the ER models presented in this book, agent bodies manifest simply as functions transforming particular environmental variables into neural inputs and neural outputs into velocity or force vectors (e.g., wheel velocity, angular joint velocity, directional velocity, ...). These functions usually involve sensory gains S_G and a motor gains M_G to scale inputs and outputs appropriately as they are fed in or read out of the network. These gains are the only bodily parameters that are evolved rather than fixed.

The outermost box in Beer's diagram (Fig. 3.4) is simulated as a virtual space of some kind in which the state and location of agents and possible external objects are stored and updated, interpreting force and velocity vectors resulting from previous world states and CTRNN outputs. The same time scale is used for both neural and environmental dynamics, which are updated at the same frequency.

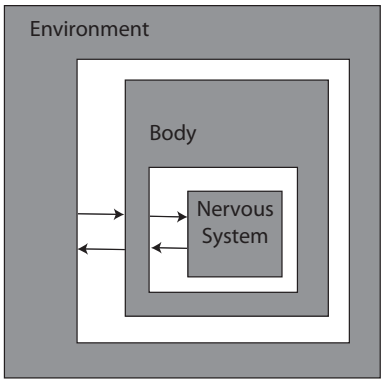

Fig. 3.4 Illustration of brain-body-environment interaction, inspired by Beer (e.g., Beer, 2003).

3.3.1.3 *Genetic Algorithm*

A genetic algorithm (GA, Holland, 1975) is an optimisation search algorithm for a parameter configuration that performs a heuristic search on the parameter space inspired by the Darwininian principles of heredity, mutation and natural selection that is similar to hill climbing search (but more random).

The search algorithm used in this book is a simple generational GA. This means that for a fixed number of generations (typically one or several thousands), a set p of individuals ($|p| = 30$ in this book) is used to generate a new generation of equal size and is then fully replaced. For each individual $i \in p$, a parent is selected with uniform probability from the $1/3$ best individuals from the previous generation according to the fitness measure F_i (i.e., truncation selection). Non-sexual reproduction was implemented, i.e., an individual's genotype is a mutated clone of the single parent's genotype. Genes are real-valued $\in [0,1]$ and vector mutation (e.g., Beer, 1996) is used as mutational operator. This means that the genotype is mutated by adding a random vector of magnitude r (magnitude Poisson distributed) in the n-dimensional genotype space to the genome. If mutation of a gene exceeds the gene boundary, it is *reflected*, i.e., the amount by which the gene boundary is exceeded is subtracted from the gene boundary to yield the new gene value.

Genes are interpreted as network parameters τ_i, θ_i and w_{ij} and as S_G and M_G. The parameter ranges vary between simulations and are specified locally. Typically, $w_{ij} \in [-8,8]$, $\theta_i \in [-3,3]$, and these values are mapped linearly to the specified target range. The minimal value for τ_i is *ca.* $20h$ and the maximum value for τ_i is in the order of magnitude of the duration of a trial or a meaningful action in the task. M_G, S_G and τ_i are mapped expo-

nentially to their target ranges, which means that the inter-individual differences that the GA works on are more fine grained for small values of M_G, S_G and τ_i than for large values of M_G, S_G and τ_i. In some cases, network structure was also modelled, i.e., genes were interpreted using step functions to determine the existence of synaptic connections c_{ij} or, in some cases, for the existence of inter-neurons n_i.

Typically, fitness evaluation is computed from several evaluation runs. In the models here presented, fitness was either averaged from several trials or an exponentially weighted fitness average was used such that for n evaluations

$$F(i) = \sum_{j=1}^{n} \left(F_j(i) \cdot 2^{-(j-1)} \cdot \frac{1}{2^{-(j-1)}} \right) \tag{3.5}$$

where $F_j(i)$ gives the fitness on the j^{th} worst evaluation trial for individual i. This evaluation technique gives more weight to worse evaluations and thereby rewards the generalisation capacity of the evolved agents. This means that it helps to avoid that evolutionary search gets stuck in a locally optimal trivial solution that stably yields a high score for some parameters of the task. At the same time, it rewards the evolution of such locally optimal behaviour as compared to no sensible behaviour at all, by still giving some fitness for solving parts of the problem.

3.3.2 *Simulation Models as Scientific Tools*

After explaining what ER simulations are and specifying the technical details of the ER simulation models presented in this book, it will now be discussed what their contribution to science consists in, preparing for an adequate evaluation of the work with respect to the methodological theme of the book.

Many ALife and ER simulation models are different from the typical formal or simulation models in other scientific disciplines, such as theoretical physics, biology or sociology. The function of scientific models is, typically, to fit and describe an empirically gathered data set, thereby generalising its structural properties and predicting future measurements. ALife modelling is a more *generative* modelling approach. In clarifying this assertion, some of the arguments and positions presented in (Rohde and Stewart, 2008; Di Paolo *et al.*, 2008; Beer, 1996; Di Paolo *et al.*, 2000; Harvey *et al.*, 2005) are reproduced here.

(Di Paolo *et al.*, 2000) argue that ALife simulation models are to be understood as 'opaque thought experiments'

> "[...] it is reasonable to understand the use of computer simulations as a kind of thought experimentation: by using the relationships between patterns in the simulation to explore

> the relationships between the theoretical terms corresponding to analogous natural patterns" (Di Paolo *et al.*, 2000).

Simulation models are guaranteed to only produce phenomena that logically result from the premises built into the model as there are no possibly interfering external variables as in complex real-world science. Thereby, they can generate proofs of concept of the kind of processes that can produce a certain kind of phenomenon under certain circumstances – or not. (Braitenberg, 1984)'s work on fictional *Vehicles* can be seen as a paradigmatic example of this kind of generative modelling approach and a predecessor of and inspiration for ALife simulation modelling.

However, an important novelty is that through the use of digital computer technology, simulation models can go beyond human cognitive limits or prejudices. How dynamical systems, in particular nonlinear dynamical systems evolve in time is extremely difficult to grasp and intuit without the help of computer simulations. A good example is (Hinton and Nowlan, 1987)'s simulation model of the Baldwin effect in evolutionary biology. Broadly, the Baldwin effect refers to facilitated integration of a biological trait into the genome by ability to learn that trait in previous generations. The mechanism had been proposed but not credited because, at first glance, it appeared to propose Lamarckianism (i.e., direct integration of acquired skills into the genome). Only with the help of a simulation model, it could be established beyond doubt that lifetime adaptation can aid the evolution of biological traits within a Darwinian framework. "A proposed mechanism that had not been perceived as convincing because it was counterintuitive and difficult to understand had been made credible with the help of a computational model" (Rohde and Stewart, 2008). As a result of this conceptual contribution, the Baldwin effect has become a widely acknowledged concept in evolutionary theory.

This power of simulation models to counter our intuitions and go beyond our imagination, at the same time, makes them more difficult to work with than 'armchair' thought experiments. This is where the 'opacity' comes in: "Due to their explanatory opacity, computer simulations must be observed and systematically explored before they are understood" (Di Paolo *et al.*, 2000). After producing a simulation result, a 'pseudo-empirical' investigation of the simulation follows, in order to understand and explain how exactly it works. Different variables are monitored over time and parameters and conditions are modified in order to discover the systematicities governing the simulation. Such exploration is, in a way, similar to hands-on scientific work, but has the benefit that the *explanandum* is fully controllable, simpler, fully accessible and experiments are easily reproducible. Therefore,

it is easier to derive general principles and formal rules governing the simulation dynamics, insights that can then be fed back into the original scientific community to inform theory building.

(Harvey *et al.*, 2005) elaborate on the scientific function of ER simulation models in cognitive science, using examples from ER simulation research on homeostatic adaptation, the origins of learning and sensorimotor development. As important features of ER simulations, they identify the *minimisation of complexity and prior modelling assumptions*. In the light of the frequent criticism of ALife modelling that it is difficult to conceive how it would scale up (e.g., Kirsh, 1991), it may seem surprising that minimalism is perceived as a merit. Many AI modelling approaches aim at approximating human or real brain complexity as closely as possible (e.g., Markram, 2006). The problem with this kind of approach is that quickly the model becomes as opaque as the original phenomenon, whilst not generating useful generalisations or abstractions.

One of the most passionate proponents of a minimal modelling approach is (Beer, 1996). When dealing with complex dynamics, even systems that seem very simple at first glance can generate surprisingly complex behaviour (e.g., Beer, 2003, 1995). Beer argues that, therefore, dynamical principles should first be properly analysed and understood in the most simple and abstracted case, to get intuitions about the kind of dynamical phenomena that exist in sensorimotor interaction, develop tools to study them and then build up complexity gradually. He talks about minimal simulation models as 'frictionless brains' in analogy to Galileo's 'frictionless planes' (Beer, 2003) that allow us to do the mental gymnastics to build intuitions, form concepts and hypotheses in order to ultimately advance with real world scientific work and explanation.

ALife simulation modelling is different from and goes beyond formal description and fitting of an empirically gathered data set because its results are more conceptual and abstract than quantitative predictions and impact on theory building as well as the scientific practices of designing experiments and interpreting data. (Webb, 2009) questions the scientific value of such merely conceptual models. Even though she is right in pointing out that ALife, as a field, is not sufficiently concerned with establishing the links between the models generated and real existing organisms, it is important to see that, for any particular model, a biological grounding of simulation results is not *a priori* necessary in order for the model to be scientifically valuable. For instance, she targets (Beer, 2003)'s model of a simple agent solving a categorical perception task by means of dynamically reshaping the attractor landscape of the agent-environment system through dynamical interaction with the environment. This

beautiful and simple model has provided an important proof of concept of this kind of dynamical process and has, thereby, directly inspired and influenced more applied work (noticeably, the models presented in chapters 6 and 7 of this book; see also the critical replies published alongside the (Webb, 2009) target article, including (Rohde, 2009)). An important point to make is that the generative modelling emphasised in this section does in no way contradict, exclude or oppose the possibility of descriptive data-driven modelling. We identify descriptive and generative modelling in psychology as "two poles [...] [that] define a continuum of dynamical approaches" (Di Paolo *et al.*, 2008).

As concerns the models presented in this book, they can be seen as examples for different roles that ER simulation models can play in scientific activity. The models of synergies (chapter 4) and of value system architectures (chapter 5) are predominantly generative models in the 'opaque thought experiment' sense outlined above. They strongly idealise the original phenomenon observed. The model of synergies (chapter 4) mainly cashes out the capacity of simulation models to exceed our cognitive grasp of nonlinear dynamics, in order to verify theoretical concepts, generate new hypotheses and suggest further experiments to empirical researchers. As such, it serves as a support structure for empirical scientific practice. The model of value system architectures (chapter 5), on the other hand, exploits pre-dominantly the fact that simulation models can take us beyond our intuitions, illustrate inconsistencies in conceptual arguments and point out implicitly held prior assumptions, which is more relevant to philosophical debate and theory building than to hands-on experimental practice. The models of perceptual crossing (chapter 6 and 7) and adaptation to sensory delays (chapter 10) also have descriptive elements. This is possible because the experimental work modelled follows a similar minimalist agenda, which means that the virtual environments in which humans are tested are the same or equivalent to those in which agents are evolved. This allows stronger analogies (see Sect. 3.4 below). Even though they also generate proofs of concept and counterintuitive insights, some direct and quantifiable predictions or measures for gathered data and future experiments result from these models. This use of ER simulation models tries to get the best of both worlds by generating concrete predictions like 'ordinary' models, as well as to contribute to the philosophical debate which surrounds the perception research modelled (see also Sect. 3.6).

3.4 Sensory Substitution and Sensorimotor Recalibration

This section introduces a line of research called 'sensory substitution' (Bach-y Rita *et al.*, 2003) and addresses how it relates to more general research in perceptual learning or sen-

sorimotor recalibration. The approach has been termed 'perceptual supplementation' (PS) by the CRED group at the Technological University of Compiègne, who have generated useful conceptual contributions identifying its potentials, but also the limitations of this kind of approach (Lenay *et al.*, 2003). The simulations presented in chapters 6, 7 and 10 model results from this strand of experimental research and chapter 9 presents empirical results using the kind of technique described.

In 1963, Bach-y-Rita *et al.* have started a research program of building prosthetic devices for blind people that allow for substitution of aspects of their visual sense, with tactile signals representing visual information (Tactile Visual Sensory Substitution, TVSS; e.g., Bach-y Rita *et al.*, 1969, 2003). Equipped with a head-mounted camera that relays pixeled images to arrays of tactile stimulators (on the belly, the fingertip, the back, the tongue,...), congenitally blind people can be trained to perform tasks that are normally considered visual tasks, such as face recognition, catching a ball (which requires 'hand-eye-coordination'), or recognising shapes. Bach-y-Rita sees this technology as a direct extension of the principle of a blind person's cane: even though the cane produces tactile stimulation of the palm of the hand, blind people use it to perceive objects at a distance. As they get used to navigating with a cane, the automated swaying movements and the vibrations in the palm of the hand that holds the cane disappear from their conscious experience and, instead, blind people perceive external objects, such as steps, doors, puddles, *etc.* In a similar way, when trained with the TVSS, subjects employ visual language to express their experiences, and optical illusions have been reproduced in subjects trained with the TVSS (Bach-y Rita *et al.*, 2003). This fascinating research program, which over the years has been applied also to other sensory disabilities (most noticeably, equilibrial disabilities) continues vividly despite Prof. Bach-y-Rita's recent lamentable death, in his own department and in other groups, who have taken up the idea and built similar devices. Different teams also explore other sensory channels, such as the auditory to visual sensory substitution in the vOICe system (Amedi *et al.*, 2007), showing that the principles of this kind of sensorimotor adaptation hold more generally. The term 'sensory substitution' has become the label for technology that records signals associated with one sensory modality and, through the use of technology, transforms it to stimulate, non-invasively, sensors of another sensory modality (Lenay *et al.*, 2003).

Apart from its practical prosthetic use to improve the lives of people with sensory disabilities, the fact that this technology works the way it works makes it a rich tool for the study of the nature and sensorimotor origins of human perceptual experience. As Hurley and

Noë remark, in TVSS "the qualitative expression of somatosensory cortex after adaptation appears to change intermodally, to take on aspects of the visual character of normal qualitative expressions of visual cortex" (Hurley and Noë, 2003). This fact seems difficult to reconcile with the reductionist ideas of functionally dedicated brain areas whose activation is the physical correlate of experiences of a certain modal quality. It thus gives evidence for their "dynamical sensorimotor hypothesis" according to which "changes in qualitative expression are to be explained not just in terms of the properties of sensory inputs and of the brain region that receives them, but in terms of dynamic patterns of interdependence between sensory stimulation and embodied activity" (Hurley and Noë, 2003).

While the second part of their argument (i.e., that changes in qualitative experience are to be explained as well in terms of dynamical patterns of sensorimotor interdependence) is in agreement with the enactive approach as it is proposed in this book, the first part of their argument (i.e., that there is an intermodal transfer of experience and that information received by tactile sensors has visual qualities) is not fully conceptually sound. This way of thinking bears some remnants of a cognitivist world view in that it presumes experience to come in one of five (or so) pre-defined modal flavours and that these get swapped over when training with sensory substitution devices.

(Lenay *et al.*, 2003) criticise the term 'sensory substitution' for the described technology as "misleading and in many ways unfortunate" (Lenay *et al.*, 2003). Under close conceptual scrutiny, it becomes clear that a) what people with sensory disabilities gain from this technology are not senses (i.e., receptors), but new perceptual qualities and that b) there is no substitution of the absent sense but rather an augmentation or supplementation of the perceptual world. Thus, what can be observed is much more interesting than simple substitution of missing sensors. 'Real' sensory substitution (e.g., cochlear or retinal implants) have received much less attention in cognitive science literature because they lack the following characteristic:

> "These tools [sensory substitution devices] make it possible to follow with precision the constitution of a new sensory modality in the adult. In particular, by providing the means to observe and reproduce the genesis of intentionality, i.e., consciousness of something as external (the 'appearance' of a phenomenon in a spatial perceptive field), these tools make it possible to conduct experimental studies in an area usually restricted to philosophical speculation" (Lenay *et al.*, 2003).

(Lenay *et al.*, 2003) propose, therefore, to use the term 'perceptual supplementation' (*suppléance perceptive*) rather than 'sensory substitution'. Bach-y-Rita acknowledges a similar conceptual limitation of the term when remarking that the applications for this tech-

nology are open-ended and "could be considered to be a form of sensory augmentation (i.e., addition of information to an existing sensory channel)" (Bach-y Rita *et al.*, 2003) rather than just a substitution, a proposal that explicitly underlies (Nagel *et al.*, 2005)'s research on human adaptation to an artificial compass sense.

Taking sensorimotor theories of perception seriously means to get rid of the obsession with sensory channels. It can only sensibly be asserted that there are three senses (chemical, mechanical and thermal) or otherwise, it has to be accepted that there are infinitely many senses. This is not to deny that certain classes of experiential qualities are associated with certain classes of perceptual activity or certain sensors. In any one case, the dependence on the physiology of certain organs can be very strong (e.g., sense of pitch) or very weak (e.g., sense of simultaneity). It is just the application of the idea that outside the cognitivist premise, no *a priori* link between the mechanical level (types of receptors, neural pathways, cortical areas) and the functional/meaning level (infinitely many senses, such as sense of colour, direction, shape, posture, time . . .) can be presumed. Such differences in quality are part of the *explanandum* and should thus not be evoked, without justification to form part of the *explanans*. Most (if not all) modalities are multisensory in the sense that they involve sensation and motion, and, thereby require integration of the kinaesthetic sense (Gapenne, forthcoming).

The term 'sensory substitution' and its interpretation in the literature has led to misunderstanding and antagonistic reactions. (Prinz, 2006)'s critical response to (Noë, 2004)'s book 'Action in Perception' exemplifies such unfortunate misunderstandings: Prinz writes that, in order for TVSS systems to provide evidence for enactive theories of perception, it must be shown that "experience of using the apparatus is like vision, and [. . .] that it takes on this visual quality in virtue of the fact that subject learn to associate its inputs with the kinds of motor responses that are usually reserved for vision" (Prinz, 2006). Prinz accepts evidence for the latter condition but "seriously doubt[s] that these subjects experience anything visual" (Prinz, 2006), pointing out that experience of distal objects through tactile sensors forms part of our natural perceptual experience already, such as "when we tap an object with a cane we feel its shape and texture; when we drive, we feel the surface of the road" (Prinz, 2006). Prinz' observations are fully in line with the positions argued by (Bach-y Rita *et al.*, 2003) and (Lenay *et al.*, 2003), who explicitly draw the connection between the technology they employ and more rudimentary devices such as a blind person's cane. This veridical observation, however, does not "[put] the Brakes on Enactive Perception" (title of Prinz, 2006) but much rather puts the brakes on the slightly misleading interpretation

of sensory substitution technology that Noë provides; an interpretation that is suggested by the misleading label 'sensory substitution'.

The question is then: how can perceptual supplementation (PS) be both, the addition of a new sensory modality and skilled-tool use? The answer to this question is counter-intuitive at first: because perceptual modalities *are themselves* skills, namely the skilled use of the tools we are born with, and whose mastery we acquire during development: our eyes, our fovea, our nose, the palms of our hands, our tongue, our fingertips, our ears, ... Similar views have been proposed by others (e.g., Myin and O'Regan, 2002; Grush, 2007; McGann, forthcoming). There are many open questions around such proposals: how do we define a modality? Is there any distinction to be drawn between using our senses and using tools? We cannot put our sensory modalities out of hand: they are always mediating our experience, always its vehicle, whereas the tools we manufacture can be both, vehicle, but also content of our experience, when we put them down and look at them.[2] Is this what distinguishes using a tool from a perceptual modality? The bottom line is that, in the absence of good definitions to distinguish skills, tool mastery, perceptual modality, *etc.*, we have to see all these on a conceptual continuum. This implies that PS research is not in any fundamental way different from ordinary research on perceptual learning, skill learning and sensorimotor recalibration. Admittedly, constitution of new modalities and recalibration of existing ones are not the same thing. However, particularly where drastic sensorimotor perturbations are involved (e.g., adaptation to prismatic vision, Kohler, 1962; Welch, 1978), there is a clear continuum in the degree to which the qualitative experience of our perception prior to the introduction of a new or modified coupling resembles the perceptual experience acquired through training. The beneficial characteristics of PS technology identified in this chapter and throughout the book, therefore, extend to other areas of research in sensorimotor recalibration and perceptual learning that take a similar minimalist approach or rely on simple simulated environments.

To put PS technology on a continuum with more established research areas in human perception is not to sell short its potential as a novel tool for cognitive science. (Lenay, 2003)'s habilitation *Ignorance et suppléance : la question de l'espace* exemplifies the merits of this approach. It presents results from a series of experiments using PS experiments to investigate the fundamental basis of spatial experience. The approach the group has taken in investigating this question is reminiscent to the minimalist approach to ER simulation modelling described in Sect. 3.3. This minimalism that the approaches share can be described

[2]This corresponds to the modi of *vorhanden* and *zuhanden* in (Heidegger, 1963).

a 'throwing as much bath water out as possible, whilst keeping the smallest possible baby' (and expression borrowed from I. Harvey, personal communication), i.e., to find the simplest possible system to bring about the effect one is interested in and distinguish it from a minimally different one that does not. Simplifying PS technology to the extreme (one photo-receptor attached to the finger of a participant's hand that produces a bit sequence of on-off tactile signals), the group have identified the true minimal condition under which the described changes in perceptual experience occur (in particular, exteriorisation, i.e., perceiving the cause of a tactile proximal stimulus to be at a distance in 3D space): a minimal movement space of two joints and continuous swaying movements as strategy have been identified to lead to the perception of a stimulus as distant and 'out there', whilst one-jointed movement or lateral displacement of the receptor evoke the sensation of proximal touch. The rules of sensorimotor contingency that underlie the perception of distance have been formally mapped out and analysed. From this starting point, further experiments are conducted, building up gradually the complexity of the task, the sensory signals and the motion possibilities.

The experiments on perceptual crossing and the origins of perceived agency by the same group are described and modelled in chapter 6 and 7. They follow a similar minimalist agenda, starting from the simplest scenario possible (one-dimensional environment, Auvray *et al.*, 2009) and incrementally complexifying the experimental set-up (two-dimensional environment, Lenay, Rohde & Stewart, in preparation) to identify differences and similarities and explain them in terms of sensorimotor dynamics. The experiment on sensorimotor recalibration of perceived simultaneity (chapters 9-11) set out to follow a similar minimalist agenda to explore the origins of experienced simultaneity.

3.5 The Study of Experience

In this section, the difficult methodological issues around studying and explaining experience as an object of enquiry are addressed: experience is an inherently subjective phenomenon, our own first person what-it-feels-like. Science, on the other hand, is about observation and measurement from a quasi-objectivist perspective. It uses third person methods of quantification and can therefore not be directly applied to subjective experience. Therefore, a purely scientific explanation of cognition that focuses on measurable variables is doomed to leave out one of its most defining characteristics, i.e., subjective qualities.

In the computational cognitivist paradigm, this problem has been widely dealt with by, more or less, ignoring it by conveniently reducing it to some physical event or correlate, even though it has been prevalent in the philosophy of mind (qualia debate). This reluctance to explicitly deal with the experiential aspect of cognition results from the historical context in which cognitive science arose, i.e., as an opposition to Behaviourism. Cognitive science could make the use of mentalistic language credible being armed with scientific rigour that introspectionist psychology was missing (cf. Sect. 2.1). While the aspiration to maintain scientific standards is honourable, it prohibits the study of first person non-measurable experience, a central aspect of cognition and essential for the definition of many mentalistic concepts and distinctions. Cognitive science, therefore, finds itself in denial, trying to deal with experiential phenomena whilst pretending not to be dealing with experience.

The neurophenomenological approach developed by (Varela, 1996) argues how, within the enactive paradigm, first and third person methods can be combined in order to interdisciplinarily tackle problems of experience. Section 3.5.1 gives a short outline of phenomenology as a first person method and introduces Varela's argument, concluding that this approach is preferable to approaches that claim to be purely scientific. Section 3.5.2 suggests that other methods in general and, in particular, perceptual judgements as in psychophysics may be applied in a similar spirit as crude 'second person methods' in some circumstances.

3.5.1 *First and Second Person Methods to Study Experience*

(Chalmers, 1995) coined the term 'the hard problem' for the paradoxical difficulty that representationalist cognitive science has in explaining the existence of experience: computational theories of mind can describe functional mechanisms that bring about physically measurable results that share certain structural similarities with physically measurable variables in the brain or human behaviour, which again correlate with the occurrence of particular classes of conscious experiences. But, having a functional and mechanistic description of this kind, the question that remains is: why should such a functional unit produce experience at all, rather than just to perform its mechanistic function without experience? This problem is also referred to as the 'qualia' problem or 'the explanatory gap' (Levine, 1983). Physical *correlates* of mental acts can, to a certain degree, be identified, but they do not *causally explain* the occurrence of conscious experience. From within an approach whose explanatory domain is the material and functional, conscious experience appears to be an unnecessary and causally irrelevant extra, an epiphenomenon. Or, if it bears a functional role, this role can be formally described, reproduced and inserted into the model as a new

functional module – but this again raises the question of why there should be any experience at all, leading to a *regressus ad infinitum*.

In a response to (Chalmers, 1995)'s statement of the hard problem, (Varela, 1996) proposes his neurophenomenological approach as a remedy. He briefly reviews existing theories of consciousness, characterising them along four axes (including the prevailing functionalist approaches; the reader is referred to this scale for details about how neurophenomenology relates to existing theories of consciousness). One of the groups is characterised as acknowledging that subjective first person experience is irreducible and also that it plays a central role in a theory of consciousness, which is the group that contains Varela's approach and the approach taken here.

Varela reappraises the classical phenomenological approach established by Husserl (e.g., Steiner, 1997, recent edition of Husserl's lifework ca. 1886-1938) during the *fin du siècle* which promotes phenomenological reduction (see below) as a method for the systematic exploration of one's own experiential world. Varela quotes Merleau-Ponty to establish a first intuition about the link between the first person study of experience and the scientific study of cognition:

> "To return to the things themselves is to return to that world which precedes knowledge, of which knowledge always speaks and in relation to which every scientific schematization is an abstract and derivative sign language, as the discipline of geography would be in relation to a forest, a prairie, a river in the countryside we knew beforehand" (Merleau-Ponty, 2002), cited in (Varela, 1996).

The reason why many cognitive scientists are uncomfortable with the phenomenological tradition is that it appears to be a variant of introspectionist psychology, which, through its lack of intersubjective and methodological standards, made it possible for Behaviourism to become powerful and prohibit the scientific consideration of mind and what happens between sensors and actuators (cf. Sect. 2.1).

There are certainly some commonalities between phenomenology and introspectionism. After all, they are both first person approaches. Varela is right, however, to point out that phenomenological reduction as a method is much more credible. Firstly, it explicates the reflexive and reductive aspect of the act of self-observation, accounting for the nature and source of the introspective activity, which introspectionism left implicit.[3] Secondly, by explicitly including methods of communication and description into the approach and acknowledging its reciprocal causal effect of shaping and modifying the experiential world,

[3]Steve Torrance (personal communication) rightly remarked that, in this sense, even the term 'introspection' is misleading: it suggests that observing the internal mind was just a shift in focus from observing the external world. The self-referential and reflexive nature of introspection would be clearer if it was called 'autospection'.

the results of phenomenological reduction can stabilise in one's own account and ultimately also become subject to social debate and inter-subjective consensus. Thirdly, Varela argues for the power of intuition, not as an erratic mood swing, but as stable common sense beyond logic that informs all aspects of our life, including scientific activity. This powerful role usually goes unacknowledged in objectivist world views and is at the root of scientist chauvinism and the discarding of first person methods. Fourthly, these standards of generating communicable descriptions, stabilising one's own experience and intuition and mastering the reflexive stance do not come naturally but require training and discipline. Phenomenological reduction is not in itself a *scientific* method of reproducible measurements. In the explication of techniques and issues, however, it certainly comes closer to scientific standards than naïve introspection.

The lack of appreciation of these merits, which, pragmatically, give it a clear competitive edge over naïve introspectionism, but not necessarily an ontologically different status, probably stems from failure to recognise just how bad naïve introspection performs in comparison. I, the author, can confirm that impression through my own personal experience. The point is not that introspection is fallible in the sense that it does not always concur with the 'objective' observer perspective – systematically and stably occurring illusions or misjudgements that bring the first and third person perspective in conflict, such as perceptual illusion or flashbulb memories (Eysenck and Keane, 2000, p. 226f) are as real an experience as me seeing the screen of my laptop in front of me right now and can be equally informative for understanding mind, if not more. The point is about the bad quality of spontaneous subjective description of experience and the lack of consistency and structure in naïve introspection. The experienced stability and consistency of our everyday perceptual and experiential world makes us believe that it is not a big deal to observe and report it.

Research with 'second person methods' (i.e., interview techniques to gather experiential reports) shows how wrong this assumption is.[4] Research on second person methods develops techniques that can, to a certain degree, compensate for the naïvety of individuals untrained in systematic observation and documentation of their experiential world and thus yield useful reports even from naïve subjects (e.g., Petitmengin, 2006; Vermersch, 1994). Petitmengin states the problem as follows:

> "How many of us would be able to precisely describe the rapid succession of mental operations he carries out to memorise a list of names or the content of an article, for example?

[4]The failure to gather useful data when straight-forwardly querying the experimental participants in the simultaneity experiment about their experience of the task (chapter 9) painfully confirmed this point: they were just baffled, shrugged and did not answer anything useful at all.

> We do not know how we go about memorising, or for that matter observing, imagining, writing a text, resolving a problem, relating to other people... or even carrying out some very practical action such as making a cup of tea. Generally speaking, we know how to carry out these actions, but we have only a very partial consciousness of how we go about doing them" (Petitmengin, 2006, p. 230).

Petitmengin gives a much more detailed account of the difficulties with untrained subjects reporting their experiences in the given source. If the reader is in doubt, it will be much easier to become convinced if he or she tries to generate a verbal report of the phenomenology of searching the cited article on the Internet – or just asking any person around them to report theirs. The result will be very poor because untrained introspectors suffer from "unstable attention, absorption in the objective, escape into representation, lack of awareness of the dimensions and level of detail to be observed, impossibility of immediate access" (Petitmengin, 2006, p. 239). Bringing together techniques from different areas, such as phenomenology, Buddhist meditation and research on consciousness taking as a mnemonic technique, Petitmengin has developed an interview technique that, so she argues, leads to reliable and verifiable experiential reports.[5] The most impressive proof of the effectiveness of this technique is from its application in non-pharmacological epilepsy therapy, where, using her interview techniques over therapeutic sessions, Petitmengin trains epileptic patients to become aware of and describe their experience of the 'aura' state preceding a seizure. Patients could thus improve their seizure anticipation and suppression skills, yielding a therapeutic effect comparable to pharmacological treatment (Petitmengin, 2005; Le Van Quyen and Petitmengin, 2002, also personal communication).

Having argued that the study of experience by skilled interviewers or skilled phenomenological reducers produces more useful and reliable experiences and experiential reports than just asking your neighbour, how can these results be linked to results from third person science without stepping into a reductionist trap? In order to explicitly link the experiential and physical aspects of cognition and to communicate this link, aspects of the experience have to be treated as observables or objects and to be included into the explanatory story. Experiential reports, as they result form second person techniques, or, if I report my own experience, from first person experiential exploration, can, to a certain extent, be treated as data in such an endeavour. It has, however, to be stressed that experience cannot be reduced to the act of reporting/measuring it. Such a step just serves as a method for interfacing two

[5]The fact that such an interview and its setting also influences and modifies experience is not *a priori* a problem. For an approach that aims at minimising this impact of the second person and comes close to 'experience in the wild' see (Hurlburt and Schwitzgebel, 2007).

types of generating knowledge, one that requires a first person approach and the other one that requires a third person approach, none of which can reduced to the other.

So, what can we say about how phenomena we experience subjectively and those we experience as objects relate? Varela remarks that "human experience [...] follows fundamental structural principles which, like space, enforces the nature of what is given to us as contents of experience" (Varela, 1996). Physical structures and regularities constrain and shape our experience. Experience may be subjective, but it is by no means arbitrary. We realise just how regular it is by studying how physical perturbations or events induce systematic changes in our experiential world, as, for instance, during development (e.g., Piaget, 1936), through pathological cases (blindsight, hemi-neglect, perceptual disabilities, PS, ...), under sensorimotor perturbation (e.g., Kohler, 1962) or through altered states of consciousness (e.g., Shanon, 2001).

Varela thus proposes a 'neurophenomenological circulation', whose objective he describes as seeking "articulations by mutual constraints between phenomena present in experience and the correlative field of phenomena established by the cognitive sciences" (Varela, 1996). He gives examples from the neuroscientific study of attention, body image, perceptual filling in, emotion, (Libet, 2004)'s work on voluntary action and his own neurophenomenological explanation of present-time consciousness (Varela, 1999).[6] Again, the most impressive demonstration of the power of this approach is to be found in its application to epileptology (Petitmengin, 2005; Le Van Quyen and Petitmengin, 2002): not only do we study how irregularities of neural activity lead to dangerous and painful seizures, we also study how they lead to altered experiences preceding the seizure (bottom-up causation). Through the skilled and systematic study of these experiences resulting from abnormal neural activity, the experiences can be transformed through behavioural therapy, which, ultimately, results in the alteration and control of neural activity (top-down causation).

An issue that is mentioned but, in my opinion, underdeveloped in Varela's account is the fact that presumably purely scientific accounts of consciousness do exactly the same thing, even if they pretend not to: "It makes us forget that so-called third-person, objective accounts are done by a community of concrete people who are embodied in their social and natural worlds as much as first-person accounts" (Varela, 1996). As a leftover from the behaviourist age, talking about experience or attempting its scientific study is an embarrassment, a cosmetic flaw, which is why the most radical followers of scientism prefer to claim experience does not exist (e.g., Churchland and Churchland, 1998). Research

[6]The latter two are presented in more detail in chapter 8.

that addresses experiential phenomena, such as the study of the neural correlates of consciousness (Metzinger, 2000), however, has to deal with it by necessity – something has to correlate, after all.

There is the clear danger that, in order to keep up the illusion of being fully scientific, research on the human mind, which is, at some level, always also research conscious experience does not explicate its methodological commitments in the first person realm and the presumed nature of its link to the physical. Ironically, the misguided aspiration for scientific rigour introduces conceptual gaps in the explanatory framework. "The line of separation between rigor and lack of it, is not to be drawn between first and third accounts, but rather on whether a description is based or not on a clear methodological ground leading to a communal validation and shared knowledge" (Varela, 1996).

3.5.2 *Perceptual Judgements as Second Person Method?*

In the conclusion of his proposal of neurophenomenology, Varela writes

> "[. . .] every good student of cognitive science who is also interested in issues at the level of mental experience, must inescapably attain a level of mastery in phenomenological examination in order to work seriously with first-person accounts" (Varela, 1996).

Many of the enactive researchers cited in this volume – including the author herself – come short of this criterion. How can perceptual experience be studied without undergoing phenomenological training? Is the approach taken here really enactivist, despite this ignorance?

A more naïve approach to experience is proposed in the following. This section promotes perceptual judgements as they are used in human psychophysics as a set of crude second person methods that, in combination with the minimal modelling and experimental approach sketched, can form part of a truly enactive and interdisciplinary explanation of certain perceptual phenomena. Going back in history, similarities between the program of psychophysics and the neurophenomenological program are identified. This comparison is not in all aspects fully developed. It is more to be seen as an emancipation against the somewhat chauvinistic statement by Varela cited above. Using perceptual judgments, you can surely not capture the richness of the perceptual experiences concerned – but this does not mean that you can say nothing about mutual constraints between the experiential and the material realm in a neurophenomenological spirit.

The original statement of the psychophysics research program through the publication of *Elemente der Psychophysik* in 1860 by Gustav Fechner (Fechner, 1966) is in some ways

strikingly similar to Varela's statement of the neurophenomenological approach. Against the dominant Cartesian currents at his time, Fechner thought of the mental and the physical as two perspectives of the same thing, like the inside and the outside of a circle, or the heliocentric as opposed to the geocentric perspective of the universe. With reference to Descartes' allegory of the mental and the physical as two clocks that are perfectly synchronised, he remarks that the easiest possibility, i.e., that it is actually just one clock, had not been taken into consideration (Fechner, 1966, p. 4). This perspective implies that asking how one realm links to the other (such as by one being reducible to the other) is an ill-posed question. He also recognises the importance of the observer status of the scientist (cf. Sect. 3.1):

> "What will appear to you as your mind from the internal standpoint, where you yourself are this mind, will, on the other hand, appear from the outside point of view as the material basis of this mind. There is a difference whether one thinks with the brain or examines the brain of a thinking person. These activities appear to be quite different, but the standpoint is quite different too, for here one is an inner, the other an outer point of view" (Fechner, 1966, p. 3).

Applying these ideas to methods of enquiry he remarks:

> "The natural sciences employ consistently the external standpoint in their consideration, the humanities the internal. The common opinions of everyday life are based on changes of the standpoints, and natural philosophy on the identity of what appears double from two standpoints. A theory of the relationship of mind and body will have to trace the relationship of the two modes of appearance of a single thing that is a unity" (Fechner, 1966, p. 5).

Fechner describes the goal of psychophysics enquiry to answer questions like: "what things belong together quantitatively and qualitatively, distant and close, in the material and the mental world? What are the laws governing their changes in the same or in the opposite directions?" (Fechner, 1966, p. 8). This formulation has clear parallels in the neurophenomenological approach.

Where the two positions deviate is in recognising the importance of closed loop dynamical brain-body-environment interactions:[7] Fechner is revealed as a localist when describing his vision of how an 'internal psychophysics' of brain physiology would help to identify the direct functional correspondents of sensations, whereas the 'external psychophysics' method he develops and applies 'only' investigates correlations that are mediated through bodily states. Similarly, the methods outlined by Fechner are very much restricted to linking sensory stimuli ('inputs') to experience and do not allow for the inclusion of actions or motion into the psychophysical story. In this sense, the original formulation of the psy-

[7] Please note that Poincaré was six years old at the time of the publication of the 'Elements of Psychophysics'.

chophysics project is in tension with the enactive or neurophenomenological approach that emphasises dynamics and circular causality across several dimensions.

Nevertheless, Fechner's painful awareness of how this method and the language he adopts lend themselves to dualistic interpretation, contrary to his own view of the nature of the link between the mental and the physical, his repeated reassurance that the proposed method produces valid results immaterial of metaphysical questions, whilst hoping and believing that this method would ultimately produce results to confirm his views in a remote future, are at least as much at odds with classical representationalism. He certainly did in no way encourage homuncular and representationalist interpretations of his approach, like Baird and Noma's statement that the key question of psychophysics was "how does the human being use sensory and cognitive mechanisms to perceive the type and amount of stimulus energy" (Baird and Noma, 1978, p. 2)

Like Varela, Fechner puts his methodological commitments in both the physical and the mental realm open on the table and makes clear how they relate. In the experiential realm, psychophysics investigates and measures perceptual *detection, identification, discrimination* and *scaling* (Ehrenstein and Ehrenstein, 1999). The techniques for measuring these perceptual judgements have been used and developed for more than a century: a powerful set of formal tools (signal detection theory, techniques for psychometric curve-fitting, ...) are associated with the discipline of psychophysics.

As stated above, the reason why we can study cognition interdisciplinarily is that, from the observer perspective, experience relates to physical constraints and sensorimotor invariances. In some cases, these constraints are so strong that they lead to reliable, verbally expressible and intra- and intersubjectively stable results without the need of an expert interviewer or experiencer. The methods to explore the experiential domain that Fechner proposes and that have been developed since do not go as deep as phenomenological reduction.[8] However, the reason why the psychophysics approach can address questions of perceptual experience is that it deliberately confines itself to experiential phenomena and judgements that are so primitive that they lead to stable results despite the naïvety of the experiencers investigated.[9] The limits of applicability of the methods are inherent: psychophysics assumes a continuous mapping between a physical variable (e.g., stimulus

[8]Please note that Husserl was one year old at the time of the publication of the 'Elements of Psychophysics'.

[9]Perceptual judgements do not always reflect experience well. For instance, in 'forced choice' paradigms, many times, participants give accurate perceptual judgements for stimuli close to detection thresholds without experiencing this accuracy – see (Dienes and Seth, forthcoming) for an overview over techniques of measurement and their contingent relation to experience or (Gallagher, 2005) on pre-noetic influences on cognitive performance. This limitation has to be born in mind and made explicit.

energy) and experienced stimulus intensity that is quantified using a perceptual response profile – this kind of link is evident in some cases, but not in others.

The major advantage of studying such basic aspects of perceptual experience *qua* perceptual judgment behaviour is that the acts of performing, observing and reporting perceptual judgements form part of everyday human life: 'do you perceive this?' or 'is this bigger than that?' are very usual questions to be asked in everyday life. Therefore, we are all trained experts in these first/second person techniques. Furthermore, the inclusion of these perceptual judgements into a scientific framework has not faced a lot of controversy: their quantification makes obvious and intuitive sense, without the need for metaphysical and ontological agreement between different researchers or between researcher and audience. Similar methods have also been used in infants ('high amplitude sucking') and animals (e.g., Melchner *et al.,* 2000), leading to speculation about their perceptual worlds without appearing to cause a lot of uproar.

The common-sense-ness of this method is both its strength and its weakness. Psychophysics can be easily hijacked by representationalists, elimininativists and behaviourists. Results can be easily integrated into any such framework, as it already appeared to have happened to Fechner 15 decades ago. Asking and recording perceptual judgements, which is sold as a second person method here, can easily be treated like just another physical variable to be explained. Using perceptual judgements, you can always retreat to a behaviourist stance. Psychophysics thus acts as a 'neutral territory': it works regardless of ideological commitment, even if the exact way of investigating phenomena using psychophysics methods and the interpretation of results will be contingent on the choice of paradigm. This uncontroversial nature of psychophysics research can also be seen as its strength: results thus generated will not encounter a lot of resistance on political grounds and may thus help to communicate and illustrate results conducted under the enactive paradigm, which, evidence permitting, will ultimately benefit its establishment and the refutation of the classical view.

Advances in technology and mathematics allow the extension of the third person methods associated with psychophysics not just to neurophysiology, but also make its incorporation into more situated and dynamical research programs possible. (Rodriguez *et al.*, 1999)'s work on neural synchrony and shape recognition, as well as (Libet, 2004)'s neuroscientific study of volitional action, which (Varela, 1996) mentions in his statement of the neurophenomenological approach, are very close to what Fechner imagined as 'internal psychophysics'. Similarly, the PS work conducted by the CRED group (e.g., Auvray *et al.*,

2009; Lenay, 2003) includes dynamical and environmental factors, but, in linking perceptual response probability distributions to physical factors, still follows a similar agenda as psychophysics.

Obviously, the point here is not to argue *against* the more sophisticated neurophenomenological approach Varela envisions – a psychophysics approach can surely not be taken towards all dimensions of experience. However, neurophenomenology using Husserl's techniques of phenomenological reduction should not be seen as a privileged method in enactive cognitive science where first person experience is concerned. There are alternatives, of which one of the most basic is proposed here, all of which have their scopes and limitations. Recording perceptual judgments does not go as deep as phenomenological reduction or the mentioned interview techniques, but it has the advantage that they work for everyone without the need for training. Also, they do not involve a transformation of experience through the act of observing it that would go beyond the kind of transformations such acts of self-observation induce on a daily basis. Similarly, there are surely experiential phenomena for which the contemplative and reflexive character of practicing phenomenological reduction is unsuitable. (Varela, 1996)'s demand to explicitly commit to methods in the first or second person realm is to be taken seriously. However, this does not imply that Husserlian techniques of exploring one's mind have to be endorsed unreservedly – other ways to account for the first person realm should be considered and developed.

3.6 Combining Experimental, Experiential and Modelling Approaches

Having introduced the empirical, synthetic and subjective methods individually, it seems quite clear how they would work together as an alternative interdisciplinary framework. In this section, the links between these different approaches are made explicit in order to discuss three issues: firstly, the differences between the classical reductionist and the non-reductionist enactive approach. Secondly, the status of simulation modelling in the enactive paradigm. Thirdly, the difference between true interdisciplinarity and mere multidisciplinarity.

In a simplified view on the classical computational cognitivism, AI modelling forms the intellectual centre-piece of a reductionist program (see Fig. 3.5 (A)): philosophy establishes the relation between 'qualia' and neural states, which ultimately results in a reduction of the mental to the physical based on functional causal role. This reduction is via a formal AI model which captures the essence of brain functionality and which, in principle, could be variably instantiated; its 'wet-ware' basis, studied by neuroscientists, is just the way cog-

nition happens to be implemented in nature. In this reductionist view, scientists can quite happily confine their work to either of these levels, only occasionally making reference to findings from levels below and above: ultimately, the functional/behavioural level does not depend on its implementation or the mental states it produces. In that sense, this approach is multidisciplinary, rather than interdisciplinary.

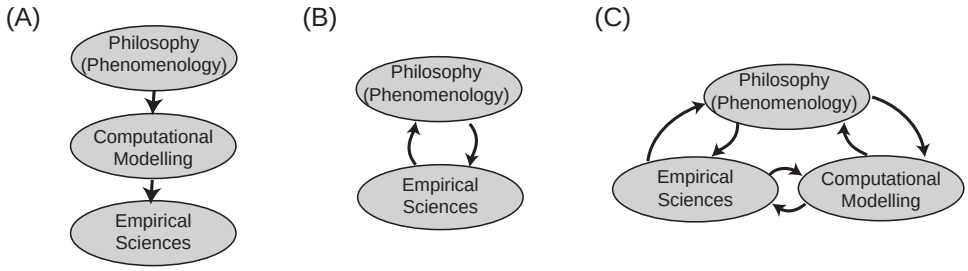

Fig. 3.5 Illustration of interplay between disciplines in (A) computationalism, (B) neurophenomenology and (C) the approach proposed in this book that includes simulation models.

The enactive paradigm as a paradigm of non-reductive naturalism, does not have an intellectual centre-piece: as argued in the previous Sect. 3.5, first/second person methods and third person methods are in an active and circular polylogue, exemplified in the neurophenomenological approach (see Fig. 3.5 (B)) and thereby truly spans levels of explanation, integrating them and requiring proper interdisciplinary activity.

Stewart identifies as one of the two basic requirements for a paradigm in cognitive science (besides resolving the mind-body problem) that "it must provide for a genuine core articulation between a multiplicity of disciplines, at the very least between psychology, linguistics and neuroscience" (Stewart, forthcoming). What is remarkable about this list is that synthetic methods or computer science, from having formed the intellectual centre-piece in the computationalist approach appear to have dropped out of the list altogether. Apart from promoting the enactive paradigm against the prevailing computational paradigm in cognitive science, reaffirming the place of computer modelling within the enactive approach to cognition, not as the centre-piece, but as an equal contributor, is one of the core objectives of this book (Fig. 3.5 (C)).

In Varela's early work, simulation modelling in the spirit outlined above (Sect. 3.3.2) formed an essential component, as most noticeably reflected in the computational model of basic autopoiesis (Varela *et al.*, 1974). From an initial enchantment with the ALife paradigm in AI, which (at least in some variants) is ideologically so close to the enactive

approach (cf. chapter 2), enthusiasm in the enactivist community appears to have cooled down significantly over the decades. The more recent formulation (Varela, 1996) and application (e.g., Le Van Quyen and Petitmengin, 2002; Rodriguez *et al.*, 1999) of the neurophenomenological approach (cf. Sect. 3.5) does not make explicit mention of computational methods or simulations at all.

Part of the responsibility for this trend is probably to be found in the ALife community, which, with few exceptions, has increasingly closed in on itself and not sought association with empirical sciences in general (cf. Webb, 2009, for a criticism) and enactive cognitive science in particular. The area has thus created a methodological bubble in which interdisciplinary links are, if it all, mainly sought with branches of chemistry, ethology and biology that do not associate themselves directly with the enactive approach or the study of cognition, even though autopoiesis theory was originally one of its main inspirations. The evident explanatory power of simulation models (cf. Sect. 3.3.2) has triggered integration of computational techniques in an enactive cognitive science that run outside the ALife paradigm (e.g., Stewart and Gapenne, 2004). However, it is undeniable that computer science is a marginalised discipline in the current enactive cognitive science.[10]

This book shows how generative ER modelling fits into Enactivism, in particular by pairing it up with equally minimal approaches to the study of human perception. As the branch of perception research outlined (Sect. 3.4) tends to use similar virtual environments as those employed in ER simulation, no strong abstractions of the behaviour modelled have to be undertaken in order to implement the envisioned interdisciplinary agenda. By virtue of this close match between model and experiment, ER simulations can be both generative models and descriptive models in the more traditional sense of computational modelling (i.e., other than their merely conceptual counterparts, they can also generate concrete and quantitative descriptors and predictions).

The second key advantage of the triangular approach envisioned here is due to the possibilities of PS and sensorimotor recalibration research as a stand-alone method (cf. Sect. 3.4): in studying the sensorimotor basis of perceptual experience, PS involves methodological circulation between empirical and experiential methods, which, in the spirit of (Varela, 1996)'s neurophenomenological approach, can naturalise aspects of perceptual experience. Explicit commitment to simple measures of perceptual experience, such as the perceptual judgements used in psychophysics, as crude first/second person methods is encouraged for the reasons given earlier. The difficult study of the dynamics of sensorimotor behaviour

[10] See (Fröse, 2007) for a discussion of the role of AI in the enactive approach.

and contingencies (cf. O'Regan and Noë, 2001) becomes more accessible, more formal and more transparent if ER models are introduced into the picture. This is due to the potential of ER models to bring us beyond our cognitive limits and prejudices (cf. Sect. 3.3).

The research presented in this book builds itself up by activating, step by step, the mutual links between the disciplines of the framework proposed (Fig. 3.5 (C)). The examples given demonstrate that the common root of ALife and the enactive paradigm has not yet been cut: the results on motor synergies (chapter 4) illustrate the mutual link between simulation modelling and the empirical experimental sciences, where models can generate descriptive concepts, proofs of concept and generate hypotheses for further experiments. The model on value system architectures (chapter 5) illustrates how simulation models can serve as extended thought experiments in philosophical and conceptual debate, pointing out implicitly held prior assumptions and counter our intuitions. The models of perceptual crossing in a one-dimensional (chapter 6) and a two-dimensional (chapter 7) environment models PS research that in itself adopts a circular and enactive method (Fig. 3.5 (B)) and therefore shows how simulation modelling can take part in a properly interdisciplinary polylogue, where all arrows in the diagram in Fig. 3.5 (C) are active. The study on adaptation to sensory delays and perceived simultaneity (chapters 9-11), finally, puts the idea to work that a cognitive scientist should really work interdisciplinarily, rather than to just contribute computer simulation models from a computer science-ivory tower. This proposal is relativised in the conclusion in chapter 12, that returns to the methodological issues this chapter opened, after the following eight chapters of application and results, with an overall optimistic outlook.

Chapter 4

Linear Synergies as a Principle in Motor Control

> The centipede was happy quite,
> Until the toad in fun
> Said 'Pray, which leg goes after which?'
> Which worked his mind to such a pitch,
> He lay distracted in a ditch,
> Considering how to run.
> (Anonymous)

This chapter presents the results from a simulation model that investigates a principle in motor control called 'motor synergy'. The term had been invented by the Russian physiologist and biologist Nicholai Bernstein (Bernstein, 1967) for systematicities between motion signals to control different effectors during one action. He proposes such systematicities as a principle that helps the nervous system to deal with redundancy in motor space. The modelling work here is directly inspired by experimental physiological work conducted by Gottlieb *et al.* in Boston and Indiana (Gottlieb *et al.*, 1997; Zaal *et al.*, 1999) on motor synergies in human target reaching. The results in this chapter have been published in (Rohde and Di Paolo, 2005). Other than the models presented later in this book (chapters 6 and 7 on perceptual crossing and chapters 8-11 on simultaneity perception), the model presented in this chapter is a strong abstraction from and idealisation of the original experiment conducted. However, in comparison to the more conceptual or philosophical model on value system architectures in the following chapter, the modelling approach taken in this chapter is still much more immediately applicable to scientific practice. The model presented in this chapter serves as an example of how simulation models can resonate with experimental research in the cognitive or behavioural sciences, with results that are meant to guide, inform and complement experimental work.

The theoretical, experimental and modelling background, as well as the research question to be addressed with the model are explained in Sect. 4.1. Section 4.2 introduces the

model, which investigates 'linear synergy' (i.e., a linear relation between torques applied to the elbow and shoulder joints) in a two-dimensional and three-dimensional simulated arm. Evolvability is compared for two dimensions of model complexity: dimensionality of Euclidean space and dimensionality of motor space (linear synergies). The results are presented in Sect. 4.3 and they show that, while dimensionality reduction through motor synergies increases evolvability in the given task, dimensionality reduction in Euclidean space decreases evolvability. These seemingly contradictory results on the usefulness of imposing and releasing constraints in the given simulation model are evaluated as to what they show for motor control and evolvability in general, as well as in the context of the experimental scientific work on human motor control in Sect. 4.4.

4.1 Motor Synergies

Motor synergies were proposed by (Bernstein, 1967) as a remedy to the degree-of-freedom (DoF) problem in motor control (Sect. 4.1.1). His biomechanical work has been the inspiration for many experimenters and modellers since it reached the English speaking world after the fall of the iron curtain in 1967 and the evidence for the existence of linear synergies in humans and animals is abundant. Section 4.1.2 presents two experimental studies that have been the direct inspiration for the model presented in this chapter and outlines the research question the model addresses.

4.1.1 *The Degree-of-Freedom Problem and Motor Synergies*

The rhyme with which this chapter starts nicely illustrates what (Bernstein, 1967) called the DoF problem. If the brain is thought of as a homuncular control organ that controls the state of all muscles and actuators centrally and simultaneously, the task it has to solve is very complex. Trying to describe animal or human behaviour in terms of joint kinematics already involves a large number of DoFs (e.g., 7 in moving an arm). This is the level of complexity aspired in main stream humanoid robotics, keeping many engineers and programmers employed full-time. The problem of controlling joint positions centrally, however, pales in comparison to the control problem of controlling a living human body centrally. Thinking of motor control in terms of individual muscles, or even motor neurons, the number of DoFs to be controlled when moving an arm quickly exceeds four digits (Bernstein, 1967). Also, while the joints used in robotics are usually exclusively sensitive to the motor signal by the robot controller, biological motor control has to be performed

in the presence of *context conditioned variability* (Bernstein, 1967, p. 246ff). The effect of a motor command is sensitive to the anatomical, mechanical and physiological context of the interaction of an agent with its environment, e.g., limb positions, passive dynamics or the state of the peripheral nervous system. Last but not least, the human and animal motor system is *redundant* with respect to the outcome of an action: there are infinitely many trajectories to proceed from a position A to a position B, a condition that Hebb has termed 'Motor Equivalence' (Hebb, 1949, p. 153ff). Humans and animals are very apt at compensating for perturbations, lesions or restraints on the motor system by using different effectors to perform the same functional behaviour.[1] This flexibility of goal-oriented motion is something that most state-of-the art robotic systems are still missing.

A homuncular view of how the body could be controlled from a central instance, like a puppet, was common at Bernstein's time, and explaining how a central organ could manage all this complexity at once seemed a big challenge.[2] Bernstein thought that *systematic relations between effectors*, a concept that he called 'motor synergy', was the answer to the DoF problem. The driver of a car can determine the position of both wheels of the car at a time because they are linked. This link imposes a constraint on the possible wheel positions. However, it only rules out useless wheel positions and does not functionally constrain the motion possibilities of the car. In a similar way, he thought mutual constraints in an organism's motor system could serve to build functional sub-units, thereby reducing the effective number of DoFs in a motor task in a beneficial way.

Motor synergies are evident in human and animal behaviour, ranging from human directional pointing (as described in the following Sect. 4.1.2) to different types of gaits, posture correction during breathing and hand motion in firing a gun (for a summary of findings see the chapters by Turvey, Fitch and Tuller in (Kelso, 1982)). Bernstein's idea of motor synergies also have strongly impacted on theory building and modelling work in cognitive science and motor control (e.g., Arbib, 1981; Grossberg and Paine, 2000; Morasso *et al.*, 1983; Sporns and Edelman, 1993).

From an enactive perspective on sensorimotor behaviour, the DoF problem, as defined by Bernstein, does not really pose itself because motor control is not thought of as the result of homuncular central planning. Also, this conception is not free from practical and conceptual problems: as argued in chapter 2, homuncular explanations typically pass the

[1] A famous example for this is the fact that characteristics of handwriting are preserved even when forcing a subject to write with its left hand, the mouth or the foot (Kandel *et al.*, 2000, p. 657).

[2] Bernstein used to demonstrate this to his students by asking them to assume the role of a homunculus and control a system he set up from sticks connected such that they had several degrees of freedom.

explanatory burden down: is explaining the brain as the 'driver of the bodily car' much easier than explaining the whole system in the first place? Also, (Weiss and Jeannerod, 1998) remark that "the context in which a motor task is executed strongly influences its organization" (Weiss and Jeannerod, 1998, p. 74). This appears to contradict the idea of functional and structural isolation of motor planning (homunculus) and execution (systematic co-activation of DoFs as functional sub-unit or building block).

However, in the light of the mentioned evidence for systematic relations between motor signals in different DoFs, questions about their nature arise: if motor synergies do not serve the purpose to decrease dimensionality for central motor planning, what is their functional role? How do they emerge from the redundant and high-dimensional movement space? Are they epiphenomenal? If they serve a purpose, how are they maintained?

4.1.2 *Directional Pointing*

The particular experimental study that inspired the simulation model here presented is a finding on *linear synergies* (a linear correlation between torques applied to the shoulder and elbow joint) in human directional pointing by (Gottlieb *et al.*, 1997). Targets were arranged spherically and equidistant from the starting position in the sagittal plane. Reaching these targets, the dynamic components of muscle torque (gravitational component removed) applied to the joints were scaled linearly with respect to each other during each target reach, with different scaling factors for different targets. This systematic relationship does not appear to result from the nature of the task, as it does not produce shortest paths or appear to satisfy any other obvious efficiency or performance criterion. For the remainder of this chapter, discussion will focus on such linear synergy. However, any systematic constraint simplifying motor space can be labelled a motor synergy.

(Zaal *et al.*, 1999) found the same systematic relationship between joint torques in infants even in the pre-reaching period, even though their attempts to grasp an object are unsuccessful. They investigated infants' reaching behaviour at several stages during their motor development, observing linear synergies throughout the stage-wise development of behaviour. Therefore, linear synergies do not appear to be the outcome of a learning process either. Zaal *et al.* conclude that "If linear synergy is used by the nervous system to reduce the controlled degrees of freedom, it will act as a strong constraint on the complex of possible coordination patterns for arm movement early in life" (Zaal *et al.*, 1999, p. 255).

Another finding that has to be born in mind is that there are behaviours in which humans learn to break linear synergy. For the case of arm movement, for instance, Weiss and

Jeannerod's review on grasping and reaching studies observes that sometimes Cartesian space dominates motor organisation, whereas in other cases (such as in (Gottlieb *et al.*, 1997)'s study), joint space dominates the organisation of trajectories (cf. Weiss and Jeannerod, 1998). Therefore, linear synergy does not appear to be a mere fixed physiological constraint on possible arm movements either.

As outlined in Sect. 3.3, one of the methodological advantages of ER modelling is that it does not presume a fixed relationship between the mechanical organisation and functional organisation. Previous modelling approaches to motor synergies (e.g., Grossberg and Paine, 2000; Sporns and Edelman, 1993; Morasso *et al.*, 1983) built in a functional role for synergies, i.e., as a movement building block for composition of complex motion. To the contrary, the ER model presented in this chapter aims at *evolving* the functional role of linear synergies in a minimally biased way to explore their functional role with minimal prior assumptions. Where do linear synergies come from? Under which circumstances do they arise? Are linear synergies epiphenomenal to a structured agent environment interaction or do they serve a particular identifiable purpose in the control architecture of evolved agents?

If linear synergies are beneficial to the organisation of the modelled task, their existence will lead to an improvement in either performance or evolvability and an exploration of this advantage can generate hypotheses about their functional role in human motor control. Such hypotheses can be tested in further experiments. The simulation compares a two-dimensional version of the task with a three-dimensional version, to investigate the relation between redundancy in DoFs and spatial complexity. Four different kinds of neural controllers are compared, with and without built-in linear synergies (details are specified in the following Sect. 4.2) to investigate their functional role in and resulting from artificial evolution. The findings are in line with (Zaal *et al.*, 1999) in suggesting that linear synergy as a built-in constraint benefits an efficient developmental process.

This exploration is also relevant for robotic engineering and the technical side of ER modelling. In order to be minimal, many Evolutionary Robotics experiments typically do not involve high levels of redundancy. The results here presented demonstrate how imposing the *right* constraints along the *right* dimensions can impact on evolvability and the nature of the solutions evolved.

4.2 Model

This section, as well as the model sections in other chapters, contain technical details that may not be accessible to readers without a training in computational methods. Such readers are encouraged to skim-read over the formulae and parameter values in this section and in the results section, trying to get the gist of the task and platform and proceed to the more accessible discussion.[3]

A robotic arm is evolved to reach to one of six target spots on a horizontal plane. The simulation has been implemented using the 'Open Dynamics Engine' (ODE, Smith, 2004) to model rigid body dynamics. The simulated arm consists of a forearm, an upper arm (each two units long) and a spherical hand (Fig. 4.1, (A)). The six target points are spread evenly on the circumference of a circle with a radius of 1.25 around the starting position of the hand (Fig. 4.1 (B)). The required reaching direction is denoted by ϕ and uniformly distributed directional noise $\in [0, \frac{1}{6}\pi]$ is added to ϕ at each trial.

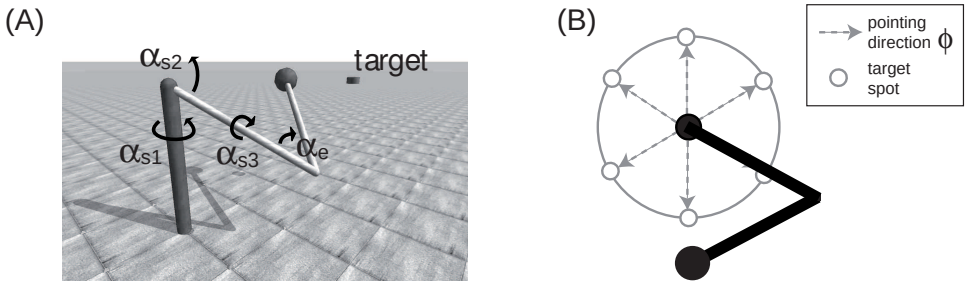

Fig. 4.1 (A) Visualisation of the simulated arm. (B) Plan view of the task (schematic).

The arm joints are referred to by their joint angle α_x (see Fig. 4.1 (A)). In order to test the effect that the number of degrees of freedom (DoFs) has on the task, experiments are run on a planar (i.e., two-dimensional) condition where both the elbow and the shoulder joint have one DoF (α_e and α_{s1}) and a three-dimensional condition, in which the elbow joint has one DoF (α_e) and the shoulder joint has, just like the human shoulder, three DoFs: rotation in the horizontal plane (α_{s1}), lifting/lowering the arm(α_{s2}) and rotation along the arm direction (α_{s3}). All joints are controlled by applying a torque M_i to the joint α_i.

The arm is constrained by plausible joint stops. Dry friction is applied at all joints. The networks have one sensory neuron for the angular position of each DoF and an additional

[3] Out of the models presented in this book, the current one is, arguably, the most complicated one. Readers should not be put off and try if they find one of the other modelling chapters more accessible.

sensory neuron for the required pointing direction $\phi \in [0, 2 \cdot \pi]$. The starting position of the hand is always at the middle of the circle, which corresponds to both the shoulder and the elbow angle starting at $\alpha_{e,s1} = 60^\circ$. The arm always starts with the elbow in the plane, even in the three-dimensional version of the task.

Some simplifications from the modelled scenario make the agent dynamics very much unlike the real-world example. In the three-dimensional environment, it is very difficult for evolution to keep the hand close to the plane, something which is automatically afforded by the two-dimensional environment. However, part of the objective of this simulation was to compare a two- and three-dimensional version of the same task. Therefore, the movement in the three-dimensional condition has been constrained such that the hand cannot deviate from the horizontal plane, meaning the possible hand trajectories are equal between the two conditions, but having more motor redundancy in the three-dimensional version. This restriction makes the movement more like moving an object across a surface (like moving a fridge magnet) than like natural human reaching movements. For similar reasons, gravity has not been modelled. These constraints reduce biological plausibility of the model. The strategies evolved are not always human-like, and it is not clear in how far the system could generate insights about concrete anatomical parameters that apply to a real-world physical system. The principal idea, i.e., how to explore the questions of redundant DoFs in a motor control task, is preserved upon introduction of these additional physically implausible constraints.

The weights of the CTRNN controllers evolved in the ranges $w_{ij} \in [-7, 7]$, the bias $\theta_i \in [-3, 3]$ and the time constant $\tau_i \in [0.1, 1.77]$ with a simulation time step of 0.01. Other parameter ranges are $M_G \in [0.1, 30]$ and $S_G \in [0.1, 20]$. Other than in most simulations, M_G and S_G were evolved individually for each DoF.

Four different neural controllers were evolved and compared for both the two-dimensional and the three-dimensional conditions. In the condition labelled as the *unconstrained* condition, a monolithic CTRNN with six hidden nodes per DoF and two output neurons for each torque signal $M_i = M_G(\sigma(a_{Mi+}) - \sigma(a_{Mi-}))$ is evolved (see network architectures in Fig. 4.2 (A)). $\sigma(x)$ is the sigmoid transfer function Eq. (3.3).

The *modularised* CTRNN has the same number of neurons, but connectivity is decreased, such that two sub-controllers generate the motor signals for each joint individually (see Fig. 4.2 (B)). They have three hidden neurons each and receive proprioceptive input only for the joint they control. However, they share the directional task input neuron. Note that in the three-dimensional condition, the shoulder sub-network still generates three motor

signals. Comparing results from the unconstrained monolithic and the modularised condition is interesting with respect to the question of *neural basis of motor synergies*. In principle, coordination between joint movements could be mediated through the environment and result from the task dynamics. If synergies emerge despite the absence of neural connections between the modules that generate motor signals for each joint, in the closed sensorimotor loop, such regularities pose a challenge to homuncular explanations.

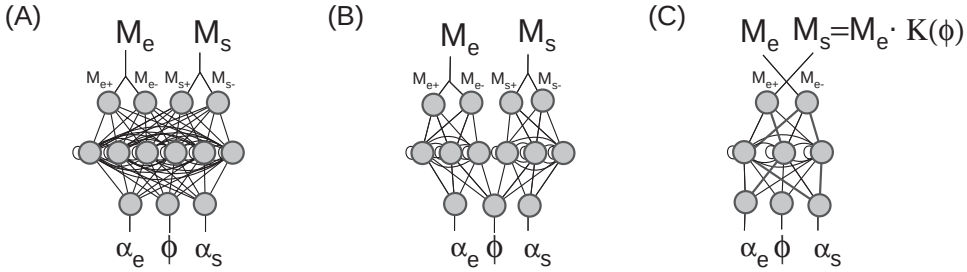

Fig. 4.2 Network diagrams for the unconstrained (A), the modularised (B) and the forced synergy (C) condition.

In the third and fourth condition investigated, a linear relation is imposed between torques applied to the elbow joint α_e and the different DoFs in the shoulder α_{si}. This type of controller is referred to as *forced synergy* controller. In these networks (see Fig. 4.2 (C)), M_e is generated by a CTRNN with three hidden nodes and all joint inputs. The other joint torques M_{sj} are scaled as a linear function $M_{sj} = K_j \cdot M_e$ where j is a DoF. $K_j = f(\phi)$ varies systematically with the desired pointing direction across trials.

Two different functional representations are used for the forced linear networks. In the *linear* forced synergy condition $K_j(\phi)$ is a simple linear function for each DoF j

$$K_j(\phi) = k_j^1 \cdot \phi + k_j^2 \tag{4.1}$$

with $k_j^i \in [-4,4]$ set genetically.

The more complex representation of the linear synergy function $K_j(\phi)$ as a Radial Basis Function Network (RBFN) is motivated by the fact that RBFNs are generic representations of continuous functions of the angle, i.e., it does not have a singularity at $\phi = 2\pi$ like Eq. (4.1). In the RBFN condition, $K_j(\phi)$ is represented by a RBFN with Gaussian RBFs

$$K_j(\phi) = \sum_{i=1}^{4} w_{Ri} \cdot e^{-\frac{\delta^2}{2 \cdot \Delta^2}} \tag{4.2}$$

where $\delta = c_i - \phi, \delta \in [-\pi, \pi]$ is the difference in direction between the evolved RBF center $c_i \in [-\pi, \pi]$ and the target direction ϕ. The width of the Gaussian RBF $\Delta \in [0.5, 1.5]$ and

the RBFN weights $w_{Ri} \in [-4,4]$ are also evolved. The absolute values of the coefficients $|k_i|$ and the absolute values of the RBFN weights $|w_{Ri}|$ are mapped exponentially.
The number of parameters evolved in each condition varies between 46 and 161 (see table in Fig. 4.3).
Trials are run for $T \in [2000,3000]$ time steps. The fitness $F_j(i)$ of an individual i on a target spot j is given by

$$F_j(i) = 1 - \frac{d_j(T,i)^2}{d_j(0,i)^2} \tag{4.3}$$

where $d_j(T,i)$ is the distance of the hand from the target spot j at the end of a trial for individual i.

	unconstrained	modularised	forced synergy (linear)	forced synergy (RBFN)
2D	109	75	53	46
3D	161	115	62	83

Fig. 4.3 Number of parameters evolved.

Networks for all conditions are evolved with either on all six target spots right from the start or, otherwise, in incremental evolution (i.e., they were evolved on just a sub-set of target spots, starting with two target spots, and the next clockwise target spot is added to the evaluation once the average performance of the population exceeds $\bar{F} = 0.4$). The evaluation of a network i on n target spots is calculated using the exponentially weighted fitness average defined in Sect. 3.3, Eq. (3.5).
Otherwise, the GA, numerical integration and CTRNN control are those described in Sect. 3.3 ($r = 0.6$ in the GA).

4.3 Results

The presentation of the results focuses on several key aspects. Evolvability is a variable that plays an important role throughout. It is mostly indicated as the number of target spots the network was evolved on in incremental evolution, as this variable corresponds to grades in performance. Section 4.3.1 compares the two- and three-dimensional version of the simulation across neural controllers focussing on the role of spatial redundancy. Section 4.3.2 compares the results from the different kinds of network controllers. The last section 4.3.3 takes a closer look at the linear synergy functions $K_j(\phi)$ evolved to solve the task.

4.3.1 *Number of Degrees of Freedom*

The problem of *motor redundancy* identified already applies in the two-dimensional version of the task: there are infinitely many trajectories to move the hand from position P_A to position P_B. However, for any position P, in this set-up, there is (due to joint stops) just one possible pair of joint angles (α_e, α_s) to realise it. In the three-dimensional set-up, due to the three DoFs in the shoulder joint, there are infinitely many shoulder positions $\alpha_{1,2,3}$ associated with a position P, even if the elbow angle α_e is not redundant. The space of motor signals to arrive at a configuration is even more redundant, due to the fact that the network generates torques, rather than angular velocities or joint positions, so different interfering forces (passive dynamics, interaction of torques applied to different joints through the body and the environment) work on each joint and affect the arm trajectory.

Averaged across 10 evolutionary runs, the motor redundancy afforded by the three-dimensional set-up provided a clear advantage in evolvability (see Fig. 4.4) in all network architectures: in the incremental evolution condition, the number of target spots reached is much higher.

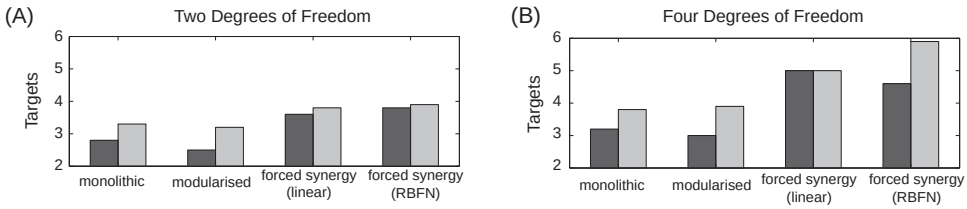

Fig. 4.4 Average number of starting positions reached in incremental evolution after 100 (dark) and 500 (light) generations across ten evolutionary runs.

Exploring the space of strategies evolved in case studies, the three-dimensional version shows a much greater variety of solutions than the two-dimensional version, where the only variation in strategies to reach a certain target spot is to temporally vary the torques applied to both joints in order to bring the two planar joints in the appropriate end positions. In contrast, the motor redundancy afforded by including two additional DoFs in the shoulder joints allows for a greater variety of strategies that exploit the additional DoFs and environmentally mediated forces. Among the monolithic and the modularised CTRNN controllers, a common strategy is to turn the arm along its length to one of the joint stops, leaving the hand in the centre of the plane, before moving to the target spot. It seems

that the positions thus reached are more suitable for evolutionary search and directional reaching than the original starting position.

To gain further insight into the mechanisms of the evolved solutions, the robustness of controllers to disabling individual DoFs was investigated. To investigate the role of passive dynamics, the different conditions were compared: in condition $F'_a(i)$ individual DoFs were 'paralysed', i.e., passive dynamics were possible, but no motor torques were applied. In condition $F'_b(i)$, individual DoFs were blocked, i.e., the joint angles were fixed at their initial position. Figure 4.5 shows the squared difference in performance $(F'_a(i) - F'_b(i))^2$ between these two conditions per DoF affected and network type. This measure indicates in how far passive dynamics contribute to the solution to a task where immobility leads to its total break-down. In the two-dimensional condition, enabling passive dynamics to work on the paralysed DoFs hardly make a difference in performance as compared to blocking the joint altogether (Fig. 4.5 (A)), i.e., passive dynamics plays a negligible role. In the three-dimensional condition, however, (Fig. 4.5 (B)), it has a noticeable impact on performance of all networks. The controllers evolved in the three-dimensional set-up, therefore, appear to make use of the motor redundancy and increased possibilities for passive dynamics in order to increase stability of the solution.

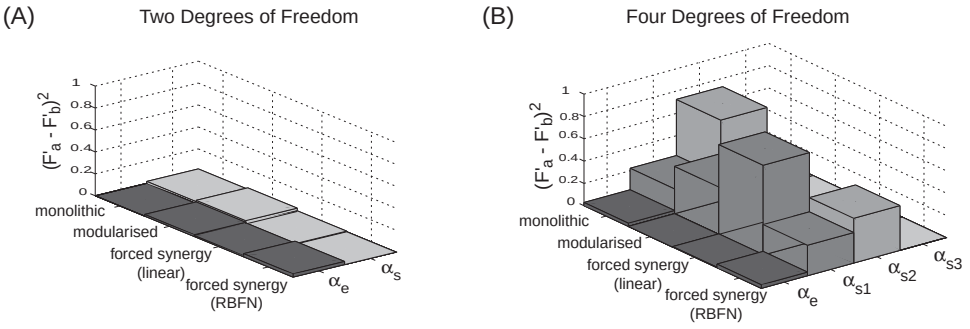

Fig. 4.5 Squared difference in normalised performance as individual joints α_i are paralysed (i.e., free to move but not driven) (F'_a) or blocked (F'_b) in example two-dimensional (A) and three-dimensional (B) agents evolved.

These findings from the simple simulation models show how, in a sensorimotor task, the inclusion of additional DoFs can increase evolvability. In the pursuit of minimalism, it is tempting to endow an agent with the minimally required sensorimotor system for a task, but such an idealisation can introduce a bias into the sensorimotor dynamics, delimit the strategies evolved and hamper evolution of high performing solutions, despite the reduction of the search space (cf. table in Fig. 4.3).

4.3.2 *Forcing Linear Synergy*

Comparing the evolution of an unconstrained monolithic or modularised CTRNN controller with the networks that were evolved to act in linear synergy, the agents forced to use linear synergies reach much higher levels of performance on average, both in the two-dimensional and in the three-dimensional condition. Figure 4.4 depicts the number of target spots that each network type evolved to solve in the incremental evolution condition. The RBFN synergy networks advance to the next goal twice as many times as the other networks. With twice as many generations, the CTRNN controllers without forced linear synergy come close but never reach the level of performance of the networks forced to act in linear synergy.

The only agents that evolve to solve the entire problem space are the RBFN synergy agents in the three-dimensional set-up; in all other conditions, evolution stagnates in a sub-optimal level of performance on a limited number of target spots, such that the population average does not exceed 0.4 to enter the next stage of incremental evolution. In the three-dimensional forced synergy condition, average performance of best individuals after 1000 generations is 0.65. Complementary non-incremental evolution led to qualitatively similar results, i.e., quicker and more successful evolution of forced synergy networks, even if, quantitatively, the overall fitness evolved was much lower in a non-incremental approach.

As explained in the model section, RBFN networks have been chosen because they appear to be particularly suitable for the task of transforming angular variables. Does this give the forced synergy networks an unjust advantage over the CTRNNs, is the superiority in evolvability and performance built-in, is it a question of design, not evolution? It could be true in the case of the RBFN, but certainly not for the case for a simple linear forced synergy condition. A simple linear function has a singularity at $\phi = 2\pi$. Given that the two-dimensional scenario is already very restricted, forcing this crude relation between task signal and required pointing angle makes it virtually impossible to generate a controller that masters the task. Despite this principal handicap, the solutions for all set-ups in which networks were forced to act in linear synergy evolved to much higher levels of performance than their unconstrained CTRNN counterparts.

To rule out the possibility that the simple CTRNN controllers (monolithic or modularised) could not cope with the presentation of the input direction as a scalar neural input, a more 'CTRNN friendly' set-up was tested, too, where controllers were provided with six different input neurons for the different target spots and no noise applied to ϕ. Still, neither in the two-dimensional nor in the three-dimensional condition did the agents advance beyond the

presentation of three target spots within 1000 generations. Something about functionally dividing the task into the generation of a torque signal and determining separately how this torque signal is scaled between the different DoFs seems particularly suitable for artificial evolution to efficiently evolve solutions for the given task.

4.3.3 *Evolved Synergies*

What kind of scaling implements the reaching to a target best? Across solutions evolved under the forced synergy condition, no general pattern, in terms of motion trajectories could be observed. Figure 4.6 depicts example RBFN synergies evolved for the three-dimensional condition. What is characteristic for many solutions is the fact that there are the different RBF centres, though there is overlap between the curves. This explains the diversity of behavioural strategies for different ranges of ϕ observed in the RBFN agents: for different targets, different DoFs are predominant in the realisation of the task.

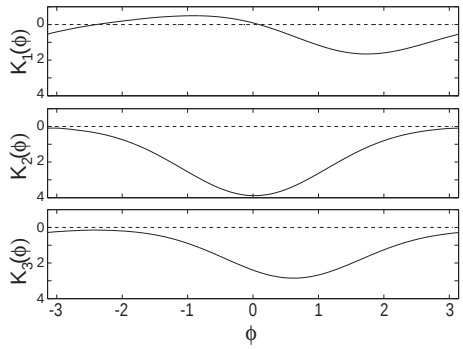

Fig. 4.6 An example evolved RBFN for a forced synergy network for the three-dimensional condition.

Imposing linear synergy increases evolvability of solutions. A possible explanation for this increase in evolvability is that such solutions are directly functionally beneficial for solving the motor task. If this was the case, an increase of linearity in torque relation could be expected as a result of evolutionary advance in the (monolithic and modularised) CTRNN controllers. Figure 4.7 (A) shows a measure of synergy in the networks that were not forced to act in synergy (i.e., the sum of squared error from linear synergy, i.e., perfect scaled co-activation across time, in these types of networks changes across evolution in the best individuals evolved for both the two- and the three-dimensional condition (average across five evolutionary runs). In the two-dimensional condition, there is a tendency to reduce this

error, i.e., to get closer to linear synergy, as performance increases. In the agents evolved for the three-dimensional networks, in contrast, linear synergy and performance appear to be completely unrelated.

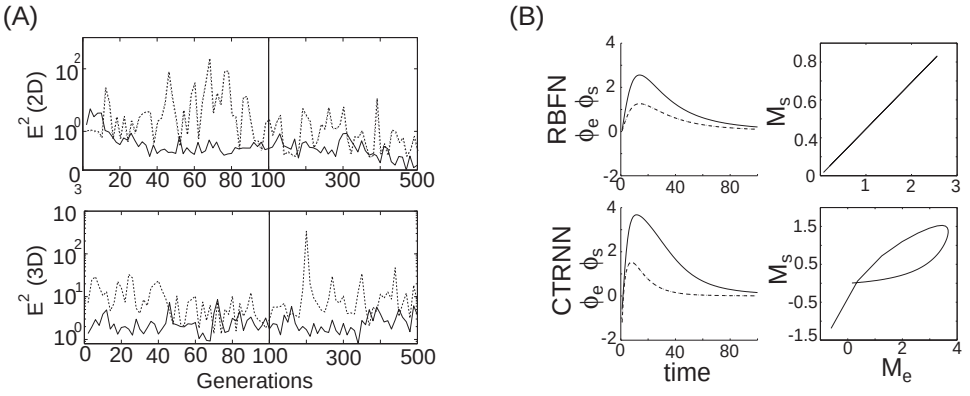

Fig. 4.7 (A) Sum of squared deviation from linear synergy across generations in the two-dimensional (top) and three-dimensional (bottom) networks. Solid: unconstrained, dashed: modularised; average across five evolutions. (Note nonlinear scales) (B): a two-dimensional monolithic CTRNN controller (bottom) applies a similar strategy as a RBFN forced synergy controller, yet the peaks in joint activation are temporally displaced, breaking synergy.

The modularised CTRNN controllers are on average much less prone to exhibit linear synergy (note logarithmic scale), even if there is a lot of variance in this variable. The reason to investigate this network architecture and compare it to the monolithic CTRNN controllers was that, if linear synergy was a generically good strategy in the task, this relation between the joint torques could have been implemented even without a neural structure controlling it, instead exploiting the environmental dynamics to achieve coordination. Given the exploitation of the environmental link for joint control using passive dynamics described in the previous Sect. 4.3.2, it is clear that the simulation used does, in principle support exploitation of environmental dynamics. However, linear synergies without a neural basis did not evolve. Also, being more disposed through neural connections to coordinate joint torques does not appear to provide the monolithic CTRNN controllers with an evolvability advantage (cf. Fig. 4.4 (A) and (B)). All these findings suggest that the magnitude of deviation from linear synergy is not an essential characteristic of a successful solution.

Figure 4.7 (B) shows how a monolithic CTRNN controller in the two-dimensional scenario applies a very similar strategy as a controller that is forced to act in linear synergy. The CTRNN controller emits motor signals to the two joints with a slight delay, as also repre-

sented by the loop in the M_e/M_s map. Such temporal displacement disrupts linear synergy as defined earlier, but this does not impact negatively on performance.

4.4 Discussion

The model abstracts strongly from the human original. Therefore, the model cannot be seen as a descriptive model in the traditional sense that can provide insight about the functional role of certain physiological features. However, the proofs of concept and hypotheses for further empirical experimentation it produces are informative. Firstly, linear synergies could not be found to be the outcome of an unconstrained evolutionary search process. Also, disconnecting controllers for different joints did not provide a disadvantage in evolvability compared to monolithic networks controlling both joints. This suggests that the mere possibility of implementing systematic relationships between effectors in a network does not provide a selective advantage.

On the other hand, imposing the constraint of linear synergy strongly improves evolvability of viable solutions, even if the function $K_j(\phi)$ that specifies the relation between the joint torques is a simple linear function (Eq. (4.1)), but even more so if this relationship is represented as a RBFN (Eq. (4.2)) that allows to define more complex and continuous functions of angles. The division of control into scaling and generation of a motor signal is suitable for evolutionary search in the given task. It is, however, unclear what exactly this benefit consists in. When analysing the ruggedness of the fitness landscape around successfully evolved individuals, no differences between the different conditions could be shown. (Decay profiles when applying mutations of increasing magnitude r had very high variance across controllers, immaterial of controller type, even if average levels of performance were comparable).

Arguably, the most interesting result from this model is that both a complication of the parameter space (i.e., adding more DoFs) and a simplification of the parameter space (i.e., forcing linear synergy) have provided independent evolutionary advantages. Thinking of the search space in numbers of parameters evolved (table in Fig. 4.3), it turns out that both the best configuration (three dimensions, RBFN synergy) and the worst configuration (two dimensions, monolithic or modularised CTRNN) are in intermediary range of evolved parameters. Improving evolvability is not a matter of scaling up or scaling down the search space, but of *reshaping the fitness landscape*. As tasks and robotic platforms become more complex, ER must produce appropriate reshaping techniques to scaffold the search process

and thereby solve the 'bootstrap problem' (Nolfi and Floreano, 2000, p. 13) and biology may be a suitable source of inspiration in searching such appropriate constraints.

The fact that both the monolithic and the modularised CTRNN controllers failed to evolve linear synergies suggests that this organisation of movements is not as such beneficial in the given task. The dramatic increase in evolvability that imposing linear synergies onto the movement space means proposes an explanation that is more in line with (Zaal *et al.*, 1999)'s conclusion that constraining the space of solutions by imposing linear synergy is a beneficial pruning of the space of behavioural possibilities for a developmental process (artificial evolution or motor development) to learn efficiently without delimiting movement possibilities severely. In order to further investigate this hypothesis, it would be interesting to study the phylogeny of linear synergy in evolutionary theory, or, as an extension to the experiments presented here, to evolve the constraints for ontogenetic development ('evo-devo' model), hypothesising that linear synergies would result from evolution in this meta-task.

Another interesting finding is that in the three-dimensional simulation, passive dynamics and redundant DoFs could be shown to be exploited, whereas in the two-dimensional version, the solutions evolved appeared to be less sensitive to environmental forces. The restriction of movement to the plane constrains behavioural possibilities much more drastically than imposing linear synergies between joint torques.

It has to be stressed that the results about the beneficial role of linear synergies do not automatically generalise to all kinds of tasks. To the contrary, it is quite obvious that, for instance, a two-wheeled robot doing obstacle avoidance (a simulation which is not redundant in DoFs) will rely on an ongoing change in the relation between the effectors. There is, however, a possible analogy to be drawn to physiological data again: as mentioned in the background Sect. 4.1, evidence from studies on human physiology suggests that linear synergy can be broken. Possibly, such a deviation from this unlearned principle of motor organisation is acquired if such variability in the relation between actuators serves the task.

Findings on systematicities between effectors, as they are ubiquitous in humans and animals, have been explored with an ER simulation model to investigate their function in an unbiased way. As concerns the scientific value of ER simulation models for the study of human behaviour and cognition, these theoretical insights generate proofs of concept (e.g., that a reshaping of motor space can aid a developmental process), which can be tested in further experimentation or explored further in simulation modelling. The descriptive con-

cept of motor synergies appears to be a useful one that can be integrated into an enactive story of motor control, even though it derives from a homuncular view on motor control. The exact functional role of motor synergies, however, remains unclear.

The feedback from the scientific community concerned with motor organisation was very positive. We communicated our results to the researchers that had directly inspired our work (Gottlieb and Zaal). In a follow-up study on joint torque covaration, researchers from the Gottlieb group refer to our simulation model as having "demonstrated that linear synergy was not a control solution converged upon by an unconstrained [...] neural network in order to reach the designated targets" (Shemmell *et al.*, 2007, p. 157) and that our model "showed that the imposition of linear synergy as a kinetic constraint significantly improved the ability of the neural network to evolve and reach the designated targets" (Shemmell *et al.*, 2007, p. 157). Also, Zaal encouraged us in personal (email) communication to extend the conceptual modelling work to include gravity into the model to gain intuition about its effect, as they had left out the gravitational component in their measure of joint torque. It is encouraging to see that idealised ER models are deemed relevant by empirical researchers working on motor control and the best way to hush critics of ALife modelling. Despite many open possibilities to extend the research on simulating motor synergies, the model here presented has not been taken further.

Within the enactive approach, no clear boundary between high-level and low-level processes can be drawn. Either way, problems of motor organisation and motor control are not in any obvious way related to our symbolic capacities, high-level cognition or human experience. As outlined in chapter 2 Sect. 2.4, embodied and dynamical approaches are sometimes criticised to be confined to such low-level behaviour. Motor control is an area where embodied thinking is nearly inevitable not a computationalist stronghold, a presumed 'representation hungry' problem. The following modelling and experimental chapters address questions that are, arguably, gaining grounds in areas that are at present underdeveloped in enactive cognitive science.

Chapter 5

An Exploration of Value Systems Architectures

The previous simulation model of motor synergies is a very applied model, whose results immediately relate to empirical science. This very tangible way of using ER simulation models gives an example of the potential of simulation models to generate proofs of concepts and to illustrate logical and mathematical states of affair beyond our cognitive grasp. By contrast, this chapter presents a simulation model whose results are of a more general and theoretical nature. It investigates the conceptual soundness of arguments proposing a certain type of control architecture for life-time adaptation. The architecture modelled is very wide-spread and features a 'value system' for self-supervised behavioural learning. The term 'value system' is borrowed from Edelman *et al.*'s work (e.g., Sporns and Edelman, 1993), but the idea is much more generally applied. The simulation model illustrates some of the implicit premises that underlie this kind of architecture and demonstrates that the adaptive capacity of such circuits can break down in closed-loop agent environment interaction, if no additional mechanisms to secure intact functioning are in place. The results from the model presented in this chapter have been partially published in (Di Paolo *et al.*, forthcoming; Rohde and Di Paolo, 2006).

The background Sect. 5.1 introduces value system architectures and reductionist approaches to value. The model and its results are presented in Sects. 5.2 and 5.3. The discussion Sect. 5.4 evaluates the results with respect to the framed question. Section 5.5 contrasts the analysed reductionist approach with ideas on autonomous sense-making and inherent valence in the enactive approach that we presented in (Di Paolo *et al.*, forthcoming; Barandiaran *et al.*, 2009), before Sect. 5.6 draws the overall conclusion and prepares for the following models by coming back to the methodological theme of the book, i.e., how computational methods (in particular ER simulations) can fill their niche in an enactive cognitive science.

5.1 Value Systems

5.1.1 *Reductionist Approaches and Value System Architectures*

In representationalist approaches, the symbol is separated from its meaning – the *signifiant* from the *signifié* – processes are not *inherently* meaningful, but are syntactic and become interpreted, as explained in chapter 2. The question of the origin of values thus has to be approached by looking for a process or entity *external* to the syntactic 'cognitive' process itself that provides meaning for the computational tokens. Many reductionist approaches refer to natural selection and survival of the fittest in Darwinian evolution as an inherently purposeful process that ensures that information processing is set up in a way that promotes genetic proliferation (e.g., Millikan, 1984): behaviour is meaningful only in so far as we can explain how it helped our ancestors to survive and reproduce in the African savannah.
This extreme reductionist perspective just sketched can be seen as one pole of a spectrum, in which a purpose precedes the living organism, a concept called *a priori semantics* here. This pole is in strong opposition to the enactive approach, in which evolution is an essential factor *shaping* the levels of mechanical processes that generate meaning but does not provide meaning itself. There are intermediary positions between these two poles that try to follow a third route, assuming that some, but not all meaning is determined evolutionarily.
A group in the intermediary range of this spectrum are the proponents of 'value system architectures'. The term is taken from Edelman *et al.*'s (e.g., Sporns and Edelman, 1993) Theory of Neuronal Group Selection (TNGS) but the kind of architecture discussed is much more widely used than this label. The term 'value system architectures' in this context denotes all those models that assume the existence of dedicated parts of the cognitive/neural architecture that have a representation of value and, therefore, can supervise learning internally, such that behavioural change is for the better.
An important feature of such architectures is that these systems are functionally and structurally isolated from the behaviour generating parts of the architecture. In many contexts, talk about value systems may be more metaphorical, i.e., even though there may be some functional differences in local structure, no strict separation is presumed.[1] However, research in AI and robotics has taken inspiration from such theories and, frequently, circuits with a strict separation of the value system from the behaviour generating systems has been implemented in self-supervised learning architectures (e.g., Sporns and Edelman, 1993; Verschure *et al.*, 1995; Snel and Hayes, 2008).

[1] At least in some context, this seems to be true for TNGS as well, which, in the first instance, is a neuroscientific theory.

Value systems are modules that generate a bipolar performance signal to evaluate sensorimotor behaviour, like an internally produced feedback signal for reinforcement learning. Sporns and Edelman define value systems as neural modules that are "already specified during embryogenesis as the result of evolutionary selection upon the phenotype" (Sporns and Edelman, 1993, p. 968). In a quasi evolutionary process of selecting the 'fittest' behaviour, such internally generated reinforcement signals direct life-time adaptation ('value-guided learning'): a value system for reaching, for instance, would become active if the hand comes close to the target.

This kind of architecture is very popular with sceptics of the traditional paradigm who argue for more embodiment and situatedness. For instance, Sporns and Edelman see this kind of an architecture as a solution towards problems of anatomical and biomechanical changes that are described as "challenging to traditional computational approaches" (Sporns and Edelman, 1993, p. 960). Pfeifer and Scheier, two pioneers of the situated and embodied approach in AI, argue that "if the agent is to be autonomous and situated, it has to have a means of 'judging' what is good for it and what is not. Such a means is provided by an agent's value system" (Pfeifer and Scheier, 1999, p. 315) and present (Verschure *et al.*, 1995)'s implementation of a TNGS architecture as the way forward in autonomous robotics.

However, there are problems with assuming an *a priori* separation of behaviour and value. Function is reduced to a local mechanism that represents an evaluation function and the design of this evaluation function is conveniently pushed off to evolution. The point this chapter aims to bring across is very similar to (Rutkowska, 1997)'s argument that "[increased] flexibility requires some more general purpose style of value" (Rutkowska, 1997, p. 292) than a value module could provide. She believes that value system architectures cannot explain adaptivity as a general phenomenon, even if value-guided learning circuits may work in specific cases. We summarise her view as follows:

> "She laments their vulnerability and their restrictive semantics consequent to the built-in evaluation criteria. A similar limitation is pointed out by Pfeifer and Scheier, who describe a 'trade-off between specificity and generality of value systems' (Pfeifer and Scheier, 1999, p. 473): A very specific value system will not lead to a high degree of flexibility in behaviour, while a very general value system will not constrain the behavioural possibilities of the agent sufficiently" (Rohde and Di Paolo, 2006).

Rutkowska goes as far as posing the question as whether a value system is a "vestigial ghost in the machine" (Rutkowska, 1997, p. 292).

Drawing a box and labelling it 'value system' seems reminiscent of first generation cognitive psychologist 'boxology'. It does not appear suitable to the post-cognitivist embodied and dynamical enactive approach outlined here. The kind of reasoning associated with value system architectures bears traces of homuncularity in assuming that what is good can be specified and represented as a function *pre factum*. As such, value system architectures suffer, in a miniature version, from those problems identified to result from the computationalist paradigm in chapter 2, Sect. 2.1: rigidity, semantic limitations, incapacity to deal with open-ended real-time change, *etc.*

Why would researchers sympathetic to embodied approaches use such a boxologist model of values? Decades of exercising a computationalist methodology persist in the language used to formulate questions and this makes it very difficult to fully let go of the baggage of implicit premises. It requires a constant attention to such issues to avoid postulating vestigial ghosts in the machine. Nobody disputes that norms exist across individuals of a species that result from natural selection. But there is a thin line between arguing that these norms are built in as parts of the mechanism, which is reductionist, and investigating the mechanism that gives rise to such norms that manifest in the relational and behavioural domain, which is not reductionist.

What does this mean for the architectures described? It means that the problems are not so much rooted in the circuits proposed but in calling parts of it a 'value system' and asserting that their meaning is built in by evolution. Where the paradigmatic confusion becomes important is when researchers take their labels of the circuits literally, when they confuse functional correlates with functional causes and propose that by placing value systems into a cognitive architecture, the problem of life-time adaptation is practically solved. The simulation model presented in this chapter illustrates just how serious such a confusion can really be. Such confusions about a complicated state of affairs can be very subtle, and many researchers concerned with concrete scientific problems in their daily routines, untrained in philosophy, may not be aware of there being a problem at all.

5.1.2 *Value System Simulation*

The simulation model described in this chapter illustrates the consequences taking a reductionist approach seriously. ER simulation models are particularly suited to investigate in an unbiased way the relation between function and mechanism, because this relation is not pre-specified but results from automated search (cf. chapter 3). (Yamauchi and Beer, 1994) address a similar problem in their evolution of learning in a fixed weight CTRNN

(this work on learning in fixed weight controllers was extended by, e.g., Tuci *et al.*, 2002; Izquierdo-Torres and Harvey, 2007). These studies show how associative learning behaviour is evolved in fixed weight controllers, refuting the intuition that fast time scale behavioural function and slow time scale modulation of this function have to be implemented by separate mechanisms (i.e., neural activity vs. synaptic plasticity). These studies show that phenomena that are distinct on a behavioural level need not be realised by separate dedicated functional mechanisms. The simulation results presented here provide a similar proof that such a functional and structural separation is not *a priori* necessary, or even beneficial. It thus aims to clarify the implications of taking a reductionist and computationalist-representationalist stance towards the problem of value or an enactive-embodied approach on the issue. By pointing out the differences between the two, the model aims at resolving the kind of paradigmatic confusion described above.

The results demonstrate how in value system architectures the proposed functional separation and localisation can lead to break-down of the adaptive principles, at least if no further mechanisms or constraints for ensuring stability are implemented. Taking the idea seriously that a local pre-defined structure generates meaning for an otherwise merely syntactic and value-agnostic architecture, it results that there is no way to make sure that the value system keeps working properly, that its input and output channels do not get re-interpreted in a variable sensorimotor context. A value signal that is actually symbolic in that it is arbitrary with respect to the meaning it bears could mean anything and the structures that obey it in performing adaptation have no way of telling what is wrong. We termed this gradual change in meaning through gradual change in sensorimotor context *semantic drift* (Rohde and Di Paolo, 2006; Di Paolo *et al.*, forthcoming).

The thrust behind the idea of pre-coded values is based on the presumption that there is a pre-specified and context-independent isomorphism between the function represented in the value-module and what is genuinely good or bad for the organism, and that value-guided learning modulates the structurally and functionally separate sensorimotor systems top-down. Proposals of value systems are based on evidence from neuroscience about certain cell assemblies (e.g., in the brain stem and the limbic system), whose neural activity modulates synaptic changes in the cortex. These assemblies have a tendency to become active when salient events in the environment are being observed. Such neural systems are postulated to implement a value system for certain circuits of value-guided learning (Edelman, 2003).

What can we conclude from such correlated activity? The first simulation model presented (Sect. 5.3.1) evolves agents to perform simple light-seeking behaviour (phototaxis) and to generate a signal bearing the characteristics of 'value systems', i.e., to correlate neural activity with behavioural saliency/success. In the case of the evolved agent this means fitness. This signal is not yet embedded into the architecture, it is just a value output signal (see Fig. 5.1 (B)). This model aims to investigate what we really can infer about functional localisation from correlated neural activity.

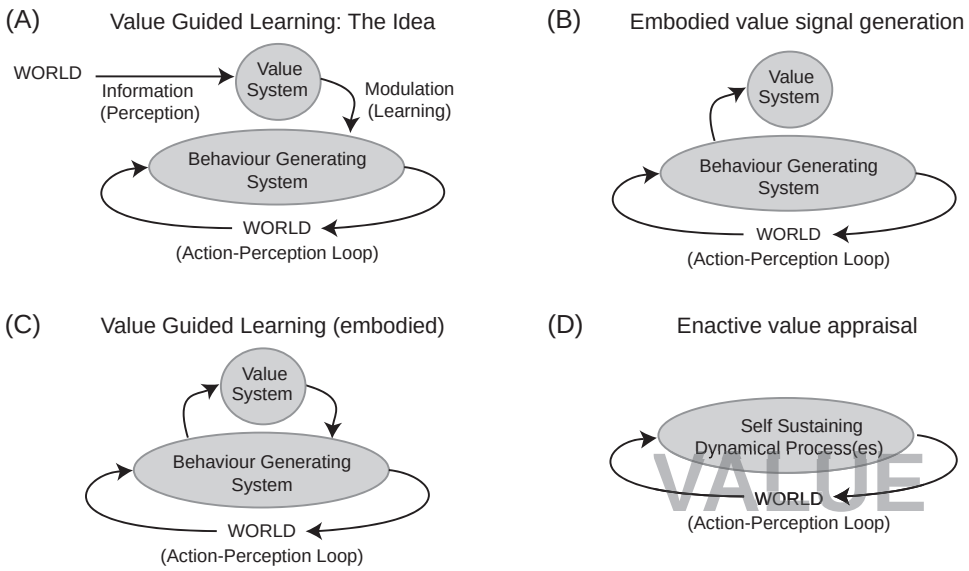

Fig. 5.1 An illustration of different views on values. (A): value-system architecture. (B): embodied value system (first simulation model). (C): value-guided learning with an embodied value system (second simulation model). This simulation shows how, if an embodied value system (as in (B)) is introduced into value-guided learning (as in (A)), semantic drift corrupts the adaptive circuitry. (D): value emergence in the enactive view -values are not localised in a neural module, they emerge from a self-sustaining material process of identity generation.

In value system architectures, value systems provide feedback for internally supervised life-time learning (see Fig. 5.1 (A)). In the second part of the simulation study (Sect. 5.3.2), the value system evolved in the first part of the simulation study is used to implement a value module in a kind of life-time learning through 'neural Darwinism' (see Fig. 5.1 (C)). In this particular simulation, artificial evolution is seen as a metaphor of ontogenetic change, not of phylogenetic evolution. This model investigates what happens to the proposed circuits of value-guided learning if embedded in closed loop interaction. The results show that the behaviour quickly gets worse as a consequence of *semantic drift*.

It is important to stress that this way of using a GA and ER simulation as an analogy for neural Darwinism is inspired by the neural Darwinism proposed by Edelman *et al.*, but differs substantially in its implementation. TNGS proposes Darwinian-style evolution as principle of neural organisation (Edelman, 1989, p. 242), where the output of value systems serves as criterion for selection of neural assemblies. Synaptic connections participating in the constitution of 'good' behaviour are strengthened. This process is akin to natural selection. However, TNGS puts much more emphasis on selection of the fittest from a large but invariant repertoire of neural populations, not on replicating the Darwinian principles of heredity and mutation. Even though the circuits proposed as part of TNGS fall into the much larger class of 'value system architectures' under study, they are not the most typical example.

The model investigates the question of the possible functional role of 'value systems' in a deliberately minimal toy-like set-up. It does not aim to model actual brain structures. It just serves to illustrate a conceptual argument of what correlated activity can mean *in principle* and what follows from the core assumptions underlying value system architectures *in principle*, if no additional assumptions are made.

5.2 Model

A circular two-wheeled agent of four units diameter is evolved to seek the light (phototaxis) and, at the same time, to generate a motor signal that correlates with its behavioural success (in analogy to a value system). Behavioural success is measured as relative distance from the light source.

In the second experiment, the internally generated value signal evolved during the first experiment is used as reinforcement signal for continued evolution of behaviour. This continued evolution is an analogy of value-guided neural learning.

The agent is controlled by a CTRNN (see Eq. (3.2)) whose structure (i.e., the connectivity C and the number of hidden neurons) is partially evolved. Connections to input neurons or from output neurons are not permitted. Input neurons can project to output neurons and to hidden neurons, hidden neurons can project to other hidden neurons and to output neurons. The network has two input neurons and five output neurons (specification below) and can have varying numbers (0-5) of hidden neurons. The existence or non-existence of hidden neurons and neural connections is determined if the corresponding values x in the artificial genome are $x > 0.7$ and $x > 0.6$ respectively (i.e. gene interpretation using step functions).

Evolution is implemented with the GA presented in Sect. 3.3 and vector mutation with $r = 0.7$. Most evolutionary runs lasted for 2000 generations. Parameter ranges are $w_{ij} \in [-8,8]$, $\theta_i \in [-3,3]$ and $\tau_i \in [16,516]$.

The agent has two light sensors $S_{L,R}$ with an angle of acceptance of 180°, which are oriented $+60°$ and $-60°$ from the direction in which the agent heads. The sensor orientation is subject to uniform directional noise $\varepsilon_d \in [-2.5°, 2.5°]$. Their activation is fed into input neurons by $I_{L,R}(t) = S_G \cdot S_{L,R}(t)$ with the evolved $S_G \in [0.1,50]$ and $S_{L,R}(t) = 1$, if the light is within the sensory range at time t and $S_{L,R}(t) = 0$ otherwise. The *binary activation of light sensors makes the fitness estimation non-trivial*, as there is no direct signal present in the sensory inputs that represent distance from the light source (e.g., light intensity). In order to generate a motor signal that corresponds to behavioural success, an active perceptual strategy has to be evolved.

The motor velocities are set instantaneously at any time t by $v_{L,R}(t) = M_G(\sigma(a_{L1,R1}(t)) - \sigma(a_{L2,R2}(t)) + \varepsilon$ where M_G is evolved $\in [0.1,50]$ and $a_{L1,L2,R1,R2}$ is the activity of the four motor neurons generating the velocity. $\varepsilon \in [0,0.2]$ is uniform motor noise. A fifth output neuron n_{M5} generates the performance estimate $E(t) = \sigma(a_{M5}(t))$ which, during the first experiment, is evolved to represent the present distance to the light source relative to the starting distance to the light source (fitness function Eq. (5.3)).

In every evaluation, the agent is presented with a sequence of 4-6 light sources that are placed at a random angle and distance $d \in [40,120]$ from the agent. Evaluation trials last $T \in [3000,4000]$ time steps. They are preceded by $T' \in [20,120]$ simulation time steps without light or fitness evaluation, to prevent the initial building up of activity in the estimator neuron from following a standardised performance curve. Each light is presented for a random time period $t_i \in [\frac{T}{5} - 100, \frac{T}{5} + 500]$ time steps. The network and the environment are simulated with $h = 1$.

The fitness $F(i)$ of an individual i is given by

$$F(i) = F_D(i) \cdot F_E(i) + \varepsilon F_D(i) \tag{5.1}$$

where $F_D(i)$ rates the phototactic behaviour and $F_E(i)$ rates the fitness prediction. The co-evolution of light seeking and estimation of performance using the product of both terms is difficult for evolutionary search to generate from scratch (noisy). The product of these two terms, rather than a weighted sum, was chosen because of local maxima in the fitness landscape. It was too easy to trade off these two criteria, e.g., to just evolve light seeking and a 'good enough' fitness estimation curve (monotonically increasing, but not sensitive to ongoing behaviour). The product forces the GA to come up with strategies that solve both

problems. The second term ($\varepsilon = 0.001$) is included to bootstrap the evolutionary process by minimally rewarding light-seeking behaviour over no sensible behaviour.
$F_D(i)$ is given by

$$F_D(i) = \frac{1-P^2}{T} \sum_0^T max\left(0, 1 - \frac{d(t)}{d(t_0)}\right) \tag{5.2}$$

where $d(t)$ is the distance between robot and light at time t and t_0 is the time of the last displacement of the light source, i.e., the reduction of distance is integrated over the trial. $P = \frac{0.125}{T} \sum_0^T \frac{v_L(t)-v_R(t)}{M_G}$ integrates the difference in velocity between the wheels and thus discourages turning.

As stated above, it was technically difficult to evolve satisfactory online estimation of performance. Online measures of performance are anti-proportional to the relative distance to the goal at any point (in analogy to Eq. (5.2)). There are trivial solutions to the problem, such as constantly outputting the fitness average or slowly building up activity in the estimator neuron, that correspond rather well to the gradual decrease of distance that characterises successful light seeking. To force a more sophisticated strategy of online performance estimation, terms rating both the absolute fitness value and its change were combined. Also, performance estimation was only rewarded if it predicted performance better than the average across a trial. A long process of trial and error led to the following mathematically somewhat complicated equation for $F_E(i)$:

$$F_E(i) = \sqrt{\max\left(0, \frac{e(\bar{d},d) - e(E,d)}{e(\bar{d},d)}\right) \cdot \max\left(0, \frac{e(0,\dot{d}) - e(\dot{E},\dot{d})}{e(0,\dot{d})}\right)} \tag{5.3}$$

with $e(x,y)$ the sum of squared error $e(x,y) = \sum_0^T (x(t) - y(t))^2$. $\bar{d}$ is the average of $d(t)$ during each trial. $\dot{d}(t)$ and $\dot{E}(t)$ are the derivatives of $d(t)$ and $E(t)$ averaged over a sliding time window $w = 250$ time steps (interval borders for the $e(x,y)$ have to be adjusted accordingly).

Fitness evaluation is exponential across $n = 6$ trials as defined in Sect. 3.3, Eq. (3.5).

In value-guided learning, internal neural modules whose activity correlates to behavioural success provide feedback for online learning processes. Such a value system was evolved in the first simulation study. To implement this idea of value-guided learning, the fitness function F_i in Eq. (5.1) is substituted for the value signal (distance estimates) E in the second simulation, such that

$$F'(i) = \sum_0^T E(t) \tag{5.4}$$

5.3 Results

5.3.1 *Co-evolution of Light-Seeking and Fitness Estimation*

This presentation of co-evolved light-seeking and fitness estimation behaviour is not concerned with evolvability or a variety of strategies evolved, as for most other ER models presented in this book. Instead, it focuses on the thorough analysis of one example agent as a paradigm case. The control network evolved for this agent is very effective yet simple and illustrates well the theoretical argument.

The network controller evolved to control the two-wheeled simulated agent is extremely simple, but astonishingly good at estimating how close the agent is to a light source, despite the minimal sensory endowment (two light sensors generating on-off signals) and the consequent ambiguity in the sensory space (i.e., any sensory pattern could occur at any distance from the light source). Even though there was the possibility for the GA to exploit nonlinear dynamics and network states as memory, the evolved controller has no hidden neurons, recurrent connections or slow time constants. Therefore, its behaviour hardly relies on internal state and its complexity is minimal, even within the already restricted range of possibilities.

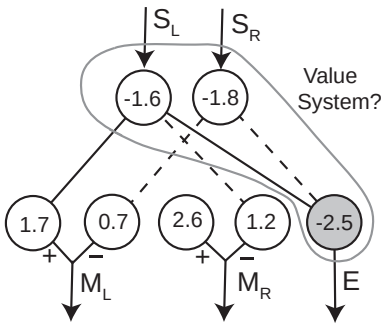

Fig. 5.2 The controller of the agent that seeks light and estimates its distance from the light. (θ in neurons, dotted lines interneural inhibition, solid lines interneural excitation.) The grey line demarcates the sub-system responsible for generating the value signal.

As a consequence of the absence of recurrent connections and hidden neurons, the neural sub-structure that generates the value signal is structurally isolated from the rest of the network dynamics (apart from being fed by the same input neurons), just like a value system in value system architectures for learning (see encircled group of three neurons in 5.2). This strict modularisation of neural substrates for evaluation and for behaviour generation

had not been built-in, but the fact that it resulted from the evolutionary process makes the analogy with value system architectures even stronger.

To understand the evaluation function the value module implements, first, an open-loop analysis was conducted. In the absence of light, or if the network receives input only on its right light sensor ($S_R = 1, S_L = 0$), it estimates $E \approx 0$. If light is perceived with both sensors, it estimates $E \approx 0.5$, and if the network receives input only in its left light sensor ($S_R = 0, S_L = 1$), the estimate reaches its maximum of $E \approx 0.8$. The judgement criteria of this value system can thus be described as 'seeing on the left eye is good, seeing on the right eye or not at all is bad'. Taken by themselves, these rules do not make sense.

Nevertheless, Fig. 5.3 (B) (bottom two plots) shows that both $E(t)$ and $\dot{E}(t)$ (dotted lines) follow with surprising accuracy the actual values $d(t)$ and $\dot{d}(t)$ (solid lines), particularly if we remember the poor sensory endowment of the agent.

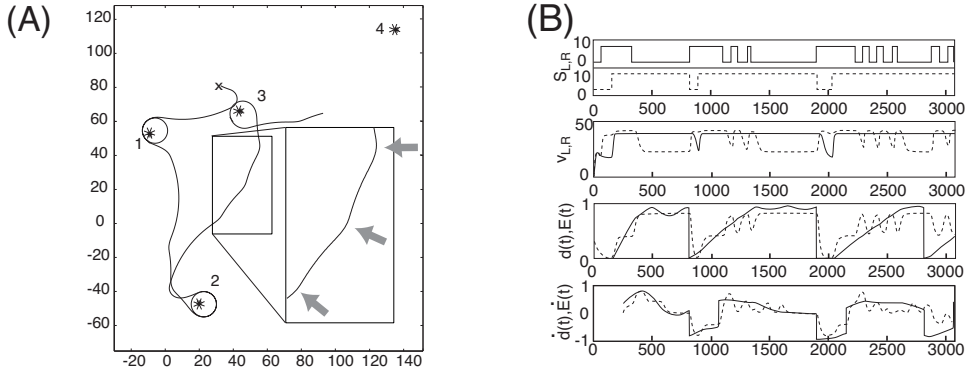

Fig. 5.3 (A) Successful light seeking trajectory for four presentations of light sources. Arrows indicate the punctuated turns during $t = 2200 - 2700$ (see text). (B) The evolution of different variables over time in the same trial (Top to bottom: $S_{L,R}$, $v_{L,R}$, $d(t)$ vs. $E(t)$, $\dot{d}(t)$ vs. $\dot{E}(t)$).

In order to explain how this accuracy in estimating the performance is achieved, it is necessary to take into consideration the agent's light seeking strategy (Fig. 5.3 (A) and (B)). The agent's phototactic behaviour is realised by the network minus the estimator neuron. In the absence of sensory stimulation, the agent slowly drives forward, slightly turning to the right. Thereby, it draws a circle that will eventually make the light source appear in its visual field, entering from the right. If $S_R = 1$ and $S_L = 0$, the 'brake' on the left motor M_L is released and induces a sharper turn to the right. This means that the light eventually crosses into the centre of the visual field of the agent, i.e., $S_R = 1$ and $S_L = 1$, which triggers the agent to release the 'brakes' on both wheels and drive almost straight, only slightly

drifting to the right. This right drift in the near straight approach behaviour means that the light source repeatedly disappears from the right sensor's angle of acceptance ($S_R = 0$ and $S_L = 1$), which induces a sharp turn to the left that brings the light source back into the range of the right light sensor ($S_R = 1$ and $S_L = 1$). Once the light source is reached, this sharp turning to the left results in circling anti-clockwise around the light source, as this ongoing sharp turning to the left does not bring the light source back into the sensory range of S_R. In combination, these phases lead to the following sequence of behaviour during the approach of a single light source:

(1) A scanning turn to the right, until $S_L = S_R = 1$.
(2) A quick approach of the light from the right side, bringing the light source in and out the sensory range of S_R (cf. the rhythmically occurring drops of sensory and motor activity in Fig. 5.3 (B)). This strategy results in the chaining of nearly straight path segments in the approach trajectory, separated by punctual left turns (arrows in Fig. 5.3 (A)).
(3) Counter-clockwise rotation around the light source during which the light source is perceived with the left sensor only.

Knowing about this light seeking strategy, it is much easier to understand how the 'value system' achieves a correct estimation of the distance: the approach behaviour only starts when the light is in range of the left light sensor, and this sensor remains activated from then on, which explains the positive response to left sensor activation $S_L = 1$. $S_L = 0$, on the other hand, implies that the light has not yet been located, which only happens in the beginning of the trials if the agent is far away from the light source, hence $E \approx 0$. The right light sensor is activated during the approach trajectory, but not once the light source is reached. Therefore, it mildly inhibits n_{M5} which results in $E \approx 0.5$ when $S_L = S_R = 1$. An additional level of accuracy during approach behaviour is achieved by keeping the light source at the boundary of the right sensor's sensory range by approaching the light at an angle from the right: the closer the agent is to the light source, the larger is the angular correction necessary to bring the light source back into its sensory range and, therefore, the longer the intermittence in right sensor stimulation (see Fig. 5.3 (A), little arrows, and (B), oscillations in sensory input and estimation). This implies that, on average, the fitness estimate is higher the closer the agent is to the light source, because the right sensor, which mildly inhibits the performance estimate, is switched off for longer intervals. When the agent has reached the light source and cycles around it, $S_R = 0$ and $S_L = 1$, and the value system produces its maximum estimate $E \approx 0.8$, expressing that the light source has been

reached. The system thus has constructed a relative distance sensor from the two light sensors that were given.

This phase of the simulation had mainly been intended to provide the basis for the second part of the simulation, i.e., an agent that generates a certain behaviour and a signal that represents behavioural success (value signal). However, it demonstrates an important theoretical point in itself: a value signal that correlates to behavioural success, even if it is generated by a neural structure that is modularised and not linked to the systems that generate motor behaviour, is not necessarily disembodied and explicable outside the sensorimotor context. This is interesting with respect to the question of neural correlates of behaviour: a neural assembly that generates a signal that correlates with behavioural success is not necessarily solely responsible for generating this signal, even if the neural structure is fully separated from the structures that generate sensorimotor behaviour. The external closure of the sensorimotor loop can contribute a link that is missing in neural connectivity.

Another event worth discussing in the trial depicted in Fig. 5.3 (A) and (B) occurs after the last displacement of the light source ($t > 2800$): as the displacement happens to bring the light source into the left visual field of the agent, it immediately enters the oscillating approach mode and its estimate therefore poorly corresponds to the actual distance measure which drops to 0. This dissonance can be seen as a possibly inevitable error due to the minimalism of the sensory equipment of the agent. However, putting oneself 'in the agent's shoes', it could also be interpreted as the superiority of the evolved estimator over the distance measure as a measure of performance: the comparably high output expresses the agent's justified optimism to be at the light source soon, which is not reflected in the distance fitness measure $F_D(i)$, which evaluates distance independent from the orientation of the agent and what it implies for behavioural success.

5.3.2 *A Caricature of 'Value-Guided Learning'*

In explaining the mechanisms of life-time learning, the proponents of TNGS (e.g., Edelman, 1989; Sporns and Edelman, 1993; Edelman, 1987) mention the following key components

(1) A neural assembly whose activation correlates to saliency of events (value system).
(2) Neural selection based on Darwinian principles that is guided by the activity in the value system.
(3) The possibility for value system learning supervised by higher order value systems.

In this sense, TNGS is a typical example of a self-supervised learning or value system architecture, by explicitly separating the value system, the selection process and the neural assemblies that generate behaviour. Many real robotic models using this kind of architecture implement only (1) and (2) (e.g., Sporns and Edelman, 1993; Verschure *et al.*, 1995). The argument here is that the third point, i.e., a mechanism that ensures that the value system works properly, is really the most important and the most difficult part of this kind of adaptive circuit – the only really adaptive part. Therefore, showing that (1) and (2) work, given that the experimenter takes care of (3) by providing a magic meaning sensor, does not explain or show very much about adaptive behaviour. The mechanisms underlying (3) are the most vague, and implementations of value system architectures do not even attempt accounting for the principles that make value systems work. Section 5.4 will come back to this issue.

The model simulates the logical consequences of the described circuits if implemented without list item 3 in place. In this simulation, the evolution of the robot controller is seen as the analogue of ontogenetic learning of sensorimotor circuitry, guided by activity of the value system. The GA is seeded with a population of the successful individual discussed in the previous Sect. 5.3.1. The only parameters that evolve in this experiment are the strengths of the three synaptic connections from sensors to motors (behaviour generating sub-system; cf. Fig. 5.2). The fitness measure F is substituted for the performance estimate $E(t)$ (Eq. (5.4)). It is important to notice that, in this set-up, the value system does not evolve, it just guides the evolutionary change of the synaptic weights to reinforce whatever behaviour leads to a high performance estimate $E(t)$. This top-down modulation by a localised value system is at the core of what has been described as 'value system architectures'.

Figure 5.4 (A) illustrates how this 'value-guided learning' results in a complete deterioration of phototactic behaviour within 50 generations (Fig. 5.4 (B)). Behaviour is altered to driving around the light source in large anti-clockwise circles, not approaching at all, which results in a deterioration in both components $F_D(i)$ and $F_E(i)$ of the fitness function, even though the value judgement, which used to be correlated with behavioural success, is maximally positive.

This deterioration is a consequence of closing the sensorimotor loop on both, the value-judgement, and the learning. As analysed previously, the value judgement relies on an active perceptual strategy. In a variable sensorimotor context, what the 'value system' rewards is simply activation of the left light sensor but not the right (cf. Sect. 5.3.1). The

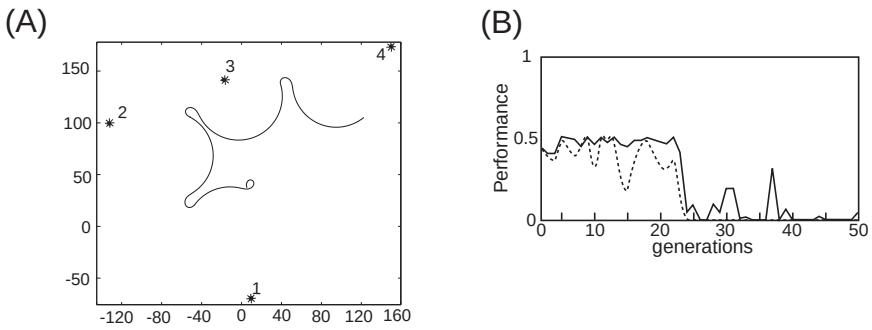

Fig. 5.4 (A) Light-avoiding trajectory of an agent after 50 generations of 'value-guided learning'. (B) The degeneration of light seeking performance F_D (solid line) and estimation performance F_E (dotted line) over generations (learning) for the same experiment.

5.4 Discussion

Value system architectures, as many related architectures proposed, presume an informationally encapsulated rigid structure to provide a meaningful signal for an otherwise meaningless process. Findings about brain areas whose activity correlates with salient events in the environment are interpreted as evidence for the existence of such value systems in the nervous system. The present simulation models show that this reasoning is not stringent: in the first experiment, it is shown that even a modularised brain area that is not directly connected to the behaviour generating neural subsystems can depend on sensorimotor dynamics through indirect linking via the agent-environment interaction. It 'measures' or 'computes' value using an active perceptual strategy. In the second simulation, it is shown how, as a consequence of this embodied strategy, a gradual change of the behavioural context induces a gradual change in judgement capacity. The behavioural plasticity that the value system itself supervises corrupts the value system's judgement, which, in turn, leads to a divergence from the desired sensorimotor behaviour and an even more pronounced change in value judgement. This phenomenon, which is a direct consequence of the existence of reciprocal causal links between value system and behaviour generating systems, is what is referred to as 'semantic drift'.

The functional integration of embodied behaviour in value judgement undermines the very concept of a value system as a top-down modulator. We cannot expect such a circuit to work immaterial of plastic changes in the environment and the brain, even if, locally, we can describe neural activity as a correlate of meaningful events. In this sense, the simulation can be seen as an illustration of the problems associated with 'hybrid' architectures that

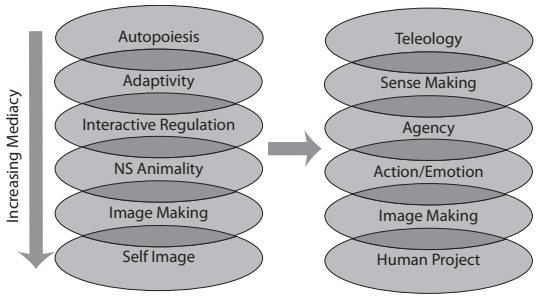

Fig. 5.5 Life-cognition continuity and the scale of increasing mediacy.

feature a central symbolic 'cognitive' control circuitry and peripheral systems that work according to more embodied principles: "If a full-blown ghost in the machine has difficulties dealing with the variability of the external world, why would a vestigial ghost in the machine not face the same difficulties dealing with the variability of its bodily environment?" (Rohde and Di Paolo, 2006). A local neural circuit implementing a mapping cannot be functionally evaluated in the open loop outside the behavioural context, because a complex nonlinear dynamical system cannot be expected to act (approximately) like a linear system that interfaces this system with the world.

The point is not to deny that value system architectures can work *if* there are additional mechanisms insuring that everything goes alright. As mentioned earlier, there is evidence about correlation between brain activity in certain neural modules and salient events in the environment (even if the existence of this correlation does not *a priori* explain anything about its function), and, if this signal is reliable, there is no reason to doubt that it could modulate behaviour. It has to be asked, however, if explaining the mechanisms to maintain the generation of a meaningful value signal is not ultimately the lion share of the explanatory work, which is conveniently pushed off.

(Sporns and Edelman, 1993) conjecture that "different value systems interact, or that hierarchies of specificity might exist" (Sporns and Edelman, 1993, p. 969). The proposal here seems to be that the maintenance and adaptation of value systems should also follow the principles of value-guided neural Darwinism. In the cited paper, this recursive application of value-guided learning circuits is not explicitly modelled. The intuition is that such a meta-value-guided learning leads to a *regressus ad infinitum* or, otherwise, require a magic (homuncular?) master-value system to end this regress.

As stated earlier, this criticism is not a criticism of TNGS in particular, but of self-supervision circuits with a dedicated 'value-system' in general. Indeed, recent work by Edelman *et al.* (e.g., Krichmar and Edelman, 2002), as well as by other groups (e.g., Doya, 2002), appears to break with the idea of neural Darwinism as fundamental principle of ontogenetic adaptation. They extend the proposed framework to include other kinds of neural plasticity and meta-modulation, proposing different kinds of adaptive circuits for different kinds of modulatory sub-systems. These models are informed by recent neuroscientific evidence and are conceptually much more complex than the simple neural Darwinism modelled in this chapter. These extensions appear to confirm Rutkowska's assumption that "[increased] flexibility requires some more general purpose style of value" (Rutkowska, 1997, p. 292) than a value module could provide.

The criticism here is a logical criticism. Such existence proofs in simulation, even though they teach us to be careful not to presuppose a functional modularity, do not exclude the empirical possibility of such structures. Maybe there are "simple criteria of saliency and adaptiveness" (Sporns and Edelman, 1993, p. 969) that can *a priori* specify what will be good and what will be bad *a posteriori* – but this will have to be proven empirically. Maybe, value system functionality can be kept intact by mechanisms of value system learning – but it has to be shown and argued how that would happen rather than to just postulate such mechanisms. Maybe, in some instances, semantic drift can even be a problem for biological instantiations of value-system architectures, not just for computational models. In a far-fetched comparison, you could think of our pleasure and pain systems as value systems, and of some forms of substance abuse as value-guided learning that is led astray by semantic drift. Usually, our pleasure systems reward behaviour that benefits our continued existence and well-being, but if you consume euphoriant drugs, these circuits may end up reinforcing behaviour that is actually harmful or even lethal. But this kind of mal-adaptive circuitry appears to be rather the exception than the rule. There is no doubt that the identified neural structures, whose activity correlates to salient events in the environment play a fundamental role in value-appraisal and adaptation – but reducing value generation to these structures seems a category mistake, a confusion of mechanism and behaviour, a reduction that cannot be justified on the basis of correlation alone.

To cut a long story short, the point of this model is not to discourage the scientific study of the described neural structures or to discourage the use of value system models if the conceptual limitations are made explicit – the point is about not confusing correlation and causation, when measuring neural activity that correlates to salient events. By postulat-

ing pre-specified value systems without explaining how they work, the explanatory burden "[b]uck [is passed] to evolution" (Rutkowska, 1997, p. 292) and the real question of why something matters to the organism, in the sense outlined above for the enactive approach, is not addressed. Surely, there are possibilities to make explicitly reductionist circuits (i.e., those that do not foresee an active maintenance of value system function) work if the reciprocal causal links on the value system are cut. There are robotic artifacts with a limited behavioural domain (Verschure *et al.*, 1995) that successfully implement the adaptive circuits proposed as part of TNGS. In these models, the value system has 'magical sensors' or privileged access to variables in the environment, and these are not affected by sensorimotor learning. But in order to be convincing as a biological theory of general adaptivity, it would be necessary to specify how such rigidly wired value systems would be realised in a living organism that is in constant material flux.

5.5 Enactive Sense Making, Value Generation, Meaning Construction

This chapter has focused so far on pointing out the problems associated with localist approaches to meaning and value. The question that remains is: how could it be any other way? In chapter 2 and throughout this chapter, there have been references to the idea of 'inherent semantics' of adaptive processes, an idea that is illustrated in Fig. 5.1 (D). This section will summarise arguments that intrinsic value can be generated by an autonomous organisation that preserves an identity. It recalls examples from autopoietic theory and the related literature, discussing them briefly rather than giving them an in depth treatment. The idea of behaviour as sense-making is very different from reductionist views that modularise and automatise functions to local syntactical processes and the model presented previously in this chapter illustrates how the two views are in tension.

(Weber and Varela, 2002) have been the first to explicitly identify intrinsic teleology, natural purposes and the possibility to imbue interactions with the environment with meaning as fundamental properties of living creatures. They combine ideas from Kant's 'Critique of Judgement' and (Jonas, 1966)'s biophilosophy in order to argue that autopoietic organisation, i.e., self-production, self-maintenance and self-repair characteristic of living organisms, not only implies basic autonomy and identity generation (This had been previously argued, e.g., Maturana and Varela, 1980). Weber and Varela argue that autopoiesis also implies genuine purposefulness of existence and of interactions with the environment. As a consequence of the organism being alive, a profane material process obeying the laws of physics, like two celestial bodies colliding or water streaming down a mountain, becomes

meaningful and can be positive, negative or ambivalent to the organism, depending on its impact on autopoietic organisation. This captivating idea of genuine intrinsic purpose of living organisms is very central to the enactive approach as it is presented here.

From recognising inherent purpose in a physical entity, however, it is not a trivial step to deduce the value of its interactions with the environment: whilst for a bacterium that follows a sugar gradient it is quite easy to judge, based on the organisation of the bacterium, that this is a good behaviour, it is much more difficult in more complex organisms: how can we explain a smoker who likes to smoke, a lemming jumping off the cliff, dolphins playing? There are clearly goals we pursue that cannot directly, if at all, be linked to our continued metabolic existence.

We propose to define value as "*the extent to which a situation affects the viability of a self-sustaining and precarious process that generates an identity*" (Di Paolo *et al.*, forthcoming). Autopoiesis, i.e., the continued self-construction of a metabolising network of processes sustaining itself in a far-from-equilibrium situation (which characterises life) is the most prominent example of such a process. But it is not the only one. More complex forms of organisation give way for multiple levels of such identity generation and, consequently, to different values which may not relate to metabolism or even generate a conflict in opposing the basic metabolic needs of the organism. (Varela, 1991, 1997) explored the idea of the organism as a 'meshwork of selfless selves' and 'patterns of life', identifying how, in phylogenetically more developed organisms, new levels of autonomous dynamics can emerge on top and alongside cellular autopoiesis. His own scientific work focused on three such levels of autonomous dynamics: autopoiesis (cellular identity), the immune system (multicellular identity) and the nervous system (neuro-cognitive identity). Varela identifies other levels of possibly identity generating processes, reaching from pre-cellular identity (self-replicating molecules) to socio-linguistic and superorganismic identity.

The important thing to notice is that these levels of autonomous dynamics (single-cellular, multi-cellular, neural, society, ecosystems...) can co-exist and come into conflict or synergy in any one individual organism's behaviour. The metaphor of the living organisation as the model physical system for cognition does not imply that everything we do has to be about survival, or has to be explained with reference to survival. Habits are strong, and so are social norms, or hormonally induced desires, and the dynamics with which they emerge, proliferate and lead to frustration or satisfaction can take life-like appearance. (Barandiaran, 2007)'s notion of 'Mental Life' makes the explicit analogy between chemical metabolical

life and mental neuro-behavioural life, where patterns of self-sustaining dynamics in the brain are autonomous in just the same sense as metabolic networks in cellular life.

The 'scale of mediacy' in Fig. 5.5 (due to Di Paolo; see also Di Paolo *et al.*, forthcoming; Barandiaran *et al.*, 2009) has been repeatedly presented in our work. It features a listing of hierarchical levels of value generation that are particularly interesting. In this scale, forms of organisation are mapped to behavioural-cognitive capacities. The scale has been inspired by (Varela, 1997, 1991) and by (Jonas, 1966), who develop similar listings of levels of organismic organisational and cognitive complexity. The underlying idea is that, with increasingly complex forms of organisation, the semantic distance between a need and the sign of its satisfaction or frustration becomes larger and more mediated: sugar has a more immediate link to metabolism (autopoiesis) than the perception of a prey's footsteps in the snow (you cannot metabolise a footstep). The sense-making activity in using the footstep as a sign of food is, therefore, more mediated. However, this does not mean that a footstep is an arbitrary symbol, in a computational or de Saussurian sense. Its meaning is still a direct result from the processes involved, not an externally specified convention. Increasingly complex levels of organismic organisation allow increasingly mediated forms of sense making, which imply more liberation of sense-making activity from immediate physical constraints – without ever separating the processes of behaviour generation from the processes of meaning generation.

Going through the list from the beginning, the first important distinction concerns the first three stages. These are not usually identified as distinct. The distinction between the first two levels is based on (Di Paolo, 2005)'s distinction between mere autopoiesis and adaptive autopoiesis, in which the recognition of environmental tendencies and according reactions form the basis for generating value and meaning that goes beyond just life and death; if I adaptively regulate, this produces the possibility of improvement, a continuity of value. A just autopoietic entity just does what it does, is robust to perturbations to a certain extent, but if something happens that means it dies. It never aims to improve the conditions for its continued existence. The distinction of the third level, i.e., of interactive regulation and agency as an elaboration of the adaptive autopoiesis (cf. Barandiaran *et al.*, 2009), is based on (Moreno and Etxeberria, 2005)'s observation that regulation only cannot be justly called agency. In order to call a living organism an agent, they argue, it has to also adaptively act on the environment. Adaptive regulators adjust their internal state in order to improve the conditions for continued existence, not the external. "An example of a just-adaptive organism is the sulphur bacterium that survives anaerobically in marine sediments

whereas bacteria swimming up a sugar gradient would, by virtue of their motion, qualify for minimal agency" (Di Paolo *et al.*, forthcoming).

The further stages that are included in the scale (Fig. 5.5) have been adopted from (Jonas, 1966)'s work. He identifies the fast motility of animals as the basis of emotions. Only animals with their fast motility can assign meaning to something at a distance and thus fear or desire the remote. They act spatially, whereas simpler organisms without fast motility and long distance perception act always on the basis of the immediate environmental and sensory surface properties, even though this may involve geometrically embedded behaviour from the observer perspective, as developed in (Barandiaran *et al.*, 2009). The last two stages are reserved to humans, who, through their general image-making capacity, and particularly their self-image-making capacity, gain the ability to regard situations objectively and define themselves as subjects. These later processes of value generation surely do not only reside in the individual and its interaction with an environment of objects but rely heavily on processes of socio-linguistic and cultural self-organisation.

In this listing, the consequence of a sign for the precarious process that generates the value/identity and the sign itself become increasingly mediated and physically detached. The consequence of increased mediacy is the liberation of ways to generate values: "For instance, only a sense-making organism is capable of deception by virtue of the mediacy of urge and satisfaction. A bacterium that swims up the 'saccharine' gradient, as it would in a sugar gradient, can be properly said to have assigned significance to a sign that is not immediately related to its metabolism, even though it is still bound to generate meanings solely based on the consequences for its metabolism" (Di Paolo *et al.*, forthcoming). This error can cost the bacterium its life. The higher the degree of mediacy, the more complex it is for the observer to interpret a sign with respect to the process(es) of identity generation from which its value emerges.

What does this mean for explanations of open ended adaptivity? The enactive study of value involves the study of generative mechanisms, as we argued for the case of autonomy (Rohde and Stewart, 2008). New forms of organismic organisations can enable new and more complex kinds of value-generating processes, and, naturally, these will be more complex for the more evolutionarily advanced species and their behaviour. No new proposal for general purpose adaptive circuitry will be suggested as an alternative to value system architectures here. The lesson to be drawn from this philosophical interlude is that the easiest and most natural way of making processes meaningful is not through a dedicated and homuncular 'meaning module' or 'value system', but through intrinsic valence of adaptive

processes, and, in order to understand this valence, biological organisation and ecological context will have to be taken into consideration, an opposition that Fig. 5.1 (D) tries to capture.

5.6 Conclusion

To conclude with a direct opposition of reductionist and enactive approaches, the former will always be *functionally limited* in their adaptive capacities, whereas the adaptation capabilities of living organisms are functionally open-ended. This is not to say that a living organisms could do anything, living organisms are limited as well. The difference is that their limitations tend to be material and physical (i.e., laws of nature), not functional (i.e., erroneous 'grounding of manipulated symbols' because the designer had not foreseen a specific situation). Organisms can adapt to situations that have never been there, because the processes that regulate their behaviour are of intrinsic valence. As we argue in (Di Paolo *et al.*, forthcoming), the most striking examples of value changes, which can shatter the functionality of established relations, are illness and other perturbations to the body (distortion or impairment). "[C]onsider a patient who, during the course of a disease, is subjected to increasing dosages of a pharmaceutical agent, with the result that he not only survives dosages of the drug that would be fatal to the average human being, but also that his metabolism relies on the medicine in a way that deprivation would cause his death" (Di Paolo *et al.*, forthcoming). The valence of the medicine here is not represented externally, as a symbol, which has to be updated by a syntactic process monitoring and parsing information. You cannot just turn a poison into a nutrient by updating a local value-function. The change in significance results from the dynamical re-organisation of the organism itself.

There are a lot of open research questions concerning the origins of value and the structures that realise life-time adaptation. How can these questions be addressed from an ER modelling perspective without stepping into a reductionist trap? One avenue would be to evolve learning behaviour, in an unbiased way, and look for value system-like structures in the evolved agent. Similarly, an approach like the one presented here could be taken (i.e., to 'force' the evolution of value systems, as in the first simulation) that incorporates the value signal into the evolution of agents for learning and investigate its functional role. The difference to the first proposal is that, in the latter case, the genetic algorithm has a value system like structure at its disposal, as a building block, and therefore would be more likely to generate a control circuitry that relies on this structure in its function. This ap-

proach serves to investigate possible functional roles of neural structures whose activity correlates with behavioural success in an unbiased way. It holds the potential to generate intuitions about the origins of such modular specialisation and about how the generation and modulation of behaviour, though functionally distinct, could be structurally integrated. This approach can be seen as combining the merits of the simulation models presented here and the work on evolving life-time adaptivity in fixed weight neural controllers (Yamauchi and Beer, 1994; Tuci *et al.*, 2002; Izquierdo-Torres and Harvey, 2007).

An important disclaimer to add here is that this kind of work would still focus on questions of localisation of function, not on questions of the origin of values. The ultimate goal is to simulate or artificially create value-generating processes in a minimally biased way. Using ER simulation models for this purpose is difficult, because ER simulation models are teleonomical – the fitness criterion is specified externally. Therefore, the purpose and function of behaviour is still fulfilling the norms of the experimenter, not of the evolved agent itself. This problem has been identified and made explicit. The reader's attention is drawn to two special issues, one on modelling autonomy (Barandiaran and Ruiz-Mirazo, 2008) and the other on modelling agency (Rohde and Ikegami, 2009), that have resulted from a series of workshops on this difficult question (noticeable contributions include Di Paolo and Iizuka, 2008; Ikegami and Suzuki, 2008; Egbert and Di Paolo, 2009; Barandiaran *et al.*, 2009). However, research on such models of processes with inherent values is still in its infancy.

Concerning the methodological theme of this book, the simulation model presented in this chapter demonstrates the kind of contribution that ER models can make to conceptual and philosophical debate: the model takes the proposed value system architectures, in their minimal form, to its logical conclusion, showing that, from the postulated principles alone, adaptation cannot be guaranteed. In making this theoretical point, the model also generates useful descriptive concepts to name the problems that occur (noticeably, the concept of 'semantic drift').

Using simulation models in order to add formal rigour to conceptual debate can be very satisfactory, because such models can address very general and far-reaching scientific questions, such as the origin and nature of value and meaning in adaptive behaviour. Such philosophical and exploratory models generate new ideas and concepts and they can challenge our intuitions or give credit to conceptual arguments whose logical soundness may otherwise be difficult to follow. The drawback of such an approach is, however, that these kind of simulation models produce less concrete results and do not directly relate to scientific practice, a concrete behaviour or task, a concrete target organism or a data set. They do

not produce hypotheses or directly suggest new experiments, as it is the case for the model of motor synergies presented in chapter 4. Both modelling approaches can be valuable, within their scopes and limits, in the study of human cognition and behaviour, as argued in chapter 3.

Arguably, philosophy is the most developed area in enactive cognitive science. What is most needed, in order to advance on the questions of mind and in order to hush critics like (Webb, 2009), is new data and hands on experimental and modelling work that establishes the usefulness of the enactive framework beyond a doubt. The remainder of the book deals with the application of ER modelling to perception research, as proposed in chapter 3, first to the problem of perceptual crossing and agency detection (chapters 6 and 7) and then about sensory delays and perceived simultaneity (chapters 8-11).

Chapter 6

Perceptual Crossing in One Dimension

The two ER models presented so far were rather different concerning their function and scientific question. The first one, modelling research in human motor control (chapter 4) has shown how ER simulations can generate proofs of concept that can rather directly resonate with hands-on scientific research. The second model, tackling a general architectural proposal in neuroscientific theory (chapter 5), demonstrates a more abstract philosophical value of ER models, i.e., to point out implicitly held prior assumptions in a theory and illustrate logical consequences from such assumptions that are counter-intuitive or difficult to understand. Applying ER modelling to simple sensorimotor perception research combines the merits of both approaches, i.e., the concreteness and 'meatiness' of the scientific ER modelling and the application to questions central to cognitive science, like the questions addressed with the theory-driven model of value systems. The question addressed here is about human perception of agency.

In this and the following chapter, the results from simulation models on the dynamics of human perceptual crossing in a one-dimensional and two-dimensional simulated environment are presented. The original work was conducted by the CRED group in Compiègne in two subsequent studies. This chapter starts with an introduction (Sect. 6.1)to the problem area and by presenting the results from the experiment on perceptual crossing in a one-dimensional simulated environment (Auvray *et al.*, 2009). The model of this study is briefly described in Sect. 6.2 and the modelling results are presented in Sect. 6.3. They are evaluated and discussed in Sect. 6.4. These results have been previously presented in (Di Paolo *et al.*, 2008).

6.1 Perceptual Crossing in a One-Dimensional Environment

The question addressed by both the model and the experiment (Auvray *et al.*, 2009) is about the role of global interaction dynamics in social interaction. Interaction of two or more individuals is a process of reciprocal causality. Such processes can lead to the emergence of dynamical patterns and global invariances that cannot be explained or understood by studying its components in isolation (cf. chapter 2). This means that phenomena dynamically emerging from interaction may not directly result from the individual capacities, intentions or actions of any of the partners. As (De Jaegher, 2007) argues in detail, the collective and global dynamics that characterise social interaction are frequently neglected when studying social cognition (in approaches such as 'theory of mind theory' or 'simulation theory'). Despite evidence to the contrary that suggests the importance of interaction dynamics in social processes (such as, for instance Kendon's findings (as presented by De Jaegher) that "synchronisation between interaction partners happens only when their mutual expectations of each other are exceptionally well attuned in the interaction" (De Jaegher, 2007, p. 149); many more examples are given in the cited source), traditional approaches focus on or even delimit themselves to explaining individual capacities.

(Auvray *et al.*, 2009) have designed an experimental paradigm to study the dynamics of social interaction in a minimal simulated environment. Two blindfolded participants are placed in separate rooms, in front of a computer. The virtual world that participants meet and interact in is one-dimensional and infinite, i.e., a tape that loops around (for technical and parameter details see (Auvray *et al.*, 2009), the model Sect. 6.2 and Fig. 6.1). Participants can move left and right on the tape, and whenever they cross an object, they receive a tactile stimulation to their fingertip through a Braille display. They are asked to indicate with a mouse-click when they believe a stimulation is caused by another feeling sensing intentional entity. Participants are told that, in the environment, apart from the other participant, there is a fixed object (fixed lure, at different locations for each of the participants) and a mobile object (the attached lure – it actually shadows the other participant's movement at a fixed distance but the participants do not know that). All of the entities have the same size in the simulated environment.

Therefore, the task is not only to distinguish moving and static objects, but to distinguish two entities that perform identical movement trajectories, only one of which is able to sense and respond to the encounter with the participant. In (Di Paolo *et al.*, 2008), we have compared this experimental set-up with Murray and Trevarthen's double-monitor experiments (Trevarthen, 1979; Nadel *et al.*, 1999), in which two months old babies were tested

for their capacities to distinguish a live interaction with their mother, mediated through a screen, from the presentation of a previously recorded interaction via this screen. The difference between the mother's behaviour on the monitor between the two conditions is only whether she senses the child and reacts to its actions or not; her expressive behaviour, i.e., her motion, language, mimics, voice *etc.* are identical between the two conditions. From the fact that babies get distressed and removed when presented with a previous recording, it is concluded that even two month old infants are sensitive to social contingency. In the light of the outlined tension between holistic and individualistic views on social interaction, the question to be asked is: does such sensitivity imply dedicated internal cognitive recognition/detection mechanisms of whether an interaction is recorded or not on behalf of the infant? Or does the difference between the two conditions emerge (partially) from the interaction, possibly involving much simpler mechanisms?

The results by (Auvray *et al.*, 2009) show that subjects are very successful at solving the task ($\approx$ 70% correct responses), without previous training and in spite of the poverty of the sensory information provided by the minimal simulated environment (a simple sequence of on-off tactile stimuli). Astonishing at first glance, the results are demystified after a simple analysis of the sensorimotor dynamics of the task and the strategy adopted by the participants to solve it. Participants search for stimulation and engage in local rhythmic scanning movements with any entity encountered on the tape. This rhythmic activity can only result in stabilised interaction with the other, not with the attached lure. When making contact with the attached lure, the lure shadows the movements of the other participant, who searches for stimulation by the other. Therefore, the lure does not act rhythmically and remain close like the participant would do in a real interaction, so the mutual search nearly inevitably results in interaction with the other participant, without requiring advanced perceptual skills. This impression is backed by an analysis of the ratio of clicks per stimulation. It reveals that the probability of clicking after encountering the attached lure is equally high as the probability to click upon encountering the other. The 70 % accuracy results not from discriminatory capacity, but from the fact that the participants are much more frequently stimulated by the other than by the attached lure, due to the fact that interaction with the other is a stable attractor in the task given the search strategy, whereas interaction with the lure is not. Even though the distinction between the fixed lure and moving entities appears to be made on an individual level (less clicks for the fixed lure per stimulation), the distinction between the attached lure and the other participant appears to result mainly from the interaction dynamics. Therefore, these results can be seen as a simple paradigm case of

how in embodied closed-loop interaction *knowing how* can be more effective in performing a perceptual distinction than *knowing that*.

6.2 Model

We decided to model the empirical study to see if an ER simulation model could provide further insights into the mechanisms and sensorimotor dynamics underlying this perceptual judgement behaviour, as it is outlined in chapter 3. By generating very simple artificial agents, and exploring the sensorimotor dynamics of the task in tractable, noiseless, idealised and fully controllable settings, we intended to support and enrich the insights gained from the experiment and to generate hypotheses for further research. Additionally, an actual synthetic proof of how the dynamics of an interaction process itself can produce agency detection behaviour (by excluding the possibility that other more complex human capacities contribute), rather than individual agency detection circuits, provides support for an interactionist approach in the study of social cognition. There have so far only been few ER models of social interaction (e.g., Di Paolo, 2000; Iizuka and Ikegami, 2004; Quinn, 2001) to provide such important proofs of concept.

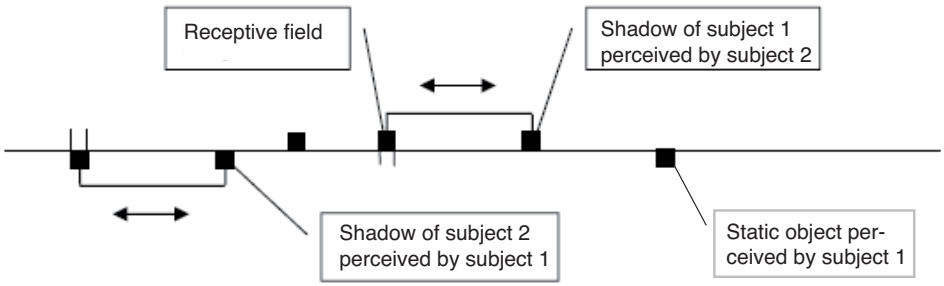

Fig. 6.1 Schematic diagram of the one-dimensional environment in the perceptual crossing experiment.

The virtual environment in the model is nearly identical to the one used in the empirical experiment. The length of the tape is $D = 600$ distance units and any entity on it (lure or participant) has a width of 4 units. One difference is that, while participants were just administered a single tactile input at any point in time, the input to the CTRNN controllers (as defined in chapter 3, Eq. (3.2)) consists of four neighbouring receptive fields of width

1 unit. The network generates two motor signals $M_{L,R}$ for left and right movement, $S_G, M_G \in [10, 1000]$ units per second.

The GA and evolutionary parameters follow generally the specifications outlined in Sect. 3.3. The network structure, however, is modified and partially evolved. The two motor neurons are treated as hidden neurons, i.e., the input neurons can connect to them directly and they can form recurrent connections with themselves or hidden neurons. The network structure (i.e., existence of up to five hidden units and synapses connecting the units) is evolved using the step functions $x > 0.7$ (for connections) and $x > 0.6$ (for hidden neurons) respectively. Other parameter ranges are $\theta_i \in [-3, 3]$, $\tau_i \in [20, 3000]$ ms, and $w_{ji} \in [-8, 8]$. In some runs, a sensory delay of 50 ms steps was applied. The trials lasted $T \in [8000, 11000]$ time steps.

Agents are tested against clones of themselves using an exponentially weighted fitness average (Eq. (3.5)) over six trials. The fitness criterion is the average relative distance $d(t)$ from the other across the trial:

$$F = \frac{1}{T} \sum_{0}^{T} 1 - \frac{d(t)}{300} \tag{6.1}$$

The task is thus to locate the other agent and spend as much time as possible as close to each other as possible while not being trapped by static objects or shadow images. This is a slightly different task than that posed to the participants, who were not given any explicit encouragement to seek the other. They were only asked to indicate their perception of another sensing entity. As the later model of the two-dimensional version of the task revealed (chapter 7), this modelling assumption biased the evolved behaviour to seek live interaction in a way that does not result naturally from the task. The reason for including this bias was to avoid the evolution of trivial but perfectly viable behaviour, i.e., to avoid interaction.

6.3 Results

First attempts to evolve agents to solve the described perceptual crossing task were unsuccessful. Evolutionary search got stuck in a local maximum, which corresponded to the behaviour to halt when crossing *any* object on the tape, be it the partner, the fixed object or the attached lure of the other. Given the simulated set-up and the fitness function, this is a comparably successful strategy: if agents first encounter each other, or if one agent runs into a partner waiting at the fixed lure, this strategy yields perfect fitness, and these are the majority of possible cases. However, it is not the optimal behaviour, as in the remaining

cases, the agents will not find each other at all, because they either both stop at their respective fixed lures, at a maximum distance from each other, or in a configuration where one agent stops on the fixed lure, and the other agent stops on its attached lure. Also, this is not a very intelligent or adaptive solution and does not resemble any of the strategies adopted by human subjects, who keep actively exploring the environment, even after they have found the other, engaging in rhythmic interaction. Only after a 50 ms sensory time delay between crossing an object on the tape and the agent's sensation was included into the model, active perceptual strategies evolved and the local fitness maximum of stopping when being stimulated by any source could be overcome.

While agents without delay evolved to simply stop, with the delay, they evolved to engage in rhythmic interaction. This means that both, the agent's discriminatory capacities are stronger, in an active sense, and that it remains distinguishable from the fixed lure for the other agent, in a passive sense. This finding (in accordance with the results from the two-dimensional model presented in the following chapter) indicates that there is a relation between oscillating scanning movements and the delay in the evolved agents. This further suggests that there may be a similar relation between the oscillatory strategies that most subjects adopt and the existent delays between sensation and reaction in humans. It seems natural to us that subjects would adopt a strategy such as oscillatory scanning. But why? It is not *a priori* necessary and even seems like a waste of energy. There are many possible explanations for this behaviour, but the model suggests that reaction time delays may play a role in shaping human crossing behaviour, like they do in the evolved agents. This hypothesis can be tested in further empirical experiments; it predicts that the phase of scanning oscillations is positively correlated with the amount of sensorimotor latencies in a task where such latencies are varied between different conditions.

The overall behavioural trajectories that the agents generate (Fig. 6.2 (A)) are similar to those generated by some human subjects (cf. Auvray *et al.*, 2009): phases of search are followed by phases of unstable rhythmic interaction with either of the lures or the other agent, until, at some point, rhythmic interaction between the partners stabilises for an extended period of time. Given the similarity of the task and virtual environment in the robotic simulation and in the empirical study, quantitative observations on the simulated data can be transferred to and tested against the human data in the spirt of more data-driven approaches to modelling. This was not true to the same extent for the more theory driven models in the earlier chapters.

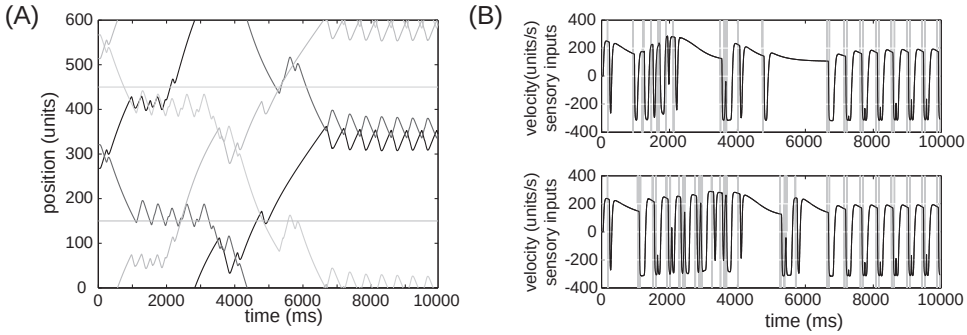

Fig. 6.2 Example behaviour evolved. (A) A trial resulting in stabilised perceptual crossing with motor noise (position across time; Agent 1 black, agent 2 dark grey; attached and fixed lures are lighter shades of grey). (B) Sensorimotor values for the behaviour depicted in (A). Agent 1 top, agent 2 bottom; velocity black, sensory inputs grey.

Monitoring the course of artificial evolution across many evolutionary runs, a consistent pattern is that avoidance of the attached lure evolves very quickly, while avoiding the fixed lure seems to take a long time (in accordance with the above reported difficulty to evolve such behaviour at all). These findings contradict the intuition that the easier task would be to recognise and avoid a static object, while distinguishing two entities that perform identical movements, only one of which responds to the perceptual encounter seems much harder. Embedding the evolved agents turns this intuition upside down.

One factor seemingly neglected in the model is proprioceptive sensation. It could be argued that detecting the invariant correlation between tactile and proprioceptive sensory input during active scanning would be a cue for distinguishing fixed objects from moving ones and that artificial agents cannot evolve this strategy because they do not have proprioception. This is, however, only superficially true. The neuro-controllers evolved allow for recurrent feedback to be used. In the simple virtual environment modelled, reafference of motor signal corresponds directly to proprioception and evolution could easily implement this strategy if it was advantageous.

A look at the data from the simulation model suggests a different explanation. The search strategy evolved in the artificial agents is to invert the movement direction once an object is sensed, thereby crossing the encountered object again, turn around, cross again, *etc.* This means that agents in interaction who both employ this strategy always cross at the same location in virtual space. There is a striking similarity of how sensation and motion evolve over time during rhythmic coordinated mutual scanning (crossing) and rhythmic scanning of a fixed object (see Fig. 6.3 (A) and (B) bottom). This coordinated activity leads to

sensations and motions changing over time in a way very similar to those that come about when investigating a fixed object (see Fig. 6.3 (B)).

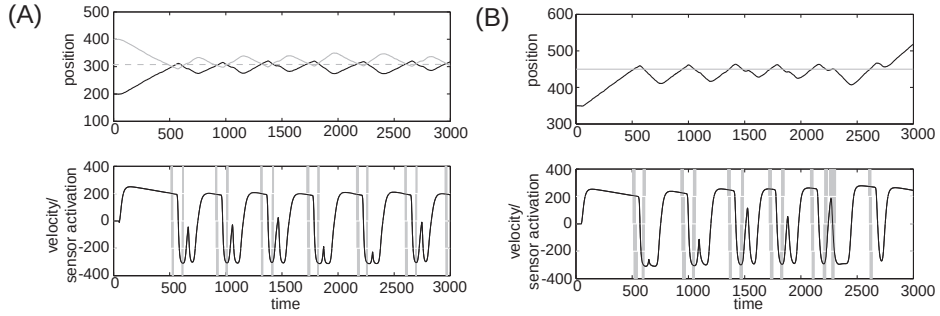

Fig. 6.3 Trajectories and sensorimotor values of interaction with a fixed object and with the other (details). (A) Stabilised perceptual crossing between two agents (trajectories and sensorimotor values; dotted line: location where perceptual crossing repeatedly takes place). (B) Scanning of a fixed object (trajectories and sensorimotor values). All diagrams include motor noise.

What is the strategy employed by the agents in order to distinguish coordinated interaction and a fixed object? The duration of the stimulus upon crossing a fixed object lasts longer than when crossing a moving agent. This is because the agent, even though it is the same size as the fixed object, moves in the opposite direction. Therefore, the simulated agent can integrate sensory stimulation over a longer period of time to perform its judgement. This yields a higher value for a static object, i.e., it is sensed as having a larger apparent size. Further support for this explanation comes from the fact that agents are quite easily tricked into making the wrong decision if the size of the static object is varied, i.e., a small object is mistaken for another agent and a larger agent is perceived as a fixed object.

The smaller perceived size in the case of perceptual crossing depends on encounters remaining in anti-phase oscillation, which is an *interactionally coordinated property* as defined in (De Jaegher, 2007). The agents co-construct the appearance of the agents being of smaller size. The changes in velocity induced by stimulation are tuned to this smaller perceived size. The close timing of the two perceptual crossings and the double drop in velocity they induce lead to coordinated oscillation around a fixed point of interaction. In turn, individuals respond to this emergent coordination by remaining in coordination with the apparently smaller object (see Fig. 6.3 (A)). In the case of the scanning of the fixed object, however, the longer sensation is integrated to reinforce the positive velocity signal when crossing over the object, i.e., to cross further over the object before crossing it on the way back back. This temporally displaces the two crossings of the object, which means

that the return trajectory is decomposed into two separate dips in velocity, disrupting the evolved stable oscillatory interaction in anti-phase. Note that also here, the integration of sensory stimulation time does not rely solely on internally integrating the sensory signal but is escalated in interaction: the temporal disruption of the oscillation can result in additional crossings, causing further sensory stimulation, which in turn reinforces the crossing velocity to the point that the agent leaps far across the object and agent escapes the attractive behaviour of rhythmic scanning (see Fig. 6.3 (B)). Such an integration of external variables and factors (position, velocity) into strategies to perform distinctions is very typical for closed-loop ER models. These kinds of sensorimotor invariances and interactive strategies are typically not considered in explicit (or even implicit) design of open-loop controllers.

6.4 Discussion

There are many viable solutions to the task, and it is rather unlikely that humans would use a strategy just as the one just described, as it appears rather specific to the conditions under which the agents were evolved. Even though the trajectories look qualitatively similar, the algorithmic preciseness with which interaction is initiated and maintained is very unlikely to be found in the human data. But, as argued in chapter 3, the point in modelling is not to recreate the original phenomenon but to identify invariant dynamical principles that remain robust upon idealisation and abstraction.

In a similar way as the model of motor synergies presented in chapter 4, the present simulation model generates a number of conceptual results that are interesting in an abstract sense. As argued in the introduction Sect. 6.1, most of the research in social cognition is individual-centred. The modelling approach taken, in contrast, does not just look at the individual capabilities, but also at phenomena that emerge during embodied and situated interaction. This broadened perspective leads to the inclusion of factors into perceptual strategies that are easily overlooked when only looking at open-loop behaviour: a task that intuitively seems difficult, i.e., to distinguish two entities with identical movement characteristics (the partner and the shadow image), becomes almost trivial, if the effects emerging from the mutual search for each other are taken into consideration. This finding already results from the minimal empirical closed-loop experiments by (Auvray *et al.*, 2009). The simulation experiments confirm this experiment and demonstrate beyond a doubt that this kind of behaviour can be realised without anything more complex going on, as the network controllers evolved are extremely simple, too simple to do anything more sophisticated than what was presented in the analysis.

Also, the simulation points out a different counter-intuitive state of affairs: distinguishing a moving entity (the other agent) from a static one, which intuitively seems very easy, is indeed a non-trivial task, if the emergent effects of interaction, i.e., anti-phase coordination, are taken into consideration. In the experiment by (Auvray *et al.*, 2009), 32.7% of the stimulation were caused by the fixed object, as opposed to 15.2% caused by the attached lure. This suggests that participants may also find the intuitively easier task of avoiding the fixed object more difficult, even if this increased difficulty does not manifest in classification mistakes (as explained in the introduction Sect. 6.1). There is evidence from both the model and the experiment that the distinction that arises mainly from interaction dynamics (which moving object is the other agent?) is more efficiently solved than the distinction that requires individual recognition capacities (is the entity I am scanning the fixed object or the other?).

With this global view on the dynamics of perceptual crossing in the investigated set-up, these insights may seem almost trivial. However, had we started from the perspective of the individual and its conscious recognition capacities (such as 'theory of mind' approaches in social cognition), these findings would be mysterious – just as (Trevarthen, 1979)'s results from the double monitor paradigm seem mysterious when focusing on the individual perspective, not on the interaction dynamics. However, in the light of the simulation results, the fact that babies would be sensitive to the social contingency of a situation does not seem that astonishing or sophisticated anymore.[1]

The close match between the experiment and its model, however, makes it possible to also generate quantitative hypotheses about the gathered data in the more traditional sense of mathematical modelling in science. The strong abstraction from the modelled phenomenon underlying the models presented in the previous chapters fiercely limited their potential for such concrete predictions of experimentally measurable results. One hypothesis that the model generates results from the described strategy of distinguishing fixed objects and anti-phase rhythmic interaction by means of integrating sensory stimulation time. The model suggests that one of the predictors for this decision will be an apparently smaller object scanned. The researchers of the CRED group favour a different explanation for this decision, i.e., "something that resists being spatially determined" (Auvray *et al.*, 2009, p. 18), which is valid for the experimental data but not for the noiseless model. Interestingly, however, the experimental data *also* supports the hypothesis generated by our model. Decreased

[1]This logic also works the other way around: when communicating Trevarthen's results to computational neuroscientists, biologists and other people familiar with dynamical systems, they have a tendency to be not in the least impressed or surprised about the baby's behaviour.

stimulation time due to opposed movement is a good predictor for when participants click ('event E6' in (Auvray *et al.*, 2009)).

The ER model successfully predicts human sensorimotor behaviour, which some researchers seem to find difficult to imagine, given that it is both simple and, at the same time, 'opaque' – much like an animal model. Which of the two valid hypotheses is true could be easily tested in further experimentation in which humans asking them to distinguish objects/agents of different size. This test helped to establish that the strategy observed in the evolved agents really was the one we seemed to recognise ('pseudo-empirical' investigation, compare chapter 3). Another quantitative prediction generated from the model has already been mentioned in Sect. 6.3, i.e., that there would be a proportional relation between sensorimotor latencies and the variation in the magnitude of oscillatory scanning. The researchers who conducted the experimental study published their results long after conducting the study and also, long after the model here presented was implemented and its results published. Referring to the simulation model presented, they write:

> "Their evolutionary robotics simulations showed similar results as the one reported in our study. Interestingly, and contrarily to any a priori prediction, Di Paolo and his colleagues found it easier to evolve agents that can distinguish between the avatar and mobile lure than agents that can distinguish between the avatar and fixed object. As a consequence, according to Di Paolo and his colleagues, in the case of social interactions, it is simply not necessary to evolve simulated agents with an individual contingency recognition strategy, given that the social process takes care by itself of inducing the individuals to produce the right behavior" (Auvray *et al.*, 2009).

It is reassuring that experimental researchers see and value the proofs of concept that ER simulation models provide for their research.

One important difference between the set-up investigated by (Auvray *et al.*, 2009) and (Trevarthen, 1979)'s double TV monitor experiments is that in the double TV monitor experiments, the baby is only either confronted with its mother or with a recording of its mother, whereas in the experiments on perceptual crossing, the other participant and its attached lure are presented at the same time. It could be argued that the dynamic distinction emerging from the interaction dynamics in the perceptual crossing experiments is specific to the set-up because of the linkage between the attached lure and the other participant. As long as the other participant is still searching, the attached lure keeps moving away, shadowing the search trajectories and making stable interaction impossible, which is not the case in the double TV monitor experiments: infants could, in principle, enter in one-sided interaction with the recording of their mother.

The simulation modelling work on the dynamics of perceptual crossing was extended in (Iizuka and Di Paolo, 2007; Di Paolo *et al.*, 2008) to a scenario that is closer to (Trevarthen, 1979)'s double TV monitor paradigm in this sense. In an equally simple ER simulation, artificial agents were evolved to distinguish between a recording with another agent and a live interaction. The resulting agents use a very simple, yet very effective active perceptual strategy. Agents oscillate around each other, in anti-phase oscillation. If a previous interaction is replayed using identical starting positions, similar behaviour is observed initially, which is a case of one-sided interaction. However, the agents sporadically induce perturbations (fast sideways 'jump') into the apparent interaction, in order to probe whether they are being followed and can thus find out if interaction is live. If the agent interacts with a recording, the break-down is irrecoverable due to the lack of mutuality in the interaction, whereas in a genuine two-sided interaction, the other agent reacts to the perturbation induced and restores rhythmic interaction. This demonstrates that perceptual distinctions between live interaction and recorded interaction, as they have been observed for infant-mother-interaction through a video link, may effectively be realised by very simple sensorimotor principles, maybe even accidentally or epiphenomenally, when a sudden unreciprocated reflex movement causes the breakdown of one-sided coordination. Behaviour that appears complicated on the surface and that seems to require elaborate information processing and internal models of personhood may thus result from very simple sensorimotor circuits. Unpublished follow-up experimental research (Di Paolo, Wood & De Jaegher; independently: Iizuka) has tested the human capacity to perform this distinction as an extension of the presented research on perceptual crossing (live perceptual crossing was suddenly replaced with a recording of the previous interaction). This research confirms that humans are sensitive to social contingency in this minimal virtual environment, and that simple action-perception strategies can produce behaviour similar to the one reported for the double TV-monitor experiments (Trevarthen, 1979).

Concerning the implementation of a dialogue between empirical studies and simulation models (Sect. 3.6), the model presented in this chapter demonstrates how this can be done: an experimental study is modelled in a computer simulation, which increases our understanding of the data obtained, because the simulation produces results that go beyond our cognitive limits and prejudices and is, at the same time, easier to understand than the original phenomenon. From these results, an extended version of the experiments is generated, which is first investigated in simulation, leading to refined hypotheses and ideas. These ideas are then tested in empirical experiments.

The experimental paradigm, despite its simplicity, is very rich and the possibilities for further research are open-ended and keep being explored experimentally and in simulation (different follow-up models of the experiment include (Martius *et al.*, 2008; Fröse and Di Paolo, 2008)). The following chapter presents a simulation model of such an experimental extension of the research by the CRED group that is a direct extension of the paradigm modelled in this chapter to a two-dimensional scenario.

Chapter 7

Perceptual Crossing in Two Dimensions

Despite, or maybe because of the simplicity of the experimental paradigm, the investigation of perceptual crossing in a minimal virtual environment serves to generate important insights into the potential role of the autonomous dynamics of interaction processes in social scenarios. The CRED group has extended the presented research to a two-dimensional scenario. The results from this experiment have not been published yet, but a combined publication of the two experiments alongside the modelling results presented in this chapter is in preparation (Lenay, Rohde & Stewart, in preparation). The model aims at elucidating, amongst other things, the role of human arm morphology in the generation of the quantitative properties of the recorded data. The results from this model have been published in (Rohde and Di Paolo, 2008).

The following Sect. 7.1 briefly introduces the extended experiment, its scientific purpose and that of the model. The model itself is described in Sect. 7.2. Three morphologically different types of artificial agents were evolved on the task and it was found that the dynamical principles that govern the task are independent from agent bodies. The realisation of these invariant principles, however is variable and depends on agent specific sensorimotor properties. Such variability in evolved solutions includes the evolution of one-dimensional oscillation along a line in a simulated arm agent, a kind of behaviour that had been observed in the participants in the original experiment as well. The results are presented in Sect. 7.3 and discussed in Sect. 7.4.

7.1 Perceptual Crossing in a Two-Dimensional Environment

Having investigated and analysed the dynamics and principles of perceptual crossing in a one-dimensional scenario (see chapter 6 and Auvray *et al.*, 2009), Lenay *et al.* (personal communication) extended the experimental set-up to a two-dimensional virtual toroidal

environment. With this modified set-up, the group wanted to test whether the experimental results transfer qualitatively or quantitatively from a one-dimensional to a two-dimensional scenario, which is by no means guaranteed: the sensorimotor contingencies afforded by the two-dimensional simulated toroidal environment are more complex and very different from those in the one-dimensional version.

A preliminary result from their study is that the data from the new version of the experiment is indeed surprisingly similar to the data obtained in the one-dimensional version. Not only do the results transfer qualitatively in terms of success (i.e., 65% correct clicks), but also the quantitative aspects of the behaviour are remarkably similar. In particular, interaction with an object or the other participant was realised by moving rhythmically back and forth along a line, reducing action to just one dimension, even though both dimensions were explored during search.

One of the hypotheses explored here in simulation is that this rhythmic one-dimensional interaction is related to the morphology of the human arm. The simulation model presented in this chapter aims to establish, amongst other things, the role of human arm morphology in the constitution of quantitative aspects of behaviour. Therefore, a simple simulated arm agent was modelled and compared to two other kinds of artificial agents, i.e., a two-wheeled robotic agent and an agent that generates a velocity vector anchored in Euclidean space, similar to a joystick (called the 'Euclidean' agent; details of the environment, tasks and agents modelled in Sect. 7.2). This latter type of agent can be seen as directly extending the agent architecture used in the model of the one-dimensional version of the experiment, whereas the sensorimotor couplings of the other two agents in the task are radically different.

The objective of comparing these different kinds of controllers is to identify common dynamical principles that derive from the task and the environment and that are relatively independent of embodiment and to distinguish them from qualitative and quantitative aspects of behaviour that are specific to a certain type of body or sensorimotor coupling.

The results point out some interesting common principles and quantitative differences. For instance, one-dimensional oscillation along a line evolved in the Euclidean and the simulated arm agents but not in the two-wheeled agents. Also, a very efficient strategy evolved, which is counter-intuitive and contrasts with the strategies employed by the human participants: agents establish stable interaction with the fixed lure and avoid the other agent. This is because the fitness function was changed from the model of the one-dimensional version of the experiment (chapter 6). Both this surprising strategy and the finding that one-

dimensional rhythmical interaction can result from arm-like agent morphology increase our understanding of the dynamics afforded by the task and lead to generalisations that can be tested by re-analysing the data and extending it through further experimentation.

7.2 Model

As it was the case in the model of the one-dimensional version of the experiment, the simulation used for the evolution of artificial agents was, apart from parameter details, identical to the one used in the original experiment.

The simulated environment is a (200×200) virtual torus, i.e., a plane that wraps around in both dimensions. In this plane, there are six different objects. Two circular simulated agents of diameter 20, two mobile lures that are attached to the agents (at a fixed distance and angle) and two fixed lures that are statically installed at $(50,50)$ and $(150,150)$ respectively (see Fig. 7.1 (A): the agents are the circular objects, the attached and fixed lures are depicted as boxes in this and the other figures, even though they are also circular of diameter 20 in the simulation). The attached lures shadow the trajectories of each of the agents at a distance of 93 units, being attached in perpendicular directions.

The only sensory signal S that the agents receive is a touch signal, i.e., if the distance d between the agent and something else is $d < 20$, an input S_G (sensory gain, evolved) is fed into the control network. Each agent can only perceive the other and one of each kind of lure, i.e., the dark agent can perceive all light objects in Fig. 7.1 (A), but not the dark ones, and vice versa, in order to make it impossible that interaction between the agents is mediated by another object that both agents perceive at the same time.

In order to investigate the role of morphology in the strategies evolved, and in particular the role of arm morphology, three different types of agents were evolved (specification below). For purpose of comparison, all three kinds of agents are controlled by structurally identical CTRNN controllers (compare chapter 3, Eq. (3.2)) with one input neuron, four fully connected interneurons and five output neurons (Fig. 7.1 (B)). Four of the output neurons regulate the two motor outputs: $M_1 = M_G(\sigma(a_{M1}) - \sigma(a_{M2}), M_2 = M_G(\sigma(a_{M3}) - \sigma(a_{M4})), M_{1,2} \in [-M_G, M_G]$ with M_G being the evolved motor gain. These outputs are interpreted as $v_{l,r}$, $v_{h,v}$ or $\omega_{e,s}$ for different agents respectively (see below). The task is to interact with something and correctly classify if the object encountered is either of the lures or the other agent. The fifth output neuron generates the classification signal M_C to indicate whether interaction is with another agent (output $M_C > 0.5$) or with one of the lures (output $M_C \leq 0.5$).

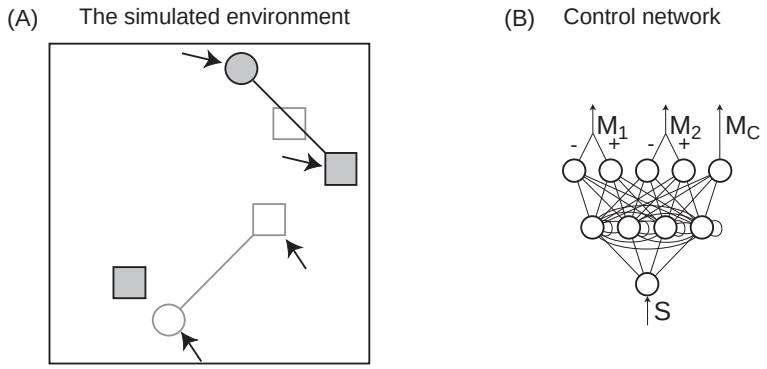

Fig. 7.1 Schematic diagram of the simulation environment and control network. (A) The simulated environment with the two agents (circles), the attached lures (boxes attached with a line) and the fixed lures (boxes). (B) The control network.

The three agent types evolved where:

- *Two-wheeled agent.* The two-wheeled agent generates the velocity $v_{l,r} = 20M_{1,2}$ for each wheel (Fig. 7.1 (A); velocities are specified in units/s).
- *Euclidean agent.* The agent referred to as the 'Euclidean' agent generates a horizontal and a vertical velocity vector $v_{h,v} = 30M_{1,2}$ that are summed up to define a vector in absolute space (Fig. 7.1 (B)). This agent can be seen as the two-dimensional analogy to the agent generating left and right movement modelled in the one-dimensional model in chapter 6.
- *Arm agent.* A simple simulated arm with two segments of length 400 units that is steered through angular velocity signals $\omega_{e,s} = 0.05M_{1,2}$ to the elbow and the shoulder joint (see Fig. 7.1, (C)). In order to approximate the dynamics of human mouse motion, the arm agent is restricted in its movements in two ways: through joint stops $\alpha_s \in [0.1\pi, 0.6\pi]$ and $\alpha_e \in [0.2\pi, \pi]$ and through the delimitation of movement to an area of 600×600 units that represents the 'desk' surface (i.e., the area within which a human participant would move the mouse), whose bottom left corner is fixed at $(-200, 200)$ taking the shoulder joint as the origin. The desk area is translated randomly with respect to both the desk area of the other agent and the simulated virtual environment to avoid that agents evolve to meet in the middle of the desk.

A problem with the simulated arm agent was that it has no way of telling where with respect to its anchoring in absolute space it is, because it has no proprioceptive sensors that represent its joint angles or any other form of telling where it is and whether it is still

moving or has run up to a joint stop. This is the reason why the arm agents did not evolve to a high level of performance (see Sect. 7.3). A modified version of the arm agent with three sensory neurons that received the joint positions as additional inputs ($S_{2,3} = S_G\theta_{e,s}$ was evolved for purposes of comparison. However, many of the original questions were already addressed with the original defect set-up, so this amended version of the arm model was not tested exhaustively. Controllers for all three kinds of agents were evolved without sensory delays and with a 100 ms sensory delay.

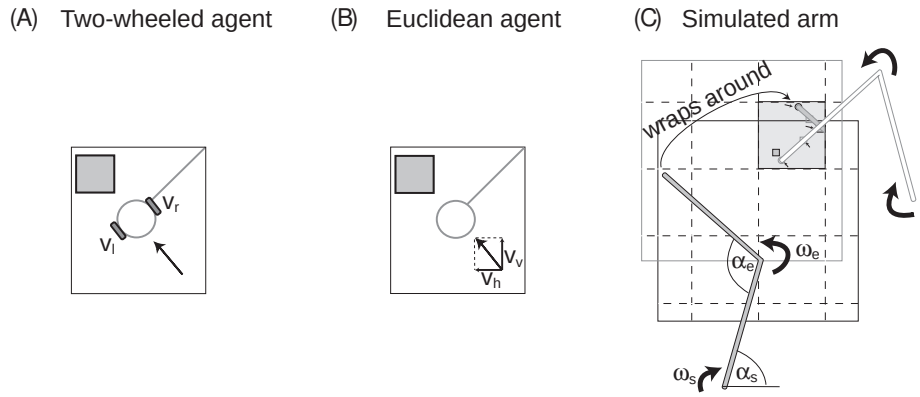

Fig. 7.2 Schematic diagram of the different types of agents evolved. Diagrams of the two-wheeled agent (A), the agent moving in Euclidean space (B) and two simulated arm agents, with the space in which they can act (C).

The GA and evolutionary parameters were those specified in Sect. 3.3 ($r = 0.6$). Evolved agent controllers (characterised by 74 parameters) are matched against clones of themselves in the task. 10 evolutionary runs over 1000 generations were performed for each agent body, with and without delay. Parameter ranges are: $S_G, M_G \in [1,50]$, $\tau_i \in [20,3000]$, $\theta_i \in [-3,3]$ and $w_{i,j} \in [-6,6]$.

Each trial lasts $T \in [6000,9000]$ ms. The starting positions are random for the wheeled and the Euclidean agent and random within the centre area for the arm agent. The starting angle for the wheeled agents is random. For the arm agent and the Euclidean agent, the relative orientation of the agents to each other is random $\in \{\frac{-\pi}{2}, 0, \frac{\pi}{2}, \pi\}$. The fitness $F(i)$

of an individual i in each trial is given by the following function

$$F(i) = \begin{cases} 1 & \text{if } (d_s \leq D) \wedge (d_o > D) \wedge (M_C > 0.5) \text{ (true positive)} \\ 1 & \text{if } (d_s > D) \wedge (d_o \leq D) \wedge (M_C \leq 0.5) \text{ (true negative)} \\ 0.25 & \text{if } (d_o < D) \wedge (d_s < D) \text{ (ambiguity)} \\ 0.1 & \text{if false classification and } S > 0 \text{ (touch)} \\ 0 & \text{else} \end{cases} \quad (7.1)$$

where $D = 30$, d_o the distance to the closest of the two lures and d_s the distance to the other agent. Agents are tested on eight trials and fitness is averaged. This fitness criterion is conceptually different from the fitness criterion used in the one-dimensional version of the simulation model. It resembles the task posed to the human participants more closely, as the agents are not evolved to interact with each other but instead to correctly indicate the presence of the other agent. Interestingly, this relaxation of the pressure to seek interaction with the other agents led to the evolution of a preference for interaction with the fixed lure, as discussed later on in this chapter.

7.3 Results

7.3.1 *Evolvability*

The wheeled agent and the Euclidean agent evolve to a much higher level of performance (see Fig. 7.3 (A)), with the best individual from the best evolutionary run achieving nearly perfect performance, whereas even the best arm agent clearly stays below a fitness of 50% (Fig. 7.3 (B)). Part of the reason for this discrepancy is that the arm agent does not have means to orient itself in space. For the Euclidean and the wheeled agents, there are simple strategies (fixed motor outputs) that allow them to scan the space (i.e., to go into a non-horizontal or non-vertical direction for the Euclidean agent or to go around in circles/spirals/curves for the wheeled agent). The arm, however, will run up to a joint stop or the edge of the desk surface if it applies any constant angular velocity to any of the joints without receiving any sensory feedback about whether it is still moving or not. This disadvantage made evolution of the arm much more difficult and subject to randomness than those of the wheeled or Euclidean agent (cf. Fig. 7.4, bottom left).

Agents were evolved with proprioceptive inputs (joint angles) for comparison and they immediately achieved much higher levels of fitness (population average/best after 1000 generations in 10 runs: 0.33/0.70) and evolution was less noisy (Fig. 7.4, bottom right). Despite this patch of the model, the arm agent did not evolve to near perfect fitness like the

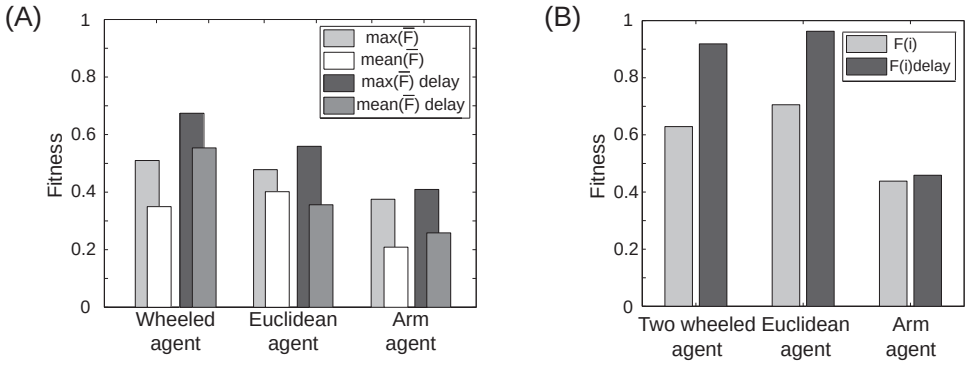

Fig. 7.3 (A) Population fitness average $\bar{F}$. Mean and maximum from 10 evolutionary runs, with and without delay. (B) Performance average across 100 evaluations for the best individual from the best evolution. Dark: 100 ms delay, light: no delay.

wheeled agent and the Euclidean agent did. Even though further exploration of how the arm model can be improved and made to approach the human example is and interesting problem as well, the question addressed with the model, i.e., the role of arm morphology in the constitution of rhythmical one-dimensional trajectories, could already be addressed using the simulation results with the sub-optimal solution.

All agents evolved to a higher level of performance with delays than without (see Fig. 7.3 (A)), as already observed for the one-dimensional scenario presented in the previous chapter. Figure 7.4 (top) depicts typical fitness evolution profiles for the wheeled agents without (left) and with (right) sensory delays. This shows that evolution without delays quickly converges to a non-optimal solution (local maximum), whereas evolution with delays converges as quickly to a near-perfect solution. The nature of this evolvability benefit provided by sensory delays is discussed in more detail in the following Sect. 7.3.2 and relates to the evolution of rhythmic interaction behaviour as opposed to search-and-stop behaviour.

7.3.2 *Behavioural Strategies Evolved*

Irrespective of agent body, two large classes of behaviour dominate the fitness landscape for the perceptual crossing task. The more successful strategy (1) is to avoid any mobile objects, search for the fixed lure, interact with it and always output 'no' ($M_C \leq 0.5$). This strategy can lead to perfect classification of encounters, and therefore to perfect fitness. Even though viable, this strategy is rather unintuitive (tongue-in-cheek, this strategy has been termed 'autistic' in Rohde and Di Paolo, 2008). It also clearly contrasts with the participants' behaviour, who avoid the fixed lure and seek interaction with each other. The

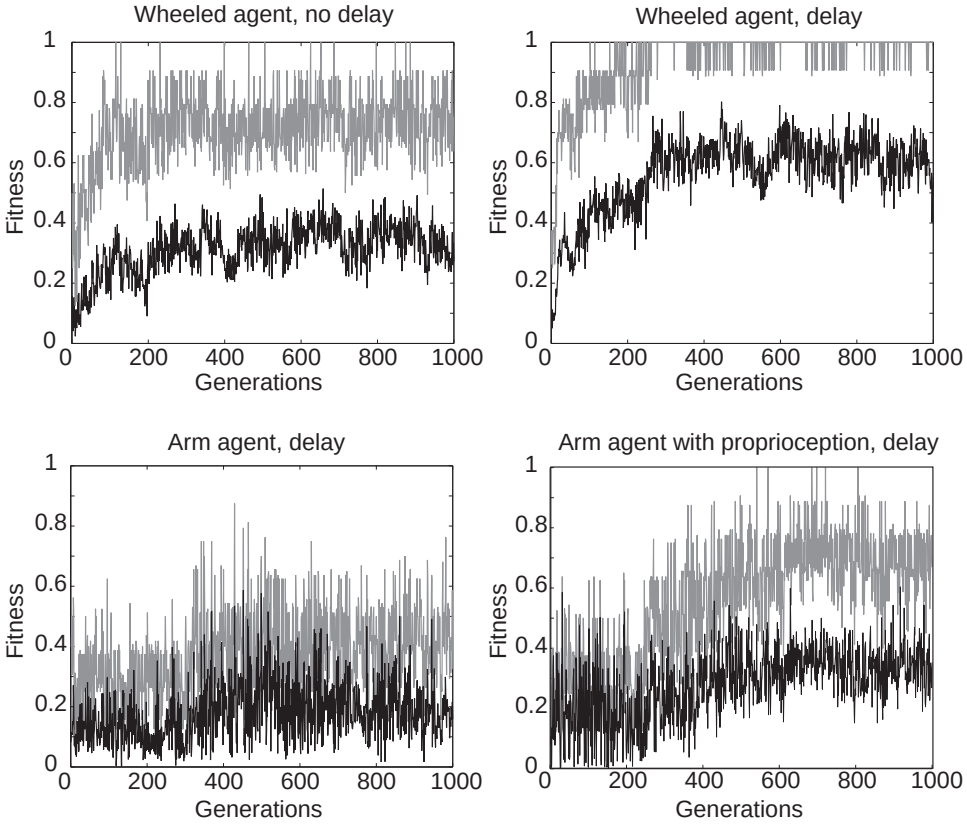

Fig. 7.4 Example evolution profiles for different agents and parameters, black: population average, grey: population best. Top left: wheeled agent, no delay (search-and-stop solution. Top right: wheeled agent, delay (rhythmic solution). Bottom left: arm agent delay (noisy). Bottom right: arm agent with delay and proprioception (less noisy).

second predominating strategy (2) is to interact indiscriminately with any entity encountered and to output 'yes' ($M_C > 0.5$) constantly. This strategy yields a fitness of up to ca. 40%. It appears that what evolved were preferences rather than discriminatory capacity; even if agents evolved to interact with all kinds of objects (strategy (2)), it appears to be more advantageous to exploit the slight combinatorial advantage of a permanent 'yes' answer over a permanent 'no' answer and not to intend a discrimination based on sensorimotor interaction with an object. The arm agents nearly exclusively evolve strategy (2), whilst the Euclidean and the wheeled agent evolve strategy (1), frequently passing during evolution through a phase of strategy (2). Only four agents (one arm, one wheeled, two Euclidean) evolved a contingent classification output triggered by stimulation (e.g., say 'yes'

if you touch something, in case you run into the other last minute and 'no' if stimulation continues over an extended period of time) additionally to a behavioural preference. The preference for interaction with the fixed lure contrasts with the experimental results and also with the synthetic results from the model presented in chapter 6, in which preference for live interaction and had been presupposed and built into the fitness function.

Both strategy (1) and strategy (2) involve localising another entity and staying close to it. Staying close can be realised, in principle, by rhythmical interaction with the target or by simply stopping where the stimulation does not cease. It appears that rhythmic behaviour is more adaptive: if we define, as an approximation, rhythmic behaviour as activity confined to a radius of $d = 50$ around an entity during the last second of a trial with at least five inversions of sensory state, we find that within each agent type for which both oscillating and non-oscillating solutions evolved, the oscillating ones were on average 9% more successful (see Fig. 7.5 (A); note that, due to the noisiness of arm evaluation, some of the rhythmic solutions evolved in arm agents with delay were not recognised by this approximate measure).

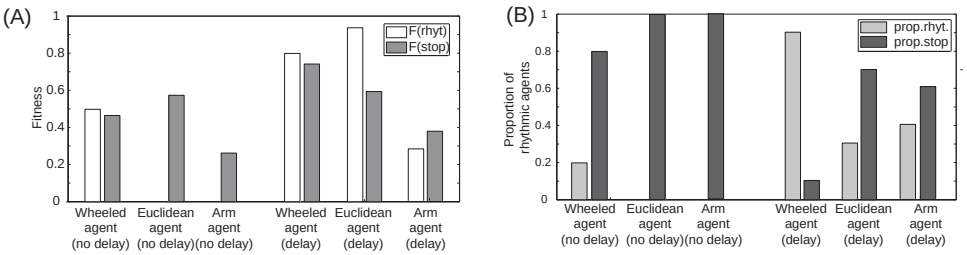

Fig. 7.5 Average of populations in which rhythmic behaviour was evolved and correlated fitness. (A) Fitness for rhythmic solutions (white) is on average much higher than that for non-rhythmic solutions (grey). (No rhythmic action was evolved for Euclidean or arm agents without delay; note that the measure for rhythmicity is an approximation as explained in Sect. 7.3.2.) (B) Proportion of agents that evolved rhythmic strategies for each of the conditions: the proportion of rhythmic solutions is much higher for evolutions with sensory delays.

The reason for the adaptive advantage of rhythmic strategies is that an agent evolved to simply stop is clueless where the stimulant has disappeared to if stimulation suddenly ceases. Such unexpected cessation can happen, e.g., when crossing an object at an unfortunate angle. It will start the search for sensation anew. An agent that interacts with an object rhythmically is moving repeatedly towards and away from its boundary and therefore has at least some capacity to relate its actions to the sensation of the object, inverting the effect of an action that makes stimulation go away. Thereby it establishes how it spatially relates to the object. With this minimal spatial interaction, if stimulation unexpectedly disappears,

the agent has at least the possibility to go into the direction of the last stimulation, which increases the probability to re-encounter the lost object.

As in the one-dimensional version of the model, integrated sensory stimulation over time that represents perceived size of the object is crucial for distinguishing fixed or mobile objects. In order to test this hypothesis, the size of the objects in the virtual environment was varied (just as in the one-dimensional version of the model). If the size of the other agent is doubled or the size of the fixed lure is divided by two, the fitness of the arm agents, who do not make the distinction between mobile or fixed objects drops only marginally altered 0.33 to 0.5/0.28 for doubled/halved respectively. These differences can be explained solely through the increased or decreased probability of making contact with another entity in the first place. For the Euclidean and wheeled agents that seek interaction with the fixed lure only, fitness deteriorates completely with these alterations, dropping from 0.69 to 0.11/0.07 and from 0.79 to 0.08/0.07 respectively, showing that their discriminative capacity is severely impaired by the alteration of size and the subsequent differences in integrated duration of stimulation during interaction.

Sensory delays seem to be crucially involved in bootstrapping the evolution of this kind of solution: rhythmic behaviour as defined above evolved to occur at least once in 10 trials in 2 of the 30 best individuals evolved without delays and in 16 out of 30 best evolved individuals with delay. With a delay, objects are only registered once an agent (in all three conditions) already shot past it. This forces agents to stop and return to the locus of stimulation, which is a more advanced behaviour and helps to overcome a local maximum in the fitness landscape, i.e., to stop upon any stimulation and start the search anew if stimulation unexpectedly ceases, which again bootstraps the evolution of effective and active perceptual strategies (cf. Fig. 7.5 (B)).

The exact realisation and behavioural dynamics vary quite substantially between conditions, as analysed in the following sections for the agents evolved with delays. The objective with this model was to explore the space of possible solutions and a detailed investigation of example agents (best agents evolved with delays) will help to understand and clarify those. In particular, it has been observed that, across agent bodies, two behavioural phases, search phase and interaction phase, can be realised variably and independent of each other.

7.3.3 *Two-Wheeled Agent*

Wheeled agents evolved a variety of strategies to search for objects in the toroidal environment: some shoot off in one direction, others drive around in large circles, arches or spirals.

When an object is encountered, interaction is either initiated immediately, or, alternatively, the agent backs off and comes back to see if the stimulating object is still there, a strategy which contributes to localising the fixed lure rather than the other agent or the attached lure in the 'autistic' solution to the task.

All wheeled agents evolved to drive in circles (of variable size) around the encountered entity, most of them aiming at a distance from the object that makes stimulation rhythmically appear and disappear. Figure 7.6 depicts a sample behaviour of the best agent evolved with average fitness $F(i) = 0.92$. Agent 1 (black solid line) is in stable interaction with the fixed lure throughout the time period depicted. Agent 2 (dotted solid line), on the other hand, is momentarily trapped in an interaction with agent 1's attached lure (black dotted line and grey solid line, t = [500,1500]). The interaction does not stabilise, because stimulation through the mobile attached lure is too intermittent, even though it is maintained over a number of crossings. The agent thus eventually abandons the lure, passes the other agent twice (both times touching it very shortly and, consequently, not performing a complete return trajectory, and then finds the fixed lure. This strategy only fails in very exceptional cases in which interaction with a mobile entity is phase-locked in a way that resembles interaction with a fixed lure.

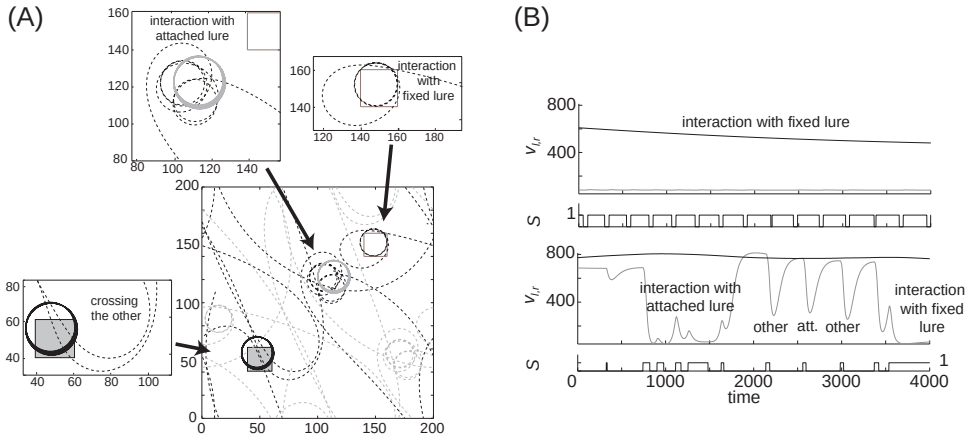

Fig. 7.6 Example trajectory and sensorimotor diagram for the best wheeled agent evolved. (A) The trajectory over the entire time period (large square) and local trajectories during significant sub-behaviours enlarged (small squares). Agent 1 solid line, agent 2 dotted line; agent movement black, movement of attached lure grey. (B) Sensorimotor diagram $v_{r,l}$ and S (rectangular) during the behaviour depicted in (A). Agent 1 top, agent 2 bottom.

7.3.4 *'Euclidean' Agent*

An architectural advantage that the Euclidean agents have is that the direction of their movement is anchored in Euclidean space. This inbuilt 'sense of direction' allows them to scan the space by applying a constant motor output, producing straight lines on the torus that wrap around it in a tight spiral (see slightly displaced lines in Fig. 7.7 (A); best Euclidean agent evolved with average fitness $F(i) = 0.96$). This is an extraordinarily efficient search strategy. Only two agents evolved to start search in a large curve.

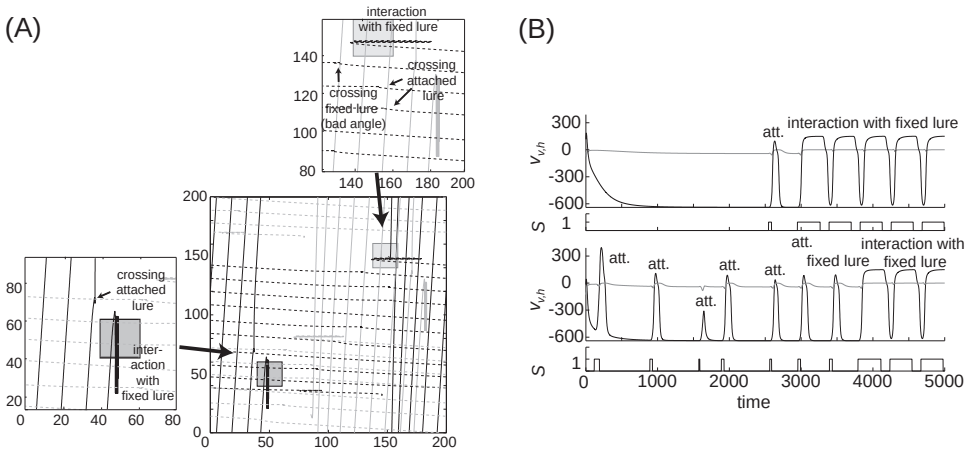

Fig. 7.7 Example trajectory and sensorimotor diagram for the best Euclidean agent evolved. (A) The trajectory over the entire time period (large square) and local trajectories during significant sub-behaviours enlarged (small squares). Agent 1 solid line, agent 2 dotted line; agent movement black, movement of attached lure grey. (B) Sensorimotor diagram $v_{h,v}$ and S (rectangular) during the behaviour depicted in (A). Agent 1 top, agent 2 bottom.

Figure 7.7 depicts the behaviour of the best agent evolved: if either of the agent encounters a mobile entity that moves perpendicularly, the stimulation is so short that the velocity is only minimally decreased ('kinks' in trajectories) and not even repeated crossing is initiated. The Euclidean agents exploit their absolute sense of direction because it constrains the angles at which they could possibly meet, due to the limited number of relative starting orientations.[1] Agents move either in parallel (unlikely to meet) or in orthogonal directions (very short stimulation).

Once contact with the fixed object is made, half of the agents evolve to simply stop upon stimulation, rather than to engage in rhythmic interaction. This tendency probably accounts for the slight population disadvantage of the Euclidean agents as compared to the wheeled

[1]This was the same for experiments with humans (they always started from the same orientation, which was identical for both).

agents. The other half evolve to rhythmically interact with the fixed lure along one dimension, implementing the 'autistic' strategy (1) to the task by making stimulation continually appear and disappear.

A behavioural pattern that only evolved in some of the Euclidean agents is to systematically destabilise even interaction with the fixed object, by slowly grinding past it (for stop solutions), or by moving further away with each oscillation (for rhythmic solutions). This strategy makes it possible to avoid interaction with mobile objects more efficiently and also breaks interaction in the rare occasions where interaction with a mobile object resembles interaction with the fixed lure. Even if this technique leads to the occasional loss of the fixed lure, due to the very efficient search strategy of the Euclidean agents, the probability to find it again quickly is very high. This strategy, as the strategy employed by the successful wheeled agents, is very effective and fails only in exceptional cases.

7.3.5 *Arm Agent*

As mentioned earlier, the arm agents evolved to much lower levels of fitness. This disadvantage is probably largely due to the fact that, other than the other two types of agents, arm agents do not have an easy way of exploring the environment. Without proprioceptive feedback, the agent has no way of telling where it is and whether it is still moving or has run up to a joint-stop or the edge of the desk. No constant output will yield any efficient search behaviour.

The agents evolved to either approach the desk edge in a large arch and then grind down the edge or to quickly go to one extreme arm position (neuron with fast τ) and then scan back in a large curve (neuron with slow τ). Both these scan behaviours fail if no object is encountered the first time this movement is executed. This enters randomness into the fitness evaluation, as behavioural success largely depends on appropriate objects lying on the path of the reflex-like movement executed by the arm. This makes evolution very noisy, as mentioned in Sect. 7.3.1.

From the original series, only one agent evolved a scanning behaviour that goes beyond the execution of one blind swaying movement: it makes use of a neural oscillator as central pattern generator (CPG). The trajectories it generates and the sensations and motions over time are depicted in Fig. 7.8.[2] This agent is the second best agent evolved, even though

[2]The trajectories generated are a bit difficult to interpret, because during each oscillation, a part of the previous path is exactly inverted by inverting velocity on one joint and decreasing angular velocity on the other joint to 0. This visualisation problem is quite common for solutions evolved in arm agents and also characterises the solution depicted in Fig. 7.9.

it has no sophisticated interaction strategy (i.e., sensation initiates the decrease of motor outputs to 0).

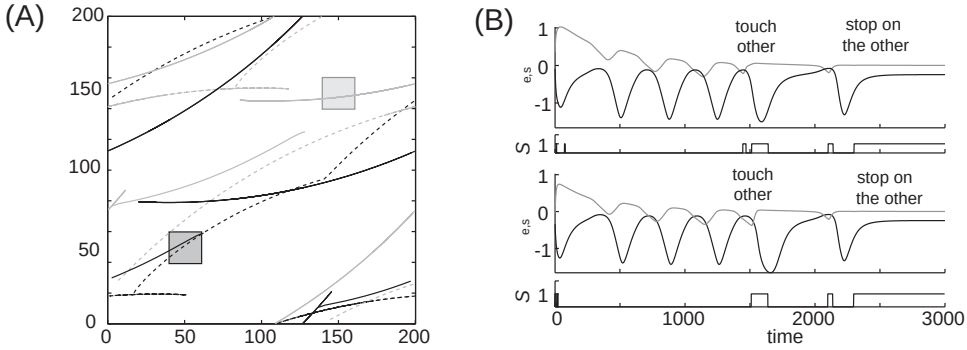

Fig. 7.8 Example trajectory and sensorimotor diagram for an arm agent that evolved a neural oscillator as central pattern generator. (A) The trajectory over the entire time period (large square). Agent 1 solid line, agent 2 dotted line; agent movement black, movement of attached lure grey. (B) Sensorimotor diagram $\omega_{e,s}$ and S (rectangular) during the behaviour depicted in (A) clearly shows the oscillatory outputs in the absence of sensory inputs. Agent 1 top, agent 2 bottom.

Nearly all arm agents evolve to rhythmically interact with any entity encountered (even if that is not always recognised by the criterion specified in Sect. 7.3.2), making the sensory stimulation constantly appear and disappear. The best agent evolved with average fitness $F(i) = 0.46$ (see trajectory and sensorimotor diagram in Fig. 7.9) implements this kind of behaviour. The rhythmic powering of one joint only leads to the exact inversion of the path just made (i.e., trajectories are difficult to follow in the figure).

As expected, the rhythmic activity in the arm agent leads to the production of near-straight oscillatory trajectories, as they were observed in human participants. The interesting aspect about this result is that, even though such trajectories did not evolve in all agent types (wheeled agents evolved to drive around in circles), it seems to be the arm-specific implementation of a general principle, i.e., the reduction of motion to oscillatory behaviour in one dimension of the output space only.

Looking at the behaviour and performance levels attained in the complementary evolution of arm agents with proprioceptive feedback reveals that, even though solutions do have higher fitness on average, arm agents with proprioception evolve still strategy (2), i.e., indiscriminate interaction. The resulting interaction behaviour is, in many ways, similar to the behaviour evolved in successful arm agents without proprioception (Fig. 7.10 (A) and (B)), even if the localisation behaviour is more successful. The additional proprioceptive

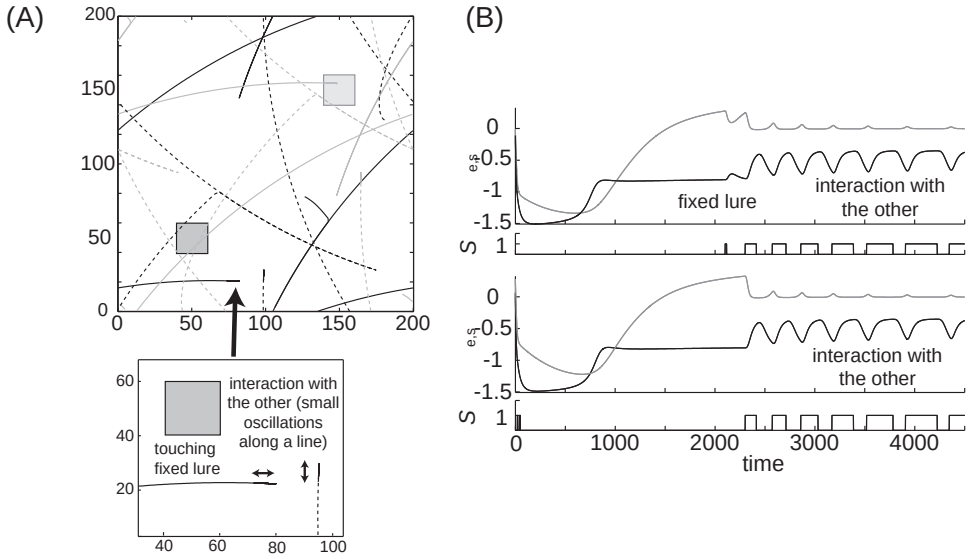

Fig. 7.9 Example trajectory and sensorimotor diagram for the best arm agent evolved. (A) The trajectory over the entire time period (large square) and the trajectory during interaction enlarged (small square). Agent 1 solid line, agent 2 dotted line; agent movement black, movement of attached lure grey. (B) Sensorimotor diagram $\omega_{e,s}$ and S (rectangular) during the behaviour depicted in (A). Agent 1 top, agent 2 bottom.

input mitigates some of the problems with noisy evolution and behavioural randomness associated with the impossibility of spatial orientation. It does, however, not lead to the evolution of perfect or near perfect solutions, such as strategy (1).

There are possibilities for further analysis of why this is so, and more ways of trying to further improve the arm agents' performance (such as longer evolution due to the larger parameter space). One of the main questions behind this model can, however, already be addressed with the sub-optimal results obtained. The results show how arm morphology produces oscillation along one dimension as the implementation of a general dynamical principle, i.e., rhythmic interaction along one dimension of motor space (see following discussion).

7.4 Discussion

A main result from this simulation model is that several dynamical principles govern the evolution of solutions to the modelled task. These hold across different agent bodies.

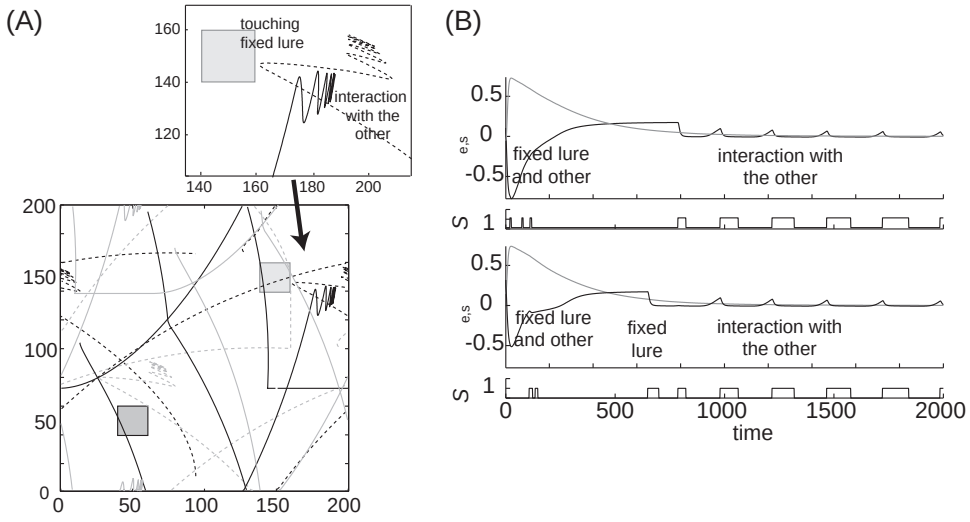

Fig. 7.10 Example trajectory and sensorimotor diagram for an arm agent evolved with proprioceptive feedback. (A) The trajectory over the entire time period (large square) and the trajectory during interaction enlarged (small square). Agent 1 solid line, agent 2 dotted line; agent movement black, movement of attached lure grey. (B) Sensorimotor diagram $\omega_{e,s}$ and S (rectangular) during the behaviour depicted in (A). Agent 1 top, agent 2 bottom.

- The search space of possible strategies is dominated by two principal solutions. (1) Avoid mobile objects, seek interaction with the fixed lure and output 'no'. (2) Interact indiscriminately and output 'yes'.
- Strategy (1) is the more successful strategy and yields nearly perfect fitness.
- Solutions that rely on rhythmic interaction are on average more robust to perturbations because they facilitate spatial localisation of the stimulant and thus yield higher fitness than solutions that rely on stopping on top of a stimulant.
- During this rhythmic interaction, one motor signal implements the oscillation, the other one is frozen and serves to adjust behaviour if necessary.
- Evolution of the superior rhythmic solutions is facilitated by the introduction of a 50ms sensory delay.
- Two different behavioural modes that can be realised variably and independently are identified: search and interaction.
- Despite the quantitative differences in how the behaviour manifests in space and time, the sensorimotor diagrams displaying sensorimotor activation over time are of remarkably similar appearance.

Apart from these commonalities, there are different quantitative properties associated with the realisation of these dynamical principles across the different agent bodies.

- The realisation of search and interaction behaviour is strongly influenced by agent morphology and the sensorimotor couplings that characterise and constrain the space of possible solutions.
- In particular, search behaviour can be particularly efficiently implemented in the Euclidean agent and is extremely difficult to evolve in the simulated arm agent. The difficulty of evolving search behaviour implies a drastic disadvantage in overall evolvability for the simulated arm agents.
- Rhythmic interaction behaviour is realised differently in all three agent types. In particular, wheeled agents circle around the object encountered, whereas the arm agent and the Euclidean agent engage in one-dimensional rhythmic interaction. In the Euclidean agent this implies oscillation along either the absolute vertical or horizontal dimension, while in the arm agent, oscillation of either of the joints results in slightly curved oscillations along the orientation of the arm.

These simulation results support the hypothesis that arm morphology plays a role in the one-dimensional rhythmic interaction observed in human participants, as the arm-specific implementation of a more general dynamical principle governing the task. They predict that in the gathered data, observed oscillations should be orthogonal to the orientation of the arm and that this oscillation should serve to establish rhythmic interaction with the encountered object or participant.

An interesting parallel with the one-dimensional version of the simulation study is that, again, sensory delays improve evolvability because they bootstrap the evolution of oscillatory scanning behaviour. This result suggests an investigation of dependencies between sensorimotor latencies and frequency of oscillation in the experimental data, just like the results presented in chapter 6. Also, integrated sensory stimulation time and how it correlates to perceived size of the object/agent appears to play a key role in distinguishing the fixed lure from the other agent. As in the one-dimensional version of the experiment, this synthetic result predicts that integrated stimulation time correlates to the decision made.

A difference between the experimental result and the modelling results presented in this chapter is that experimental participants seek interaction with the other participant, whereas, in the simulation the dominating strategy (1) is an ‘autistic’ strategy in which agents avoid each other and seek for the fixed lure. This surprising result also contrasts with the earlier simulation model, for which agents had been required to seek interaction with

one another, presuming a preference for live interaction. From these results, we concluded that perceptual crossing is, given the task, a nearly inevitable result from the mutual search of the agents/participants for each other (see chapter 6), even if this simulation already hinted towards the difficulty to avoid the static lure. In the light of the present simulation results it becomes clear that, leaving aside motivational factors (such as boredom), the dynamics of the task do not favour perceptual crossing, but much rather interaction with the static lure, and that perceptual crossing is established despite this strong basin of attraction. The results have been fed back to the researchers of the CRED group, who have conducted the experiment. They found that the simulation results clarified the role of morphology in the recorded behaviour and the evolution of autistic behaviour pointed them to an implicit presupposition in their formulation of the task. Further simulations to investigate the dynamical principles of the task have been suggested. Moreover, they have started to analyse the data gathered in order to test some of the principles that the model suggested to be relevant. Unfortunately, postural data had not been recorded in the experiment with humans, such that the orientation of oscillatory movements with respect to arm posture cannot be directly investigated. As a first approximation, however, they tested whether there is a direction-specificity in the oscillatory behaviour in absolute space. If there is no such specificity, it is highly unlikely that human arm morphology plays a role in bringing about one-dimensional oscillations. It appears that some subjects exhibit such a fixed orientation in their one-dimensional scanning, whereas others do not (no clear result yet). Other predictions from the model that are being evaluated in the data include the occurrence of oscillations during interaction and the occurrence of return trajectories after losing contact. It is not yet clear in how far these factors pointed out by the simulation model bear significance in the human data. In any case, the fact that the model has enriched and guided the analysis of the human data by suggesting potentially relevant variables and factors and that it provides the proofs of concept to back such suggestions up is, in itself, encouraging. A publication about the joint simulation and modelling results is in preparation (Lenay, Rohde & Stewart, in preparation).

The four simulation models presented in the previous chapters have addressed different kinds of research questions. The model of linear synergies (chapter 4) aimed at exploring a concept from human motor control research in strongly minimised and idealised settings, in order to generate hypotheses for further experiments and to generate proof that the postulated principles can work in theory. In a more philosophical endeavour, the model of value system architectures presented in chapter 5 caricatured a neural architecture pro-

posed as a mechanism for general behavioural adaptation, pointing out implicit premises underlying the proposed principles. The previous chapter and this chapter have applied minimal ER modelling to findings from PS research, proposed in chapter 3. As argued in Sect. 3.6, the close match between experiment and simulation allows a much stronger analogy between model and experiment that serves to generate quantitative predictions about experimental data from previous and future experiments, alongside with the more abstract proofs of concept and counter-intuitive insights resulting from ER as a tool for thinking in theory-building.

All four models have generated valuable contributions to the problem area they address. Arguably, none of the concrete simulation results add groundbreaking new insights to their respective field. However, they help to bring in an embodied, dynamical and enactive perspective into research practice and point out the non-obvious. Thereby, they show in how far this kind of modelling approach can be valuable in principle, not only for robotics and research on simple animals, but also for studying human level cognition, perception and behaviour.

The following chapters (8-11) present the results from a study on the adaptation to sensory delays and perceived simultaneity that combines experimental and simulation modelling work. The hypothesis put forward in chapter 3 was that a researcher should work across disciplines herself. Insofar, this interdisciplinary study can be seen as a test of what it buys to not only provide the models for an ongoing research program, from the outside, but to combine these different methods in person. Chapter 12 assesses the different modelling approaches presented in this book in the light of the overarching methodological theme.

Chapter 8

The Embodiment of Time

This chapter is the first of four chapters about time perception and time cognition. It is entirely conceptual and does not involve any modelling or experimental work in itself. It prepares the ground for the experimental study on adaptation to sensory delays presented in the following chapter and its model in the chapter thereafter. An overview about interesting work on time cognition and time perception from a multitude of sources is given. The conceptual links between the covered material are identified and explained.

The subjective perception and experience of time and its relation to temporally co-ordinated real-time behaviour are curious problems and possibly among the hardest in the study of human cognition. Time is ubiquitous. Findings about time perception and its embodiment presented in this chapter stem from disciplines as diverse as phenomenology, neuroscience, anthropology, psychophysics, philosophy, linguistics and psychology. Each of the sections below would deserve an entire book; the collage raises more questions than it provides answers. It is clear that important if not crucial perspectives are left out or incomplete, and likely that some of the conclusions drawn are either naïvely wrong or stating what others have found out more quickly and described in better words. This is an inevitable problem when dealing with a question like time perception and temporality, which is a phenomenon short of being as complex as mind itself. Most researchers working with the mind will have dealt with time or temporality at some point during their career. Giving a complete inter-disciplinary review of work on time is next to impossible.

The reason to attempt such a broad review in spite of this difficulty is that the views presented have shaped the experimental hypothesis investigated in the study on perceived simultaneity (chapter 9) and the perspective on time underlying it. The synthesis of recurring themes and links within this variety of research on time in disjoint disciplines at different points in history fuels a constructivist stance towards time perception. This chapter fuses a number of independent sources that all contradict our intuition, namely that mental time is

simple and logical. Some of these sources are old, are outside the natural sciences or are not very accessibly written and probably unknown to or deemed irrelevant by a large proportion of contemporary cognitive scientists. In order to grasp the gist of the experiments presented subsequently, it is helpful to name the sources of inspiration, their interpretation and how it is reflected in the approach taken to study the problem of sensorimotor recalibration of perceived simultaneity and sensory delays.

The chapter starts gently by decomposing Cartesian intuitions about what the experience of time is and how this view relates to traditional approaches in cognitive science to explain time perception (Sect. 8.1). For the largest part of this chapter (Sect. 8.2), the work of other thinkers and scientists is cited in order to oppose such a traditional and naïve view and replace it with a multi-tiered and rich picture of time perception and temporal behaviour. Since Kant's *Critique of Pure Reason* (Kant, 1974), and possibly even before, many authors have realised that our perception of the world flows, and that this is the most elementary and irreducible form of temporal experience. This changing flow, however, is a very primordial, low-level and unreflected form of temporality. On the other hand, time is one of the most abstract, ubiquitous and elegant constructs that the human mind reliably develops. Logical and mathematical transformations of temporal properties and relations are possible. From the enactive perspective, the question to be asked is the following: what is it in our bodies and our interactions with the world that gives rise to the peculiar categorisation of encounters into those that are present, those that are past and those that are future? How do we come to impose an absolute and irreversible order relation on all the events of our world? How do we distinguish events (i.e., temporal entities) from objects (i.e., spatial entities)? This question phrases a whole research program, rather than a research problem. The theoretical and broad perspective taken in this chapter is applied to the concrete problem of delay adaptation and simultaneity in the following chapters 9-11 that conclude the results part of this book. The final chapter 12 revisits the body of data presented in this book in the light of the underlying theme: the re-introduction of computer modelling into an enactive and embodied approach to cognition, by means of ER simulation modelling.

8.1 Newton Meets Descartes: The Classical Approach

What is a caricatured naïve stance towards time cognition? Crudely speaking, it assumes that there is an objective time in the world, a Newtonian time arrow, that imposes a global order on events (before, at the same time, after) and defines absolute temporal distances between temporal events (a day before, five seconds after). A naïve representationalist

and objectivist perspective on time cognition assumes then that time cognition is basically about having an internal clock, a mechanism to properly measure the timing of external events for our internal mental recreation of the world (e.g., Gibbon and Church, 1984).

This approach has indeed been implemented in early AI systems and models of time cognition. They use, e.g., temporal logic that extends propositional logic to include a time variable or tense stamps for each proposition (Allen, 1984). Similarly, indexicality with time stamps is used in formal semantics to disambiguate temporal language (Heim and Kratzer, 1998). The advantages of this view are (a) that it appeals to our intuition of what time is and how it works and (b) its simplicity. The disadvantage, however, is that with this view, one runs into three entire classes of drastic problems that are described in the following in a little bit more detail: ontological problems about the nature of real time; technical problems about computers acting in real time; a failure to account for the phenomenon of mental time.

Firstly, both Einstein's relativity theory and quantum mechanics have challenged our naïve intuitions about the objective reality. Experientially, time appears to us as an arrow and space as a three-dimensional Euclidean coordinate system that contains matter and objects with defined boundaries, in agreement with Newtonian physics. We are tempted to believe that this view corresponds to an objective, observer-independent reality. With the insights of modern physics, however, the most basic dimensions in which we perceive the world – time, causality, spatial extension, *etc.* – are shaken. Einstein's relativity theory has counter-intuitive consequences, such as the possibility of order inversion, time dilation and size contraction, all of which depend on the inertial system in which an observer is located. Heisenberg's uncertainty principle in quantum mechanics has led Schrödinger to think up the well-known and mind-boggling thought experiment about a cat that is both alive and dead, in order to criticise the Copenhagen interpretation. In reality, the counter-intuitive results from modern physics do not impact on our everyday lives – they concern events at very high velocities or on nanoscopic scales. Yet, they leave us wondering what can be said about a 'world out there', in the absence of us, the sense-making creatures, that only pick up on the structures that concern us, certain time-scales, certain spatial dimensions, certain forms of energy, *etc.*

(Bitbol, 2001) argues that the seeming paradoxes of quantum mechanics stem from the prevalent representationalist-dualist epistemology. If a constructivist epistemology with a "two-way set of relations between theories of knowledge and scientific theories" (Bitbol, 2001) is adopted, they can be resolved. This means, however, to accept that the observer

is an essential part of the scientific story, that the natural sciences do indeed not describe nature, but much rather the "interplay between nature and ourselves" (Bitbol, 2001). Venturing briefly into the domain of metaphysics, Bitbol's argument, which is only partially reproduced here, shows that the very idea of an observer-independent reality or universal scientific truth is flawed, even in the 'hardest' science of all, i.e., physics. The epistemological constructivism in physics that Bitbol describes comprises the observer-dependence of time (Bitbol, 1988).

Rejecting an objectivist world-view, obviously, does in no way contradict the construction and usage of clocks as tools for time measurement or the concept of an absolute time arrow and Newtonian physics as helpful mental constructs. Indeed, dynamical systems theory, which, as argued in chapter 3, is one of the prime mathematical and scientific tools for the enactive approach, employs Newtonian absolute time as an *a priori* variable, an atom of explanation. What is important is that it has to be made explicit that time as a useful mental and technical tool does not possess any kind of ontological priority or reality over the rest of our useful mental constructs and, therefore, at some level, requires explanation, just like all the others.

The second point is about the problems that GOFAI systems have with acting in real-time. These have been described by critics of the computationalist paradigm many times and have already been addressed in chapter 2. A system that exists in time and aims to represent the passing of time gets into trouble coordinating the internal and external time arrow. As (Cantwell-Smith, 1996) points out, in the case of a clock, this coordination is all it does and the closer the clock comes to mimicking the natural processes that were chosen to define temporal units, the better the clock. In the case of a digital computer, things are more difficult, because the formal language in which it is defined (automata theory) disregards real time, which means that any Turing machine can be instantiated in different ways that are temporally contingent, by adding an external clock with arbitrary time scale or exactness to the computational process. The implicit premise in (Turing, 1950)'s 'Computing Machinery and Intelligence' is that exact timing is irrelevant to intelligence. This premise has been criticised many times by different authors. To name but a few: Cantwell-Smith's criticism that "[traditional models of inference] take the temporality of inference to be independent of the temporality of the semantic domain" and that these need to be at least partially coordinated (Cantwell-Smith, 1996, p. 259); van Gelder's diagnosis that the computational hypothesis treats "time as discrete order" rather than a real-valued variable in his plea for the dynamical hypothesis in cognitive science (van Gelder, 1998, p. 6); Harvey *et*

al.'s observation that computational systems are "a rather specialised and bizarre subset" of dynamical systems which are characterised by the fact that "updates are done discretely in sequence, with no direct references to any time interval" and are thus instantiated with accidental real-time properties (Harvey *et al.*, 2005, p. 6). All these authors come to the same conclusion: the need for embodiment and embeddedness in real-time interaction and a formalism that unifies model-external and model-internal time. This realisation is already half the way towards an enactive approach, even if a mitigation of the shortcomings in computational systems by inclusion of an explicit clock and partial co-ordination is a half-blooded possibility (e.g., Clark, 1998; Cantwell-Smith, 1996).

The third point is the most obvious point and can even be argued against a dyed-in-the wool objectivist. Even if it were the case that time was basically Newtonian and even if there were no problems of synchronising the represented time in a Turing Machine with this real time, a simple fact is that *mental time does not work that way*. To take the most trivial example, everybody knows that in our experience, sometimes, time flies and sometimes, the hours go incredibly slowly. This is but one and one of the less interesting examples of how our mental time behaves strangely and at odds with Newtonian physics. There is simply no evidence for a central, linear and dedicated internal clock mechanisms, and many authors in cognitive science, even if they do not affiliate with enactive or constructivist approaches, have developed proto-constructivist views on time perception on the basis of empirical evidence. For instance, (Ivry and Schlerf, 2008), in a recent review of evidence and models of time perception, conclude that "neuropsychological research generally has promoted models in which time is represented by dedicated neural systems", whereas "recent physiological and computational studies have highlighted how temporal information is reflected in the intrinsic dynamics of neural activity". They refer to a recent model of psychophysical duration judgements (Karmarkar and Buonomano, 2007) that uses the inherent dynamical repertoire of a big recurrent neural network and predicts, amongst other things, nonlinear interactions between perceptual judgements in humans (see Sect. 8.6 below). In a similar way, but from a more phenomenal perspective, (Dennett and Kinsbourne, 1992) argue that it is erroneous to suppose that "there must be some place in the brain where 'it all comes together' in a multi-modal representation or display" and that "there is no one place in the brain through which all these causal trains must pass in order to deposit their contents 'in consciousness'". Both the intrinsic models of time perception described in (Ivry and Schlerf, 2008) and the multiple drafts model proposed in (Dennett and Kinsbourne, 1992) are in some ways similar to the enactive view on temporality developed here.

Taking these three classes of problems together, they suggest one common thing: why make the effort of modelling an internal clock, a separate and dedicated mechanism, to temporally tag cognitive events and strive for coordination of internal and external time, when this is not even what we humans do? Are we not still perfectly able to act in real-time despite our messed-up mental time that resists logic and Newtonian physics, despite lack of biological evidence for a central dedicated clock mechanism? A Newtonian-Cartesian-cognitivist approach smoothes over the real puzzles and mysteries of time cognition even before scientific work starts. The classical computationalist modeller will end up wasting her time solving artificially induced technical problems resulting from the choice of formal language, trying to co-ordinate internal and external time, but not address any of the real questions. The puzzle of how coordination is achieved in the light of latencies is passed on to a presumed homunculus that works with the skillfully constructed internal representation of external time. By contrast, a constructivist enactive approach sets out to find meaningful sensorimotor invariances, circuits *knowing how* to predict and coordinate, rather than *knowing that* temporal relations exists. It takes into consideration the natural habitat and evolutionary history of the human species, and thus tries to explain what leads us to construct our perception of time so stably across different domains of time. Such an approach is infinitely more difficult, yet infinitely more satisfactory.

8.2 Time and its Many Dimensions in our Mind

The remainder of this chapter attempts to represent in a text a complex landscape of evidence and ideas that thinkers and scientists have expressed on time cognition and perception and how they relate. Starting off with merely phenomenological descriptions of time (Sect. 8.3), that refers predominantly to James' work (which, in turn, had been explicitly influenced by Husserl's). It also makes reference to the work of Husserl, Merleau-Ponty and other 'real' phenomenologists. Section 8.4 stays within the realm of conceptual contemplation, but focuses on those thinkers that explicitly link mental time to physical processes, such as Kant and Piaget. Section 8.5 presents empirical approaches that rely in some form on verbal experiential reports, such as Núñez' anthropological work, Shanon's research on altered states of consciousness and Piaget's experiments in children's cognitive development. Section 8.6 presents evidence from cognitive neuroscience, the psychology of perception and psychophysics, which makes direct reference to physical and physiological processes which may play a role in the constitution of primitive time experience. An attempt to bring these diverse perspectives together is undertaken in Sect. 8.7.

Before starting this journey, some conceptual distinctions have to be made that recur across authors and disciplines, to prime the reader and ease the task of seeing the connections in this broad spectrum of work. Firstly, nearly all researchers that have seriously dealt with explaining temporal experience have remarked that *there is a primitive/intuitive temporal dimension inherent in our flow of consciousness and that it is different from our cognitive and symbolic conception of time*. However, there is a multitude of ideas about the exact nature of either and how levels of sophistication are structured and relate. Secondly, *a spatial metaphor of time* seems absolutely indispensable to any analysis of time and this link between space and time has frequently been made explicit. It seems that the question of how the conception of space and the conception of time relate is of crucial importance in an enactive approach to mind. Thirdly, a close look at the notions of *knowledge and time* reveals that they are intricately linked, in a story that includes also the concepts of *agency and possibility*. This last point is possibly the most obscure, tacit and least developed of the three.

The reader who expects a coherent theory of time and temporality will be disappointed. The picture that emerges is one of 'thought in progress'. Extensive scientific and conceptual work will be necessary to come up with a theory of time perception. All that this chapter does is to phrase questions, from which such an extensive endeavour can start. An attempt to hint at an answer to some of them is undertaken in chapter 11.

8.3 Phenomenology

The most fundamental observation on the phenomenology of time perception is that the "cognized present is no knife-edge, but a saddle-back, with a certain breadth of its own on which we sit perched, and from which we look in two directions into time" (James, 1890). Were our flow of experience but a chaining of punctual moments, as our Newtonian-Cartesian intuition has us believe, *our experience would change, but we could never experience any change*. The just-past is always still present, as is that which is about to come. This dynamic of 'retentions' and 'protentions' in our experience of the present has been analysed and described in detail by Husserl (in Steiner, 1997). Other thinkers mentioned in this context share and extend the observation that the present is 'specious' in this sense.

These extended chunks of present do not change continuously in our experience. They do not flow like a river, but instead switch abruptly, discretely, switching their overlapping yet different meaningful content. "The discreteness is, however, merely due to the fact that our successive acts of recognition or apperception of what it is are discrete. The sensa-

tion is as continuous as any sensation can be." (James, 1890). This observation rephrases Husserl's distinction between the bottom two (of three) layers of time phenomenology. (Varela, 1999) refers to these in an interpretation of Husserl's work as the subtle 'absolute flow of consciousness' (level one) and the immanent flow of meaningful moments (level two). So, from a continuous and changing flow of primitive sensation, we construct and chain moments of recognition that are *meaningful* in the most rudimentary form. These discrete and chained moments are not of arbitrary length. It is cognitively impossible to grasp and experience an extended time span as a single integrated percept. James observes that, in this point, there is an interesting qualitative difference between the phenomenology of time and that of space. Even though we can zoom in or zoom out of space according to need and experience an entire landscape as an integrated phenomenon, containing objects that are kilometres apart, as well as focus on microscopic events, blanking out the rest, this is not possible for time experience, which was a 'myopic' sense: "The durations we have practically most to deal with – minutes, hours, and days – have to be symbolically conceived, and constructed by mental addition, after the fashion of those extents of hundreds of miles and upward, which in the field of space are beyond the range of most men's practical interests altogether." (James, 1890). This additional layer of time phenomenology is the same as Husserl's third layer, which Varela calls the symbolic-narrative (third level). The distinction between these three layers is important when phrasing research questions concerned with mental time. As the following sections will show, these layers function and can be modulated more or less independently from each other. Therefore, it has to be made clear which of the layers is addressed and how. The naïve Cartesian illusion that time is one coherent variable in our mind, a central clock, already is challenged by this layered structure of temporal experience.

Only through the construction of the third symbolic level of time phenomenology, a fundamental and interesting issue enters the stage: the apparent paradox of experienced pastness. Supposedly, at any moment in time, only the present is real, not the past (nor the future). The memory of the past, and the anticipation of the future, are manifestations of the past and the future in the present, a kind of 'trace' as (Merleau-Ponty, 2002) calls it. The question then, as (James, 1890) points out, is: "But how do these things get their pastness?" A memory cannot *be* the past because the past does not exist anymore. If a memory, instead, was a retrieval of the original train of discrete chunks of subjective experience, it would feel as present as it did when it was lived. The memory wears the sign of the past-made-present, and what this pastness consists of is a mystery. Again, this problem is not obvious

from a representationalist perspective, where past can be represented by tagging bits of information temporally, which, arguably, does not do justice to the just described experience of pastness. As (Merleau-Ponty, 2002) points out, it is our capacity to remember and expect and thereby experientially change the direction of the flow of consciousness, which allows us to think of time as time. Paradoxically, through this conceptualisation of time, it ceases to be temporal: "It is spatial, since its moments are spread out before thought" (Merleau-Ponty, 2002, p. 482).

These two observations – the three layer structure of temporal experience and the spatial metaphor of time in the symbolic layer of time perception, which makes mental travel to memories and future anticipations possible, are the most crucial insights from phenomenology for the present purposes. The phenomenologists have observed many more. Among them, there is the distinction between future and past; rhythmicity in the primitive flow of time; meaning, intentionality and objects of time. Valuable though these contributions are, Varela's critical remark that "[we] still lack a phenomenology of internal time consciousness where the reductive gestures and the textural base of the experience figure explicitly and fully" (Varela, 1999) is adequate. The expert phenomenologist reader is asked to bear with an impatient scientist author that has dealt with the material only superficially. I dare to argue that for the scientist interested in a particular temporal phenomenon, the phenomenologist's viewpoints are too abstract, too general, treating temporal experience across time scales, modalities, tasks, behaviours, levels of abstraction and, therefore, deal with a mental time that is disembodied and detached from physics. As (Varela, 1999) points out, Husserl's prime example of listening to a melody is developed without any mention of whether this melody is familiar, of which kind of emotional effect it has, where it is heard, in a large room, a small room, an open space, sitting, standing up, *etc.* All these factors clearly hold the potential to impact on the temporal experience of the piece.

Phenomenology as a discipline never aspired to be scientific, never aspired naturalisation. In a scientific endeavour to explain time perception and experienced temporality, some of the phenomenologist's observations are invaluable in realising the poverty and inadequacy of a vulgar Cartesian intuition about mental time. Also, they provide the vocabulary to name distinctions between different aspects and levels of mental time. Phenomenology does have its niche in the interdisciplinary study of mind, with possibilities and limitations. But it it is not all there is to temporal experience, material and physical processes are equally important. As this chapter proceeds, work presented becomes increasingly concrete and scientific, bringing in that other side.

8.4 The Construction of Time

The preceding summary of our experience of time hardly made reference to the idea of a simple four dimensional physical Newtonian-Cartesian time-space. However, even if the phenomenology of time perception does not follow the laws of Newtonian physics, the world does – at least, most of the time. Newtonian time itself and its geometrical and mathematical properties, applied to the real world around us, are very powerful in explaining, understanding and predicting it. Drawing on thoughts by Kant and Piaget, this section, which is still conceptual, tries to explain the link between mental time and physical time and the role of the spatial metaphor of the time as arrow.

To start with the discussion of time in (Kant, 1974)'s *Critique of Pure Reason*, in the transcendental aesthetics, Kant assigns a special status to time and space, calling them the *a priori* formal conditions of *Anschauung* (perception). In an at least proto-constructivist fashion, Kant stresses again and again that time and space are not objectively real, in the sense that they are not observer-independent properties of the *Welt an sich* (world in itself). Time is nothing but the form of our inner senses, our experience of our changing self.[1] As such, time has 'empirical reality', 'subjective reality' for Kant, and it makes the perception/imagination of self possible, the reflexive subjective experience of subjectivity itself, as an object.[2] This description of registered change in inner subjective state resonates strongly with the phenomenologists' identification of the primitive and immanent levels of time experience. However, Husserl observes that reflexive experience of change as change is, in itself, atemporal and, therefore, not part of the immanent flow of time (in Steiner, 1997, p. 327). Also, Kant's idea that temporality of direct subjective experience is necessary for the experience of self resonates with (Heidegger, 1963)'s idea that temporality is necessary for concernful existence.

For Kant, the irreducible reality of subjective time *a priori* as a changing flow does not contain or imply the categorical and relational properties that characterise our grown adult conception of time. Time, at this level, is not a property of the experienced exterior. It is not a property of gestalt, location, *etc.*, but instead it determines the relation of experience in our inner state only. This lack of a gestalt of our inner state is compensated for by the construction of a metaphor such as time as an arrow that goes to infinity, chaining

[1] "Die Zeit ist nichts anders, als die Form des innern Sinnes, d.i. des Anschauens unserer selbs und unsers innern Zustandes" (Kant, 1974, p. 80f).

[2] "[Die Zeit] hat also subjektive Realität in Ansehung der innern Erfahrung, h. i. ich habe wirklich die Vorstellung von der Zeit und *meinen* Bestimmungen in ihr. Sie ist also wirklich nicht als Objekt sondern als die Vorstellungsart meiner selbst als Objekts anzusehen" (Kant, 1974, p. 83).

'manifolds'.[3] Only thereby, time is projected into the world and becomes a property not just of self and subjective experience, but of the objects around us. Crucially, Kant sees the construction of time as an object, as a dimension of the objective world as a strictly logical process: apart from 'empirical reality', time possesses 'transcendental ideality'. He supports his claim with the fact that mathematical and logical laws hold for time and space, which are strictly intersubjectively valid and thus not really *a posteriori*, but, what he calls 'synthetic judgements *a priori*'.[4]

Further on in the Critique of Pure Reason, Kant also hints towards some of the relations between time and space and their geometrical properties that he thinks form the basis for the synthetic judgements *a priori* that constitute transcendentally ideal concepts of time and space. In the analogies of experience (transcendental analytics), Kant explains how the concepts of constancy, succession and simultaneity (*Beharrlichkeit, Folge und Zugleichsein*) result from connecting distinct experiences in subjective time. For instance, he points out that simultaneity in time is given if the order in which objects are perceived is arbitrary or reversible, for if the order in which they were experienced was fixed, they would be successive and not simultaneous.[5] At the same time, he asserts that the rules of constancy, succession and simultaneity are *a priori* and necessary for experience to happen at all.[6] This identification of reversibility as characteristic to distinguish space and time is essential to fully understand what the spatial metaphor of time as an arrow really means. (Merleau-Ponty, 2002) remarked that it is the possibility to anticipate and remember, which allows us to travel freely in both directions on time's arrow and to thus spatialise and objectify time, to overcome the 'myopic' character of time that (James, 1890) observed (i.e., that only moments of short duration can be directly experienced as coherent percept, cf. Sect. 8.3). In Kant's view of space and time, only when things other than oneself move

[3]"Denn die Zeit kann keine Bestimmung äußerer Erscheinungen sein; sie gehöret weder zu einer Gestalt, oder Lage, *etc.*, dagegen bestimmt sie das Verhältnis der Vorstellung in unserm innern Zustande. Und, eben weil diese innre Anschauung keine Gestalt gibt, suchen wir auch diesen Mangel durch Analogien zu ersetzen, und stellen die Zeitfolge durch eine ins Unendliche fortgehende Linie vor, in welcher das Mannigfaltige eine Reihe ausmacht" (Kant, 1974, p. 80f).

[4]'Synthetic judgments *a priori*' can be roughly understood as judgments that are independent of experience, necessary and universal, without being directly and obviously tautological.

[5]"und darum weil die Wahrnehmungen dieser Gegenstände einander wechselseitig folgen können, sage ich, sie existieren zugleich" (Kant, 1974, p. 242) or later "Woran erkennt man aber: daß sie in einer und derselben Zeit sind? Wenn die Ordnung in der Synthesis der Apprehensionen dieses Mannigfaltigen gleichgültig ist, d.i. von A, durch B, C, E, auf E, oder auch umgekehrt von E zu A gehen kann. Denn, wäre sie in der Zeit nach einander (in der Ordnung, die von A anhebt, und in E endigt), so ist es unmöglich, die Apprehension in der Wahrnehmung von E anzuheben, um rückwärts zu A fortzugehen, weil A zur vergangenen Zeit gehört, und also kein Gegenstand der Apprehension mehr sein kann" (Kant, 1974, p. 243).

[6]"Daher werden drei Regeln aller Zeitverhältnisse der Erscheinungen, wornach jeder ihr Dasein in aller Erfahrung vorangehen, und diese allererst möglich machen" (Kant, 1974, p. 217).

around, the conceptions of time and space get in contact and in conflict and require the construction of relations such as movement, velocity/speed, simultaneity and causality in order to distinguish them and disambiguate, which results from the processes and experiences described by Kant in his analogies of experience.

Kant is right in pointing out that our temporal and spatial experience, in its most rudimentary form, cannot be stripped off our experience and imagined away in the way that other aesthetic qualities, such as hardness or colour can be stripped off. He is also right in pointing out that temporal experience is tied even closer into experience than spatial experience: space is a property of the exterior, but subjectivity is experienced non-spatially, yet temporally. This is why Cartesian fantasies of brains in vats or *The Matrix* are happy to place the *res cogitans* in an illusory fantasy world, hiding the 'real' world as regards its spatial surroundings; the time line, however, in which the deception takes place is maintained, because it is the *a priori* form of the subject. If temporal coherence is lost, self is lost. However, what is debatable is the privileged character that Kant assigns to the constructed and projected 'transcendentally ideal' time (and space): the elaborate observations by the phenomenologists, as well as the empirical data presented in the following Sects. 8.5 and 8.6 show that there are many and variable factors contributing to the conception of time (culture, sensorimotor dynamics, development, intact functioning of the brain, *etc.*). The logical properties of time are, to a degree, contingent on these factors. The processes of construction Kant describes, which underlie the synthetic judgements *a priori* that lead to the transcendentally ideal notion of time, can they not be interrupted? The empirical study of the construction of time shows that unusual circumstances can lead to experiences of time, even on the abstract symbolic level, that violate logical constraints.

Piaget's views, expressed nearly two centuries later, are in many ways akin to Kant's. He distinguishes *intuitive time* and *operational time*. Intuitive time, for Piaget, is "limited to successions and durations given by direct perception." (Piaget, 1969, p. 2), which seems to broadly correspond to what Kant describes as *a priori* "Ansehung der innern Erfahrung' (observation of inner experience) (Kant, 1974, p. 83), whilst operational time "is the operational co-ordination of the motions themselves" (Piaget, 1969, p. 3) and builds on the active, successive construction of the relations between simultaneity, succession and duration.

Both, Kant and Piaget, distinguish two and only two modi of time, the primitive and the constructed (intuitive vs. operational in Piaget, empirical vs. transcendental in Kant). This contrasts with the more fine-grained view of the phenomenologists, who pick up on the even

more subtle distinction between the primitive and the immanent flow of time. Where both Piaget and Kant go beyond the phenomenologists, however, is in attempting to explain how the primitive and the constructed level of time experience relate, how one is constructed on top of the other, making reference to the body, to action and to the external world.

Piaget hypothesises that space and geometrical relationships have to be constructed prior to the development of a more sophisticated concept of time: "It is only once [space] has already been constructed, that time can be conceived as an independent system" (Piaget, 1969, p. 2). Piaget seems to assign ontological priority to the conception of space over the concept of time ("In the course of its construction, time remains a simple dimension inseparable from space" (Piaget, 1969, p. 2), whereas "space suffices for the co-ordination of simultaneous positions" (Piaget, 1969, p. 2) and "space is above all a system of concrete operations, inseparable from the experiences to which they give rise and which they transform" (Piaget, 1969, p. 1). Time, then, on the basis of a pre-existing conception of objective space, defines the relation of change in space: "as soon as displacements are introduced they bring in their train distinct and therefore successive spatial states whose co-ordination is nothing other than time itself." (Piaget, 1969, p. 2). The important addition Piaget thus makes is that the construction of time is a stage-wise developmental process that relies on a history of sensorimotor interactions with the world, not a disembodied process of mathematical deduction. The construction of time comes after the *knowing how* to act in a coordinated manner in the real world.

Both Piaget and Kant rush over a number of steps along the way of how temporal experience is constructed. While Kant has focused too much on the logical-mathematical side of space and time, neglecting the real-world processes underlying it, Piaget fails to address the complexity and reciprocity of the steps that lead towards the construction of sophisticated concepts of both time and space, prioritising space. By contrast, Kant describes how, from primitive temporal experience (i.e., the experience of change) and primitive spatial experience (i.e., the experience of inside/outside), more operational conceptions of both time and space are bootstrapped, in a process of mutual co-construction. Combining the mutuality and graduality of Kant's account and the embodied developmental perspective of Piaget, a good starting point to understand the evidence presented below is gained. Comparing species, cultures, developmental stages, pathological experience of time, *etc.*, it will be possible to understand the mutual dependencies between levels of temporality and spatiality and how they rely on one another, at least intuitively.

The enactive approach sees the living organism and its evolutionary and developmental history as the physical basis of mind. (Stewart, forthcoming) points out that "developmental systems have to make do with piecemeal step-by-step tinkering and cannot be redesigned from scratch". He argues that the changing constraints that evolution puts on development (and *vice versa*) require explanation. In (Barandiaran *et al.*, 2009), we sketch out a hierarchy of spatio-temporal complexity of behaviour and cognition, picking target organisms for crucial transitions in phylogenetic evolutionary history. The hierarchy is based on the constraints and possibilities afforded by the organisation of the nervous and the sensorimotor system. For instance, in the bacterium *E Coli*, no reversibility of action is possible. Picking up Kant's point that reversibility of experience is what distinguishes spatial from temporal sensation (earlier this section), we can infer that for *E Coli*, time and space do not exist as distinct factors in its behavioural domain. Following Kant's reasoning, the only distinction between time and space possible from the perspective of *E Coli* is the most fundamental one (change, inside-outside). The organism can never be clear if a change in sensation is caused by the bacterium itself or by an outside force, if this change cannot be reversed at will.[7] Other organisms (e.g., some insects) may have access to topological order, i.e., they are able to reproduce a sequence of sensory states, but not to the metric properties of space and time. Yet other organisms (some vertebrates) perceive and exploit the metric properties, but are unable to spatialise and symbolise time as arrow, which, probably, is an exclusively human skill that requires symbolic cognitive capacities and enables us to mentally travel back and forth in time at will. The sophistication of space and time, that starts from the primitive *a priori* forms of perception and reaches the pinnacle in human reflexive and symbolic abstracted space-time, is a process of gradual co-construction in both development and evolution.

8.5 Findings on Cognitive Concepts of Time

After an extensive conceptual analysis, the focus now shifts over to empirical research. This section presents work from Piaget's developmental psychology of the conception of time, from linguistic/anthropological work on the conceptual metaphors of 'time as space'

[7]This is besides the point of whether we want to talk about bacteria intelligence or bacteria cognition at all, given that what bacteria do does not involve any reflexive self-awareness or the like – as outlined in chapter 2, the enactive approach sees the mind and life on a continuum, and single celled bacterial life is at the far bottom end of autonomous living organisation. Recall (Jonas, 1966)'s and (Weber, 2003)'s argument that our own experience as living organisms helps us to understand meaning and value in other organisms, even if they are incapable of reflexive self-awareness.

and from Shanon's work on temporal experience under the influence of the psychedelic Ayahuasca potion.[8] These examples serve to ground the preceding theoretical contemplations in empirical research. However, they remain on the conscious and symbolic level, as the data is generated from verbal reports. The more subtle and basic aspects of time experience and time perception that can be disrupted by simple physical manipulations on the macro level are treated in the next section.

Out of the experiments about the construction of the child's conception of time, those about succession and simultaneity in physical time are particularly revealing. The experimental paradigm used in both cases is the simultaneous motion of two figures at different velocities, either stopping simultaneously or successively. Due to the difference in velocity, this scenario can lead to different spatial configurations once both figures have stopped, where the spatial order does not necessarily reflect the temporal order in which the figures stopped. This contingency between temporal order and spatial organisation leads to characteristic confusions in children at certain stages of development, when asked about spatial displacements, temporal orderings and how these two relate. In what Piaget calls 'stage I' "successions and durations remain undifferentiated from distances [...] and differences in speed are thought to preclude synchronous processes and lead to confused estimates of duration." (Piaget, 1969, p. 85). The following excerpt from an experiment with a four year old child will help to gain an impression of the kind of errors children make at stage I. The child is presented with a situation in which a yellow figure is made to stop earlier than a blue figure, with the blue figure still stopping spatially less far than the yellow figure (child's responses in italics):

> "Did they stop at the same time? *No.* Which one stopped first? *The blue one.* Which moved longer? *The yellow one.* [...] But which one stopped first? *The yellow one. No, it was the blue one, the yellow one went on longer.* Let's do it again. (The race is re-run.) *The yellow one stopped first, the blue one was still moving, so the yellow one went on longer.* But did one stop before the other? *The blue one*" (Piaget, 1969, p. 86).

Children at this developmental stage are incapable of detecting or correcting their confusion of temporal and spatial differences, and do not seem to be bothered by the logical contradictions either. Their symbolic temporal experience is incomplete.

To pre-empt objections that there could be just a linguistic confusion about the spatial metaphor in time, Piaget mentions that he and his colleagues also asked the children further, less ambiguous and more intuitive questions, coming to the same results of the children confusing temporal and spatial order. Interestingly, the systematic mistakes disappear at the

[8] Ayahuasca is a psychedelic herbal brew used in a ritual context in many indigenous cultures in Amazonia.

same developmental stage, if the figures are made to move into opposing directions such that the spatial and the temporal dimension are more clearly separate in the scene. Piaget then describes how children pass through later developmental stages in which they would still make misjudgements of the described type, but be able to correct them when being pointed, in dialogue, to the logical contradictions in their report, before finally arriving at a 'transcendentally ideal', in Kant's sense, conception of time. These stages and the mistakes they entail are the same if the child itself is made to run against the experimenter, as an agent of the scene, rather than as a mere observer.

From these results, it is clear that what we experience as different layers of time perception phenomenologically also corresponds to different levels of behaviour and physical processes. A child that is developmentally advanced enough to lead these kinds of interviews has an intact capacity to register change and the immanent chaining of meaningful moments. It is also sufficiently symbolically developed to use and understand language with compositional structure. However, it lacks the maturity to experience space and time as transcendentally ideal. The children have clearly learned to name temporal properties of objects in the world and describe order relations, but these concepts remain fuzzy and intermingled with those of space. Even though children are perfectly able to coordinate their actions in the real world, there is no clear distinction for them between those changes in a previously registered flow of consciousness that are really reversible (spatial) and those that are only mentally reversible (temporal). The symbolic layer of time experience, with its mathematical and logical properties, is not yet fully developed.

Another interesting turn on the story of how time and space relate in our symbolic conception comes from the use of spatial language as a metaphor for time in an across-culture comparison. (Lakoff and Johnson, 2003) report that spatial language is used metaphorically to talk about time in nearly all languages: usually, the future is seen as being in front, whereas the past is conceived of as behind, in expressions such as: 'the time will come when . . . ' or 'in the weeks ahead of us' (Lakoff and Johnson, 2003, p. 42). Lakoff and Johnson's work shows that such 'conceptual metaphors' are used systematically and consistently across cultures, and they interpret this systematic occurrence as a sign of an inherent semantic link between the concepts, not just as a verbal shorthand. All inter- and intra-cultural inconsistencies in the metaphor they encountered could be assimilated into a universal story by including an aspect of agency in the metaphor, i.e., to conceive time passing as motion, which can be instantiated either as time being the moving object or us as moving in time (Lakoff and Johnson, 2003, p. 41-45).

A very interesting deviation from the described conceptual metaphor has been described by (Núñez and Sweetser, 2006) to occur in the Aymara language spoken by indigenous people in certain parts of the Andes. In a crude simplification of Núñez and Sweetser's findings, the Aymara language is to date the only reported language in which the *time is space* metaphor is directionally inverted (i.e., the past is conceptualised as in front of the speaker and the future as behind). Most intriguingly, Núñez and Sweetser have also found that the accompanying gestures of the Aymara speakers comply with this use of language (e.g., an Aymara speaker would point forwards when using the Aymara word for forward and when referring to the past) and that this seeming spatial inversion of temporal gestures is preserved when native Aymara speakers speak the Andes dialect of Spanish. They partially adapt the Spanish grammar to match the conceptual metaphor.

The usual *time is space* metaphor, as described by Lakoff and Johnson, appears to naturally link to the processes of spatialisation and temporalisation through embodied experience, as we have analysed them so far. The Aymaran people's use of the metaphor in the inverse direction is counter-intuitive and hard to conceive. At first glance, it is also hard to even integrate it into the story of embodied construction of time and space through our development and from our embodied interactions with the environment.

Núñez and Sweetser have a very interesting explanation for this exceptional use of the *time is space* conceptual metaphor in language and gesture that reconciles it with Lakoff and Johnson's ideas: they observe that Aymaran spatial metaphors for time never involve any self-motion. Whilst the *time is space* metaphor in most languages involves movement along a path or a river (either by the subject or by an agent-time itself), leaving behind visited (past and known) stations and discovering the new behind the next corner, the spatial metaphor of time for the Aymaran people is a static one. In this static spatial metaphor of time, the space in front of the subject is visible, which means it is known. The space behind the speaker, on the other hand, is unknown. Things occurring behind the speakers back can go undetected and surprise the speaker, just as the future has potential for surprises. Therefore, a conceptual metaphor of 'seeing is knowing', in combination with the fact that what is seen is in the front, overwrites the metaphor of time as motion along a path. Interestingly, the authors also point out that the Aymaran culture assigns importance to personal testimony and that they discredit talking about the speculative/unknown, which is marked by a reluctance to talk about the future in general.

These interesting findings show two things: firstly, it is impossible to talk about *the* spatial metaphor of time. There are variable structural similarities at different levels of meaning

and interpretation between the two. The common *time is space* metaphor actually has to be elaborated to be a *time is motion along a path* metaphor, whilst the metaphor of *the past is known and the future is unknown*, in combination with a *seeing is knowing* metaphor, can lead to an alternative interpretation of how location corresponds to a conceptualisation of time, a more passive and backward looking one. Secondly, it emphasises an aspect of temporal conception that has not yet been treated in depth, i.e., the criteria for distinguishing the past from the future in our present experience of time (i.e., distinguishing memories from anticipations). What characterises the past is that it does not change, neither by its own accounts, nor by an agent's own influence. Only by taking agency out of the picture, the Aymaran people make it possible to conceive of what is in front as the past. The future, however, is open, it can change, and it can be changed through intentional action, and in this sense, it is not yet real, not part of this present world. This relates to Merleau-Ponty's remark that the future seems to only exist "by analogy" (Merleau-Ponty, 2002, p. 481), by a guess that this moment will pass and turn into past like all the ones before it, being replaced by another one that is yet unknown. The present, then, logically, is what is jammed in between the two: it comprises all that is in its making: there is no more the possibility to take influence on it, but what exactly it entails has still to be verified by experience (this view is elaborated in chapter 12).

As a third lesson from Núñez' and Sweetser's results, we should be gently reminded that our conception of time is not just contingent on developmental phase, that it is not only phenomenologically more complex and multi-faceted than Kant seems to acknowledge, but that, in its complex structure, temporal experience will also have a strong cultural component. (Evans, 2004) lists nine different (yet related) meanings of the word/concept 'time' in English, four of which he claims to be "secondary lexical concepts", i.e., they are cultural constructs, that are not rooted in universals of human experience. The extent to which such culturally contingent conceptions of symbolic time influence time experience and temporal behaviour is not clear.

A complementary line of research is (Shanon, 2001)'s research on temporal perception under the influence of the psychedelic potion Ayahuasca. Shanon's research aims at highlighting those aspects of our conscious experience of the world that we take for granted, because they are always there. These aspects can be studied by investigating how altered, abnormal states of consciousness lead to distortion and break-down of what we think of as natural and normal and which, as a consequence, brings what is normal to our attention. As concerns the experience of time, there are a number of alterations observed in both, natives

of the Amazon forest (from those cultures in which Ayahuasca is traditionally used in a ritual context) and naïve European and North American participants. Shanon describes some comparably gentle alterations of temporal experience (change of rate of experienced time, change in perceived distance to past or future events, relocation of 'present' in the illusion to witness/live past events). These examples are interesting, because they point out those factors in our cognitive time that can be topologically distorted whilst leaving the general logic and order of time intact.

More related to the previous analysis of the nature of the concept of time and space are experiences that induce the feeling of timelessness, eternity and the confusion of perception, memory and anticipation. There is a more general effect of Ayahuasca that the real and the unreal get blurred, and confusion of memory and anticipation can be seen as the time-specific experience of this blur. In relation to the preceding analysis that the distinction between the actual and the possible is essential for the experience of pastness, presentness and futureness and temporality as different from spatiality, this blur induced by Ayahuasca is important. In the limit case, the blurring of these boundaries results in states of consciousness which can be seen as a *completion of the time is space metaphor*. As Shanon puts it, "the temporal may, in a fashion, be reduced to the spatial" (Shanon, 2001, p. 47). To quote a report from such a vision:

> "In front of me I saw the space of all possibilities. The possibilities were there like objects in physical space. Choosing, I realized, is tantamount to the taking of a particular path in this space. It does not, however, consist in the generation of intrinsically new states of affairs" (Shanon, 2001, p. 47).

Shanon reports that such an 'out of time' experience is frequently accompanied with the feeling of omniscience, stripping the future off its speculative and open character. Resonating with (Heidegger, 1963)'s ideas of temporality being the basis for concernful existence, the stepping out of time coincides with a loss of concern, temporality becomes irrelevant: a side effect "is the taking of things less seriously and with more tolerance, forgiveness and also a (benevolent) sense of humour" (Shanon, 2001). This experience of eternity and the complete spatialisation of time is a perfect instantiation of what Husserl describes as God's consciousness, a "limit-notion of temporal analysis: God's infinite consciousness contains all times at once. This infinite consciousness is a-temporal" (in Steiner, 1997, p. 40).[9] And just as Husserl realises that "even a divine consciousness would have to progress tempo-

[9]My translation: "...Limes-Begriff der Zeitanalysen: 'Gottes unendliches Bewußtsein umfaßt alle Zeit 'zugleich'. Dieses Bewußtsein ist unzeitlich." (in Steiner, 1997, p. 40).

rally" (in Steiner, 1997, p. 40)[10], the description of experienced eternity under Ayahuasca is constrained in the same way: "Further, it should be noted that while traveling in the space of possibilities takes time, the possibilities themselves are there, given in an ever-present atemporal space." (Shanon, 2001, p. 47). In this experience of being outside time, all external agency and forces disappear, and thereby all uncertainty. Time is spatialised and loses its meaning. However, even as the constructed notion of time in many of its dimensions collapses, the flow of time *a priori*, the primitive and the immanent level of temporal experience, persist.

The reports from various scientific approaches to temporal experiences of the constructed type all seem alien to the healthy sober adult westerner. They help to illustrate what constitutes mental time, what regularities govern it and how time relates to space, as well as to knowledge, subjectivity, possibility and concern. They also illustrate which of the qualities of everyday temporal experience are contingent and can disappear, even if this disappearance is at the cost of logical consistency, self-concern or the concept of time itself. Thereby, they give us an impression of what is left: contemplating all these examples, we may get a better idea of the absolute flow of consciousness that survives all these bizarre transformations and seems indeed necessarily linked to all human experience. This is something that dedicated central clock approaches will be unable to account for. The symbolic level of time experience is constructed from and constrained by this primitive and the immanent flow of time. In telling a complete story of time cognition, it will not only be necessary to investigate the differences between these levels and how they can be altered, but also how they relate.

8.6 The Brain, the World and Time Perception

The previous section has given an impression of how temporal experience on the symbolic level is variable. Returning to the starting point, the intuitive Cartesian idea of an internal clock, one coherent, abstract and logical representation of time in our mind, the analysis so far has helped to separate some aspects from what we mistakenly and intuitively conceive of as a unified irreducible temporal experience. Coming back to the phenomenological analysis (Sect. 8.3), however, there was a further distinction between the immanent flow and the primitive flow, both of which have to be distinguished from the symbolic level of time experience. In order to empirically investigate the link between these more primitive

[10]My translation: "Selbst ein göttliches Bewußtsein müßte notwendig zeitförmig verlaufen" (in Steiner, 1997, p. 40).

levels of time consciousness, i.e., how the immanent flow of temporal object-events is constructed from the *a priori* primitive flow of consciousness, the level of intervention has to be scaled down accordingly. Human consciousness and socio-linguistic awareness can be surpassed to a certain degree if you mess with certain aspects of physics directly. As it is the case for the entire chapter, this section is not an exhaustive literature review, but just a presentation of few selected examples, to make a point about how the immanent flow of temporal experience can be modulated in controlled ways through physical manipulations of the environment.

In *The specious present: a Neurophenomenology of time consciousness'* (Varela, 1999) sets out to link the three levels of temporal experience identified by Husserl to dynamical properties of the human brain. From the phenomenological analysis, we recapitulate *that* there are three levels of time experience (see Sect. 8.3):

- the primitive and continuous flow of sensations
- the discrete chaining of meaningful 'nows' as the immanent flow of experience
- the symbolically constructed narrative time level that exceeds in duration our experience of the present.

However, there is no reason given yet as to why that should be the case – why not just one level? Why not infinitely many?

Varela attempts to fill this gap, starting off with quantification of the temporal duration of changes in each level. The continuous flow of sensations is identified with the duration of several tens of milliseconds. This is the time scale in which we humans can make minimal perceptual discriminations about temporal order (even if exact resolution varies across modalities), the time of micro-saccades and the time scale of inter-neural events (action potentials). The brain acts as a bottleneck there, the physiological limits of our body imply that changes that happen faster than the fastest meaningful processes in the brain and the body simply do not exist as part of our perceptual world. On top of this time scale of the continuous flow of sensation, the second level of time consciousness, Husserl's 'immanent flow of time', is constructed. According to Varela, the time scale of this level is in the scale of around 1 s (the same ballpark as (James, 1890)'s 'specious present' of 3 s), a time scale which corresponds to the time necessary to integrate several of the atomic sensations identified as the units of the primitive flow of consciousness, and the time that assemblies of neurons need to integrate and coordinate their activities across the cortex. This is the level of recognised change, the level in which experience becomes subjective and present, in a very rudimentary form meaningful. Varela calls this the scale of 'temporal object-events'

to translate Husserl's *Zeitobjekte* (Varela, 1999). Perception of present at this immanent level has meaningful contents, even if those are not objects or events in a transcendental, abstract, reflexive sense. This immanent level of time experience is what Varela focuses on in his analysis, i.e., its construction and delimitation from the first 'primitive flow' level. The third level is the level of "descriptive-narrative assessments" (Varela, 1999). Varela links it to our linguistic capacities and calls it the level of "continuity of a self that breaks down under intoxication or in pathologies such as schizophrenia or Korsakoff's syndrome" (Varela, 1999). This assessment is in line with the analysis given in the previous section, about the kind of perturbations that can take influence on this level of time experience that is reflected in language and conscious thought.

Varela thus draws a picture in which qualitatively different biological/physiological processes on different spatio-temporal scales (local neural activity, integrated neural activity across populations and socio-linguistic behaviour/long term neural learning) recursively build up the three layers of temporal experience. This is why there are three layers, not one, not more. In his neurophenomenological account, Varela focuses his account on the distinction between the first and the second layer, because it is still very difficult to directly link neuro-physiological processes to the macro-phenomena that shape the third symbolic level (development, culture, personal history, ...). The kind of approaches presented in the previous section appear more promising and insightful at this stage.[11]

One important merit of Varela's neurophenomenological approach is that, despite the strict delimitation of the three layers in qualitative terms, he resists the temptation to identify a 'magical number' of neural meaning, a unit of the 'neural currency', like (James, 1890)' 3 s or (Libet, 2004)'s 500 ms. Naturally, processes on different time scales and of different exact duration can be equally meaningful to a living organism. Varela's story naturally clusters such events into the three levels, by their capacity to influence the physiological processes that underlie the three layers. This also implies that some events, whose duration is at the transition between these time scales, can affect both of the neighbouring time scales at whose transition they are to be localised. This observation becomes relevant again in in chapter 11, when the results from the combined experimental and modelling study on delay adaptation and recalibration of perceived simultaneity are discussed.

A particularly fundamental line of evidence on the physiological basis of temporal experience is (Libet, 2004)'s work on neuro-sensory and neuro-motor latencies and how they

[11](Rosenfield, 1988)'s 'The invention of memory' should be mentioned here as a noteworthy exception; the book presents some very interesting constructivist ideas on memory as traces that is based on neuroscientific and neuropsychological results.

lead to lags between a neuro-cortical event and a corresponding experiential correlate. By means of direct cortical stimulation of variable length in neurosurgical patients in the late 1950s, Libet found that a cortically administered stimulus was only registered and reported if it persisted for at least 500 ms. Libet found this to be "surprisingly long for a neural function" (Libet, 2004, p. 39). This led him to the conclusion that "*awareness of our sensory world is substantially delayed* from its actual occurrence" and that we are thus "always a little bit late" (Libet, 2004, p. 70). Libet found a delay of nearly identical length to pass between the neural potential recorded from the pre-motor cortex ('Readiness Potential'), that marks a 'point of no return' in motor decision making, and the awareness of having committed to this decision (time of awareness of the decision was measured by the subjects' reference to a clock). This temporal discrepancy between the externally measured neural event of decision making and the conscious correlate challenges our intuition that experienced time is coordinated and synchronised with 'objective physical time' as it is measured by a clock.

Libet's experiments show another example of how, by means of measuring judgements and correlated neuro-physiological processes, links can be established between the physical and the experiential domain. As already remarked in chapter 3 (Sect. 3.5), Libet's approach implements what (Fechner, 1966) envisioned as 'internal psychophysics', but was not possible at his time because of technical limitations.

Even prior to Libet, Grey Walter conducted a very related study (unfortunately, this study has not been published in technical detail: (Dennett and Kinsbourne, 1992) refer to a talk given to the Ostler society in Oxford University in 1963). The way the study is described in (Dennett and Kinsbourne, 1992) is the following: Walter instructed neurosurgical patients to press a dummy button. He told them the button press would trigger a flip-over in projection slides. He recorded a signal that preceded the actual press of the button (he calls it contingent negative variation, CNV, but it is analogous to Libet's Readiness Potential) with electrodes directly from the motor cortex of the patients. He then used this signal in real-time to trigger the change in projection slides even before the button press was performed. Thereby, he closed the sensorimotor loop on the temporal effects that Libet reports. In Libet's version of the experiment, the astonishment is on the side of the observer, the scientist, who registers an inconsistency between her measurement of neural activity and the subject's report of registering the decision. Walter's experiment introduces this discrepancy into the experiential world of the subject himself, by including the measurement in a sensorimotor context. This experiment can be seen as an early predecessor of

real-time brain computer interfacing. The literature (e.g., Dennett and Kinsbourne, 1992) reports that subjects did not take credit for this action of flipping over the slides when activity in the pre-motor cortex was measured. They expected a second change in slides to occur as a result of their action, and they felt the decision they were about to make had been pre-empted – despite the fact that, on a neural level, the decision had been already made by that time. The artificial shortening of neuro-motor latencies led to the break-down of perceptual integration between the intended action and the observed effect. The logically and temporally impossible reversal of temporal order of cause and effect destroys the experience of ownership of the action. By bringing the discrepancy between mental time and 'real' objective time to the subject's own attention, it ceases to be a concern only of the observing scientist, it becomes a concern to the subject himself, with very interesting consequences to the experience of the event.

As concerns the more traditional discipline of 'external psychophysics', there is also a vast corpus of work on time perception. Besides the fundamental work on duration judgements and temporal order judgements that provides us with an idea of the temporal sensitivity/granularity of our sensory modalities, there are a number of perceptual illusions that are very telling about the processes and factors that underlie the integration of the immanent flow of temporal object-events. For instance in 'backward masking' (or 'retroactive masking'), a stimulus (peripheral (e.g., Herzog, 2007) or cerebral (e.g., Libet, 2004)) is administered to an experimental participant, which suppresses the awareness of a previously administered stimulus. The interesting thing here is that one event can suppress the perceptual experience of another one that has already passed – an apparent violation of the rule that the effect has to come before the cause. Similarly, in apparent motion (also called the 'psi-phenomenon'; e.g., Gepshtein and Kubovy, 2007), two discrete subsequent and displaced presentations of visual stimuli are perceived as a continuous motion from the location of the first to the second (which is the reason why we can experience a film as continually moving pictures, rather than as a discrete chain, which it 'really' is). Therefore, experienced motion is contingent on the presentation of the motion endpoint, even though, experientially, perception of motion along the path appears to precede the perception of the motion endpoint. Another interesting effect is the so-called flash-lag-effect (FLE): if subjects are presented with a moving bar, half of which is constantly illuminated and half of which is flashing, the flashing part of a moving bar appears to lag behind a constantly illuminated part, with the spatial distance being a function of the velocity of the bar (cf. Nijhawan, 1994).

There is an abundance of such findings about nonlinearities in the experience of time on the immanent level of time perception: distortion of temporal order or duration judgements have been observed in relation to factors as different as saccadic eye-movements (e.g., Morrone *et al.*, 2005; Yarrow *et al.*, 2001) and repetition of stimulus (e.g., Pariyadath and Eagleman, 2007). Overviews are given, e.g., in (Ivry and Schlerf, 2008; Eagleman *et al.*, 2005). What does this mean for time cognition, time perception and temporally co-ordinated behaviour?

From a naïve representationalist stance, such nonlinearities pose mysteries and logical problems. As outlined above, such a stance conceives of mental time as a centrally represented quantity that relies on linear processing and tagging of sensory inputs and aims at internally representing physical time with the highest possible accuracy. Any inaccuracy is expected to lead to consequent behavioural inaccuracies and a break-down of behavioural coordination, and the fact that this is not always the case leads to surprise. One example for such an objectivist fallacy is Libet's interpretation of his own observations: "so we have a strange paradox: neural activity requirements in the brain indicate that the experience or awareness of a skin stimulus cannot appear until after some 500 ms, yet, subjectively, we believe it was experienced without such a delay" (Libet, 2004, p. 72). In order to resolve this paradox, Libet proposes mechanisms that backdate experience to the time of their 'real occurrence'.

In a similarly representationalist spirit, Nijhawan explains the FLE as the result of a neural delay compensation mechanism that infers an object's 'real' position such that "the perceived location [...] is closer to the object's physical location than might be expected from neurophysiological estimates" (Nijhawan, 1994, p. 257) to make real time interaction possible. He argues that this mechanism works in the case of the predictable constantly illuminated segment of the bar, but not for the less predictable strobed segment. (Eagleman and Sejnowski, 2002)'s attempt to refute this interpretation is marked by a similarly objectivist-representationalist logic: in an attempt to keep time and space strictly separated in the effect, they argue that the FLE may have been misconceivably considered a temporal illusion. The effect could instead be a spatial illusion that results from inaccuracies in the inference processes that the brain performs to determine the location of the constantly illuminated part of the bar and the temporal cost of performing this computation.

Immaterial of the evidence the different positions in this controversy are based on, from a constructivist position the problem to be solved there, i.e., what is the 'real' temporal and spatial properties of the 'internal representation' of the stimulus, is fully artificial. The

FLE manifests as a lag, which is a spatio-temporal phenomenon, not a spatial one, not a temporal one, and it is neither possible nor necessary to tease the two dimensions apart, neither in the neural nor in the mental domain. From a constructivist perspective, the distinction between temporal and spatial phenomena can only be performed on the basis of meaningful differences between the two that manifest in the behavioural and mental domain of the subject itself. The paradox exists for the experimenter who expects a representation of his own experience of time and space on a symbolic level (a box and an arrow) in the mind and in the head of the subject and thereby turns the constructivist question of the origins of spatiality and temporality upside down.

From a constructivist perspective, no coordination other than that of real physical behaviour in the real physical world is necessary. Similarly, distinctions between spatial and temporal phenomena are not required on a mechanistic level, as long as the relevant distinctions can be made behaviourally, where required. However, it does not require a fully-fledged epistemological constructivism or commitment to the radically enactive approach proposed in this book to recognise the fallacies of naïve representationalism. For instance, (Nijhawan, 2004) has recently contradicted his own previous view. In the revision of his earlier stance he argues that "the 'real' in the ['*vdt*-lag' premise] is an unobservable quantity" because, in closed loop interaction, "many features of 'real' objects 'out there' (e.g., position) are due to descending (internal) neural signals, processes that are related to feed-forward motor control and to Helmholtz's notion of reafference. The view that emerges is that an output of one modality (e.g., object-position given by the visual system) can be related (compared) to the output of another modality (e.g., hand-position given by the motor system), but not to some idealistic 'really' given position"(Nijhawan, 2004). This view predicts an effect similar to the FLE to occur in motion, which Nijhawan confirmed empirically (Nijhawan and Kirschfeld, 2003).

Similarly, (Dennett and Kinsbourne, 1992) point out that, even within a representationalist stance, Libet's interpretation of his own results comes down to a confusion of content (what is represented, i.e., temporal information) and vehicle (what represents, i.e., neural signals with temporal properties). They comment that, on a macroscopic level (i.e., the symbolic-narrative level of time experience), these two – perceived order of stimuli and physical order of correlated neural events – coincide, which leads to the presumption that this should be necessarily the case. This confusion is what they call the "Cartesian trap" (Dennett and Kinsbourne, 1992). By contrast, the mentioned irregularities occur for "events that [are] constricted by unusually narrow time-frames of a few hundred milliseconds" (i.e.,

the immanent level of time experience) and "[a]t this scale [...] the standard presumption breaks down" (Dennett and Kinsbourne, 1992).

Whilst their representationalist approach makes sense for the kind of phenomena they discuss (Libet's results and others akin to the psi-phenomenon), i.e., phenomena in which mental time is 'intact' according to Newtonian standards, it does not serve to explain the "distortions and disruptions of time perception" on the micro-level that (Ivry and Schlerf, 2008) and others have observed. In these cases, the physiological properties of the nervous system shuffle up the 'real order' of events not only on a mechanistic level but also on a mental level, which refutes the presumed arbitrariness of the symbol that Dennett and Kinsbourne invoke when they claim independence between content and vehicle. Dennett and Kinsbourne are right in pointing out that there is no reason to expect neural processes to veridically represent 'real time' – where they err is when they presume that, for a *mental* representation, this should still be strictly the case.

An interesting computational model of time perception that challenges this primacy of mental time is presented in (Karmarkar and Buonomano, 2007). They implement the idea that, once complex dynamically coordinated processes happen, you can just *read out* time, rather than to explicitly measure and keep track. Their intrinsic model of time perception implements a large, randomly connected neural network as a dynamical repertoire, similar to the idea of reservoir computation (e.g., Maass *et al.*, 2002; Jaeger and Haas, 2004) but very closely models the physiology in the relevant areas. Due to its dynamic complexity, the reservoir contains traces of all temporal patterns one could possibly be interested in is intrinsically contained. They train four output neurons to read out the relevant intrinsic dynamics to perform duration judgements. This model successfully predicts nonlinear interactions between duration judgements in humans, depending on inter-stimulus intervals and multiple stimulus presentation. The nonlinear interactions result from transient dynamics and consequent initial sensitivity of the dynamical system and cannot straight-forwardly be explained in linear models. The striking lesson that this model teaches us is that a 'representation' or measurement of time can be a cheap epiphenomenon of ongoing activity dynamics in any neural population, even a randomly coupled one. This model shows us that, from a dynamical systems point of view, temporal coordination can precede temporal measurement, rather than to rely on a clock mechanism as a necessary pre-requisite, as it appears in the computationalist paradigm.

An example from the domain of spatial cognition that serves well to illustrate the fallacies of naïve representationalism in simple sensorimotor behaviour and perceptual experience

criticises the conventional view that 'vision for perception' and 'vision for action' are processed and encoded separately (a functional separation that is usually reduced to the ventral and the dorsal stream in terms of neural mechanism (Milner and Goodale, 1995)). Defendants of this traditional view observe that perceptual misjudgements do not usually lead to motor misalignments, which leads them to the conclusion that the 'incorrect' visual representation has to be processed separately from the 'correct' motor representation. Recent approaches (e.g., Dassonville and Bala, 2004; Li and Matin, 2005) have shown that this logic is not stringent from a closed-loop perspective: by contrasting open-loop perceptual experiments leading to spatial irregularities in perception with complementary open-loop motor experiments leading to inverse spatial irregularities in motion, these researchers came up with what was recently termed the "two-wrongs hypothesis" (Dassonville *et al.*, 2009). This hypothesis states that an 'inaccurate perception' does not necessarily lead to detrimental effects on action, if the motor system cancels out for the systematic perceptual error with an according systematic motor error. Even though examples so far focus on spatial phenomena (and despite the fact that this view is still homuncular in its essence) the general lesson also applies to temporal phenomena. A discrepancy between the observer's frame of reference and the subject's frame of reference is not necessarily a problem.

As pointed out earlier, the *a priori* intertwinement of space and time in cognition and behaviour and the *a posteriori* constructions of a distinction is one of the hallmarks of a constructivist approach as opposed to a representationalist approach (for instance, conceiving of the FLE as a spatio-temporal effect, rather than a spatial or temporal effect). In his ecological perception approach, Gibson postulates that "we have accepted space-perception as a valid problem, but have been uncomfortable about time-perception. We have attempted to keep separate the problem of detecting patterns (objects) and that of detecting sequences (events). And hence the equivalence of pattern and sequence, of space and time, has seemed to be a puzzle which had better be swept under the rug than confronted" (Gibson, 1982, p. 174). Taking into consideration the sensory physiology of humans, Gibson characterises the situation as follows

> "The eyes of primates and men work by scanning – that is, by pointing the foveas at the parts of a scene in succession. The eyes of rabbits and horses do not, for they see nearly all the way around at once and have retinas with little foveation. Does this mean that a horse can perceive his environment, whereas a man can apprehend it only with the aid of memory? I once thought so on the theory that successive retinal images must be integrated by memory, but this now seems to be wrong. It is truer to suppose that a visual system can substitute sequential vision for panoramic vision, time for space. Looking around is equivalent to seeing around, with the added advantage of being able to look closely. It is no

> harder for a brain to integrate a temporal arrangement than a spatial arrangement" (Gibson, 1982, p. 174).

Gibson's insight and his conclusion that "the perception of space is incomprehensible unless we tackle it as the problem of space-time" (Gibson, 1982, p. 175) resonates with Lenay's assessment that "if perception is constituted at the core of a closed sensorimotor loop, enriching perception [...] should be equally possible by means of enriching the sensory inputs at any moment or by means of enriching the repertoire of possible actions" (Lenay, 2003, p. 57)[12], which has been investigated by means of experimentation with receptive field parallelism.

The given examples from cognitive neuroscience and behavioural psychophysics show two things very clearly. Firstly, the immanent level of temporal experience may be immune to the kinds of manipulations described in the previous section,[13] but it can be influenced, distorted and brought to break-down by a different class of manipulations, specific to the time scale on which it is constituted. In turn, these manipulations are impotent in affecting the third and constructed layer of time perception, or only to the extent that it recursively relies on the immanent layer, which is reflected in (Dennett and Kinsbourne, 1992)'s distinction between microscopic and macroscopic events. The other main conclusion to be drawn from this analysis is the fundamentally different perspective that representationalist and constructivist approaches have on the irregularities observed. Constructivist perspectives try to explain the origins of temporal or spatial experience or time and space perception from the bottom-up and try to ground these distinctions in the characteristics of embodied and situated interaction with the environment. On the other hand, objectivist-representationalist approaches already contain such conceptions as an explanatory premise: Newtonian concepts of time or space that characterise the observers conception of the world are invoked as *a priori* target outcomes for processes of internal representation, whereby they entangle themselves in chains of apparent paradoxes that result from the artificial problem of coordinating internal time and external time.

[12] My translation: "En effet, si la perception se constitue au cœur du couplage sensorimoteur, elle doit pouvoir être enrichie [...] aussi bien par un enrichissement de l'entrée sensorielle délivrée à chaque instant, que par un enrichissement du répertoire des actions possibles" (Lenay, 2003, p. 57).

[13] e.g., drugs or stages of cognitive development: even if, under the influence of psycho-active substances, I feel I can travel forwards and backwards on the arrow of time at will, this travel will still be experienced as a chaining of moments, of spatio-temporal object-events.

8.7 Time Experience

This itinerary across issues and disciplines is a strain on the reader. The objective of the previous summary is certainly not to give an exhaustive cross-disciplinary account of time cognition. Each of the sections introduced only a small number of selected findings from very different areas concerned with time cognition, temporal experience and time perception. However, sketching the landscape of methods, perspectives and findings, it is possible to identify connections and make them explicit, indicating the directions in which to venture when addressing a problem within the area of time cognition and time experience. This section aims at integrating the potpourri of results into a somewhat more coherent picture.

We can distinguish three dimensions according to which we can characterise approaches to time cognition and temporality. Firstly, there are the three levels of temporal experience identified by the phenomenologists. Whilst the philosophical approaches sketched in Sects. 8.3 and 8.4 span these levels, the empirical findings are more or less confined to the realm of the descriptive-narrative level of time experience (Sect. 8.5) or the immanent level of time experience (Sect. 8.6) respectively. Secondly, there is a methodological continuum, from a mere first person approach (Sect. 8.3) to a conceptual-contemplative approach making links to the physical world (Sect. 8.4) to data-driven approaches that use second person methods (Sect. 8.5) and third person methods (Sect. 8.6) either proportionally or exclusively. Thirdly, there is the ideological dimension, reaching from radical computationalist approaches (e.g., Eagleman and Sejnowski, 2002) over intermediary positions (e.g., Gibson, 1982; Nijhawan, 2004) to a radical constructivist-enactive perspective as the one proposed in this book or Varela's (e.g., Varela, 1996) work.

Furthermore, the issues mentioned at the beginning of this section, i.e., levels of time experience, the intertwinement of time and space and the role of the known, the unknown and the possible (e.g., in the Aymaran culture or in visual prediction), recurred across the accounts given. However, now we have both the vocabulary and a rudimentary empirical basis to address them with a set of more specific questions: what is the relation between pastness and knowledge? What are the appropriate methods to investigate experienced simultaneity of local events? How can empirical findings obtained with a particular method be fitted into the landscape drawn? What are the structural similarities between the processes that shape the narrative-descriptive level of time experience and those that shape the immanent level of time experience? What is the origin of experienced order, on which level does it take place and what do disruptions of experienced order teach us? Equipped with this conceptual ammunition, the problem of experienced simultaneity and adaptation to

sensory delays can be approached in an informed way. The following two chapters present an experimental study and its ER model on this topic, which is evaluated in the context of the material presented here in 11.

Chapter 9

An Experiment on Adaptation to Tactile Delays

Following a somewhat dazzling cruise through different issues concerning time and temporality on different levels and across disciplines, this chapter concentrates on a specific topic in the area of time perception: experienced simultaneity and its plasticity through adaptation to sensory delays. The topic is introduced in the light of the previous analysis, followed by the presentation of results from an experimental study on adaptation to tactile delays. This study had been conducted in collaboration with the CRED group in Compiègne. The experiment tests the hypothesis that time pressure is necessary to yield an adaptation effect. This hypothesis is based on previous research that has shown that adaptation to sensory delays only occurs in some experiments, not in others.
Given that the data presented in this chapter does not support this hypothesis, the experiment is difficult to interpret in terms of the problem of perceived simultaneity. However, in the light of the methodological theme of this book, it is worthwhile to present the research as an example for the practice of designing and conducting an experiment and engaging in complementary ER modelling. The following chapter presents the ER model of the experiment. The combined insights gained from the ER simulation model and its application to the experimental data are evaluated in the context of the preceding analysis of embodied time cognition in chapter 11.

9.1 Adaptation to Sensory Delays and the Experience of Simultaneity

In a recent study, (Cunningham *et al.*, 2001a) report patterns of adaptation to artificially prolonged sensory delays in human participants in a visuo-motor task that are similar to those obtained in experiments with spatial displacement through prism glasses (e.g., Welch, 1978). Firstly, over training, the initially impaired performance is recovered and the annoying delay disappears from conscious experience. Secondly, re-adaptation to the normal

condition is marked by a strong negative after-effect, i.e., participants' performance on the unperturbed condition without delay is worse after training with a 200 ms visual delay. Although their study focuses on the behavioural aspects of the task, the authors report as anecdotal evidence that several subjects spontaneously reported that "when the delay was removed, the plane appeared to move before the mouse did – effect appeared to come before the cause" (Cunningham *et al.*, 2001a, p. 533).

Such patterns of behavioural adaptation appear plausible in the light of the analysis given in the previous chapter. A recalibration of experienced simultaneity seems a logical reaction to the manipulation of the sensorimotor coupling. The rules of sensorimotor invariance that correlate with the experience of presentness are changed by means of the increased sensorimotor latency. A time span during which the subject cannot take further influence on a process it has initiated, for all practical purposes, does not exist 'as a future', and may just as well disappear from consciousness. Such a view corresponds well to (Libet, 2004)'s result about systematic neuro-behavioural latencies that the experimenter can observe, but that are, in contrary, not part of the subject's own temporal experience. What is the use of perceiving that one is always a little bit late, if there is nothing one can do about it?

When the reverse manipulation is performed, i.e., sensorimotor latencies are shortened back to the original value, not only does the performance fall dramatically below the level initially measured without delays, also the experience of presentness is brought to a breakdown or into logical conflict. This reversal of perceived cause and effect appears reminiscent of Grey Walter's results from the 1960's (as reported in Dennett and Kinsbourne, 1992) about artificial shortening of inherent neuro-motor latencies. Walter brought the inherent neuro-motor latencies involved in motor decision making to the subject's attention using real time brain computer interface, which leads the subjects to reject ownership of the consequent action, even though it is just minimally (hundreds of milliseconds) faster than the naturally executed action (cf. Sect. 8.6 in the previous chapter). Given these known patterns, why is (Cunningham *et al.*, 2001a)'s result so surprising?

What makes (Cunningham *et al.*, 2001a)'s findings so interesting is that, at several occasions, similar adaptation effects had been hypothesised, but had failed to occur. This repeated failure to obtain sensorimotor recalibration to sensory delays even led (Smith and Smith, 1962) to conclude that adaptation to sensory delays is impossible in principle. Also, following up on Cunningham *et al.*'s reported results, (Stetson *et al.*, 2006) tried to reproduce the effect in a minimalist psychophysics set-up, but only produced partial readjustment of perceived simultaneity, which is of the order of magnitude (tens of milliseconds) of

recalibration effects in merely passive recalibration paradigms (e.g., Fujisaki *et al.*, 2004). This is in line with the corpus of previous and later studies on perceptuo-motor tasks in which adaptation effects to sensory delays failed to occur, such as visuo-motor tracking (Kennedy *et al.*, 2009), telesurgery (Thompson *et al.*, 1999) or remote manipulation (Ferell, 1965). Similarly, (Held *et al.*, 1966) report that visual delays produce a disruption of adaptation to spatial displacement. This non-exhaustive listing contains studies with delays within the range of less than 100 ms to over 1 s, from different modalities, from active and passive conditions and from different behavioural task domains. What is it about Cunningham *et al.*'s study that makes them different from those previous and later studies that failed to produce the described adaptation effect?

The authors themselves hypothesise that the observed adaptation effect is due to the time pressure in the task that makes the delay meaningful for the solution of the task:

> "[...] it has been clearly demonstrated that sensorimotor adaptation requires subjects to be exposed to the consequences of the discrepancy [...]. Thus, it is of central importance to note that subjects in previous studies slowed down when the delay was present. [...] This is crucial because slowing down can effectively eliminate the consequences of the delay" (Cunningham *et al.*, 2001a, p. 534).

This observation relies on a definition of adaptation that the authors adopted from (Welch, 1978) as '*semi-permanent*' change in perception that eliminates behavioural errors and/or the registration of a perturbation. Furthermore, the authors measure adaptation through the *negative after-effect* which they call the "most common measure of adaptation" (Cunningham *et al.*, 2001a, p. 533). A negative after-effect is the reduced ability to accurately perform the task when returning to the original condition of the task, prior to the introduction of a perturbation and to training (usually, this involves an inverse behavioural error to the one that occurred when the perturbation was first introduced).

Slowing down, as a compensatory strategy, may help to improve performance on a given task with sensory delays to a certain extent. It is, however, not a strategy that produces a negative after-effect or semi-permanent adaptation, but instead a cognitive compensation strategy. In a follow-up study in a multimodal task (Cunningham *et al.*, 2001b), the reported adaptation could be reproduced under time pressure. In a delayed vestibular feedback condition (Cunningham *et al.*, 2001c), however, only partial adaptation was found.

A representationalist interpretation of the observed recalibration effect is exemplified in (Stetson *et al.*, 2006)'s hypothesis that "sensory events appearing at a consistent delay after motor actions are interpreted as consequences of those actions, and the brain recalibrates timing judgments to make them consistent with a prior expectation that sensory feedback

will follow motor actions without delay" (Stetson *et al.*, 2006, p. 651). This open-loop perspective does not assign significance to the nature of the task, the subject's goals or the properties of sensorimotor couplings. The authors presume that adaptation proceeds automatically and based on statistical and correlational properties of the inputs alone; that a process external to the behaviour itself infers causality on the basis of the input statistics. The authors do not make mention of the failure of the adaptation effect to occur in previous studies, or assess what the partial adaptation they gain implies for real-time coordinated behaviour, i.e., if it would actually help to mitigate problems brought about from increased sensorimotor latencies. Their best guess towards why the adaptation they obtained was only partial is the hypothesis that "it may be that motor-sensory timing shifts of 100 ms are beyond the hardware limitations of the calibration mechanisms" (Stetson *et al.*, 2006, p. 656). The task in (Cunningham *et al.*, 2001a)'s experimental paradigm is of a fundamentally different nature: it relies on the significance of the perturbation within the closed sensorimotor loop (reward in the task relies on real-time delay compensation). Also, by means of active exploration, the statistics of the input patterns are brought about by the subject itself, allowing the subject to recognise the causal links between the efferent signal and reafferent stimulation in a spatially embedded and continuous way.

In a more ecological perspective, adaptation would not seem advisable in the paradigm that (Stetson *et al.*, 2006) developed. In our day to day life, there are numerous events that involve systematically correlated latencies that are due to external causal sources (throwing a stone and hearing it drop, pushing a pendulum and seeing it swing back, *etc.*). From our experience, we know, that we can still intervene and modify the course of events while the temporally extended process unfolds. This is not the same for our inherent sensorimotor latencies. We cannot take influence on the course of our actions in execution, and, therefore, such sensorimotor latencies do not exist in any meaningful way to us as organisms. It makes sense to make the time that passes between us deciding to act and us perceiving the outcome of this action part of our memory, something that has already passed. This factor, the possibility and the intention to intervene do not figure in in the open-loop approach that (Stetson *et al.*, 2006) adopt. Therefore, the outlined scientific problem is a possibility for the enactive approach to elucidate what is going on in a more embodied and embedded context and thereby add variables and factors that open-loop approaches, starting from an information processing perspective, deem irrelevant and leave out.

This problem and the different ways of approaching it and their relative success form the starting point of the experimental study presented. The objective was to reproduce the

findings reported in (Cunningham *et al.*, 2001a) in a minimal sensorimotor task. The experimental and modelling approach described and developed in chapter 3 was pursued, in order to find the minimal conditions for semi-permanent adaptation to sensory delays and distinguish them from conditions in which the adaptation is not produced. The experiment tests the hypothesis that time pressure in a closed-loop sensorimotor task with online control is necessary and sufficient to produce semi-permanent adaptation to sensory delays. The active component makes the delay a meaningful discrepancy and the time pressure is what requires adaptation, rather than just cognitive compensation by slowing down.

To delimit the problem in the terms developed in the previous chapter, this effect occurs at the level of the immanent flow of time at the scale of temporal object events. The reported results have no effect on macroscopic symbolically constructed time-scales. Along the methodological dimension, the project clearly focuses on third person methods. Measurable behavioural variables, such as negative after-effects, are seen as indicators of the perceptual world and its adaptation. No perceptual judgments ('crude' phenomenological data, cf. chapter 3) are recorded. A questionnaire had been handed out to ask subjects for a description of their experience of the task and their experience of the strategies they adopt. However, due to the difficulties that untrained individuals have with verbally describing their experiences, they did not produce useful results. The study is thus confined to behavioural data. It investigates the plasticity of experienced nowness within the level of the immanent and continuous flow of time, not its qualitative distinction from the other levels (in terms of neurophysiological or functional processes). As concerns the ideological dimension, the hypothesis adopted and the approach pursued reflect the enactive and constructivist perspective underlying this book.

9.2 Methods

The study is only briefly presented here. The reader who is interested in the technical details of the experiment is referred to the dissertation on which this book is based (Rohde, 2008). The project was conducted during a placement in the CRED group in Compiègne, using the audio-tactile feedback platform Tactos (Gapenne *et al.*, 2003) and with their help and advice.[1] The Tactos system links participants' motion in a simulated environment (movement of mouse, stylus, *etc.*) systematically to patterns of tactile stimulation on a Braille display (see Fig. 9.1). It can be used as a perceptual supplementation device, as

[1]Noticeably, the support of Olivier Gapenne, Dominique Aubert, John Stewart and Charles Lenay should be mentioned here.

outlined in chapter 3 and (Lenay *et al.*, 2003), to investigate the perceptive qualities that result from training with devices providing previously unfamiliar sensorimotor couplings. The aim was to find the minimal conditions under which the semi-permanent adaptation to sensory delays takes place that (Cunningham *et al.*, 2001a) report and to distinguish them from similar experimental conditions in which this adaptation does not occur. The experimental set-up in Cunningham *et al.*'s experiment is already simple: participants move along one dimension (mouse movement to the left and right) in order to avoid evenly spaced obstacles. These obstacles are arranged in a field that participants traverse at a fixed linear velocity from the bottom to the top (i.e., orthogonal to the direction in which they can move with the mouse). However, despite this restricted possibility for movement in the simple task, the visual inputs provided are comparably rich and difficult to interpret in terms of possible sensorimotor circuitry to bring about the behaviour. Besides the non-delayed proprioceptive/reafferent feedback about self-movement and the position of the mouse, the screen provides a visual representation of the field of obstacles and the location of the airplane. The airplane is delayed by an additional 200 ms in the delay condition to which participants are supposed to adapt.

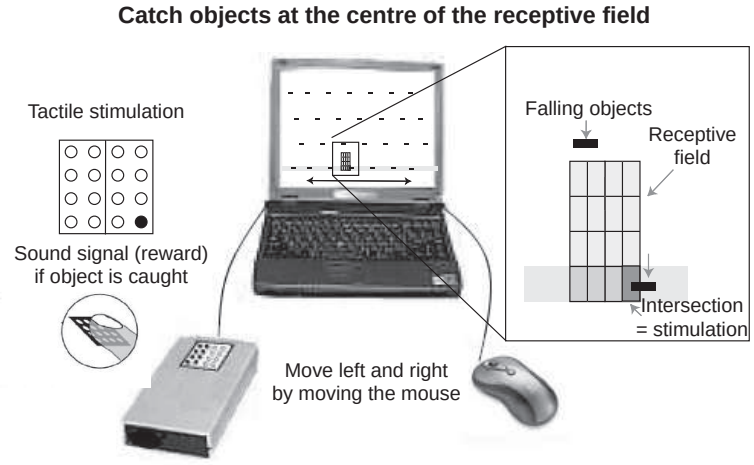

Fig. 9.1 The Tactos tactile feedback platform. Task: objects have to be located in the centre of the receptive field when they reach the bottom line.

The visual sense is a very complex sense and it is difficult to explain what in the complex and informationally rich representation has been exploited to solve this task. Therefore, in order to find the minimal conditions for adaptation to sensory delays and be able to anal-

yse the sensorimotor dynamics of adaptation, the most important part is to simplify the sensory component. As part of the simplification, visual feedback was replaced through audio-tactile signals. Participants were blindfolded. They received tactile stimulation via a Braille display (see exact specification below) and auditory signals indicated object velocity as well as reward for successful behaviour. Participants could move left and right on a tape, while objects fell down with variable velocities from the vertical dimension. The left and right dimension wrapped around, i.e., the tape on which they moved was infinite (see Fig. 9.2). When the receptive field (see Fig. 9.1) intersected with an object, the Braille display represented this intersection to the subject's fingertip (height coded for distance till impact, width coded for whether the object was to the left, to the right or in the centre). Subjects could catch these objects by positioning themselves directly under it when it touched the bottom line. Subjects could thus only catch one object per line of objects.

Lateral distance between objects corresponded to ca. 0.5 cm on the desk. In terms of virtual distance units, this comes down to 28 units. This compares to a width of 4 for each of the objects and a width of 16 for the respective field. Depending on the object velocity, there was a time window of 1-4 s from when the object first was in the reach of the virtual perceptive field, during which subjects could perform this positioning action. This small time window brings the time-pressure to the task, which we hypothesised to be essential for adaptation to sensory delays. If subjects achieved to position themselves under the object, they were given an auditory reward signal. Auditory pulses counting down till the object reached the bottom line indicated the velocity of the current row of objects. Even though this information was also present in the tactile stimulation patterns, the auditory pulses made it possible for subjects to perceive the velocity of an object when they were not currently in tactile contact with it.

Due to a technical problem, the operating system's mouse acceleration was applied to the mouse movement, such that the recorded mouse movement trajectories are spatially distorted. This distortion does not appear to make a difference to the general outcome/behaviour, but it means that analysis of any spatial aspects of the behaviour elicited is likely to be not fully accurate.

In the delay conditions of the experiment, both the tactile and the auditory signal were delayed by 250 ms additional to the inevitable delay of $\approx$ 35 ms the computer induced and that is present (though not perceivable) in all conditions. This delay is of a similar magnitude to the one (Cunningham *et al.*, 2001a) use in their task. A delay of this magnitude is large enough to be perceptible to most subjects.

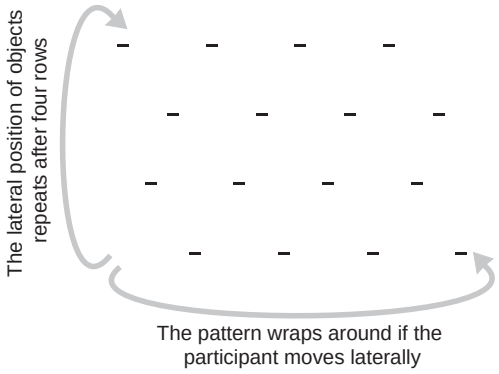

Fig. 9.2 The repetitive lateral displacement of rows of objects.

The experiment was performed on 20 unpaid subjects of different age-groups (mostly graduate students) and both sexes that participated in the experiment as a part of a cognitive science conference in Bordeaux (ARCo'06). The experiment consisted of five experimental phases that, altogether, lasted 30-45 minutes. After familiarisation with the task and the set-up, subjects were assigned to one of three velocity groups on the basis of performance. Subjects from within one group were tested on the same sequence of 32 object velocities four times, prior to training on both the undelayed and the delayed condition and after two blocks of ca. five minutes of training, first in the delay condition, then, as post-test, in the non-delayed condition.

Participants had been informed in advance about the delay and knew, cognitively, whether they were dealing with a sensory delay or not at any moment. Despite this information, some subjects reported that they did not perceive the delay as delay, but rather as 'something wrong'. Some subjects even reported that they only experienced that it was indeed a delay when they returned to the original condition.

Performance F (in allegory to 'fitness' in ER simulations) is defined as

$$F = \frac{1}{32}\sum_{1}^{32}|d_h| < 4 \tag{9.1}$$

where d_h the distance between perceptive field centre and object margin at the time the object reaches the bottom line. Behavioural data (i.e., motion, position, sensory stimulation) was recorded in order to analyse it for closed-loop behavioural correlates of the hypothesised perceptual changes.

9.3 Results

Figure 9.3 depicts the performance profile of participants with and without delays, before and after training. We expected a decline in performance between pre-test and post-test (negative after-effect). A repeated measures ANOVA (with experimental phase as factor) confirmed that the change in performance across phases is significant ($F(3,57) = 23.96; p = 0.4 \cdot 10^{-9}$). However, pair-wise comparison showed that the only significant differences are the drop of performance when the delay is introduced ($p = 0.4 \cdot 10^{-7}$) and its improvement when the delay is relieved ($p = 0.5 \cdot 10^{-5}$). Performance markedly recovered with training (mean improvement of 0.08 comparing the delay-test with the adaptation-test). However, this recovery was not statistically significant ($p = 0.062$). At these earlier phases of the experiment, some individuals followed already very unexpected patterns in their performance: some maintained their level of performance when the delay was introduced, or even got slightly better with it, whilst others got worse with training, which explains the fact that performance recovery is not significant.

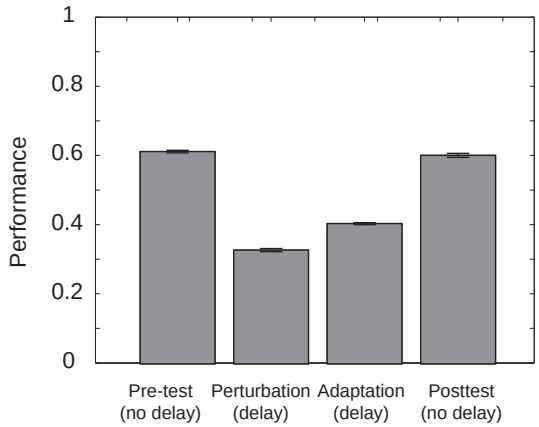

Fig. 9.3 Participants' performance with and without the delay before and after training with sensory delays (error bars: standard error of the mean). There is no significant after-effect and not even a significant improvement in performance. (The small error bars across participants are misleading; the patterns of change in performance across the experimental phases differed a lot from participant to participant, which is why changes were not statistically significant.)

While it may still be argued that there is a trend for recovery which is masked by noise, there is clearly no evidence for a negative after-effect. Performance decreased between pre- and post-test by a negligible 0.01, the participants' performance stayed literally unaltered, so the main hypothesis was not supported. How should the failure of this experiment be

interpreted? Is it true, as (Stetson *et al.*, 2006) argue, that it is impossible for participants to adapt to delays of this magnitude? Does time pressure not play a role in this kind of recalibration? Was it too easy to achieve cognitive compensation, rather than semi-permanent perceptuo-motor adaptation, and, for that reason, no after-effect occurred?

Eyeballing the movement data, there appear to be some changes in behaviour induced by the training with delays that do not impact on performance. Figure 9.4 shows an example trajectory of a subject that showed no deterioration of performance between pre-test and post-test, but whose sensorimotor strategy changed across the different phases of the experiment. During the pre-test and the adaptation-test (measurement at a phase where subjects are familiar with the current sensorimotor latencies), the subject was more exploratory and actively scanned the objects several times before halting and catching them. During the delay-test and the post-test, when the manipulations were unfamiliar, the subject reacted using a more careful, hesitant strategy. This particular pattern is not a trend to be found in many subjects – some reacted just the opposite way, others did not react at all to the changes in sensorimotor coupling. However, what it shows is that the performance measure in the task does not capture such adaptive changes, changes that may still correlate with perceptual changes of the kind we were interested in.

In trying to understand what (if anything at all) happens in a systematic way across the different phases of the experiment, different variables describing the behaviour were investigated (velocity, number of crossings with the object, proportion of time spent in motion, ...). However, at first glance, there appears to be a general trend to become more rigid in behavioural strategy when the unpleasant delay is introduced, a trend that is carried over to the post-test. The only descriptive variable that changed significantly between pre-test and post-test is the average time spent in motion before stopping and catching an object, which decreased already from the moment the delay was introduced from 546 ms to 483 ms and stayed at about that level. Such a marginal change in a single variable, discovered through *post-hoc* data analysis, does not provide a strong basis to argue for systematic adaptation effects or characteristic strategies on a behavioural level. On the basis of the behavioural data alone, no trend, explanation, message or lesson could be derived.

As stated earlier, given that the tested hypothesis is not supported by the data, it is difficult to interpret them with respect to the problem of delay adaptation and perceived simultaneity. However, another objective of the project had been to test the usefulness of combining and co-developing minimal behavioural experiments with humans and ER simulation models (as outlined in Sect. 3.6), where ER simulation modelling should serve to clarify issues

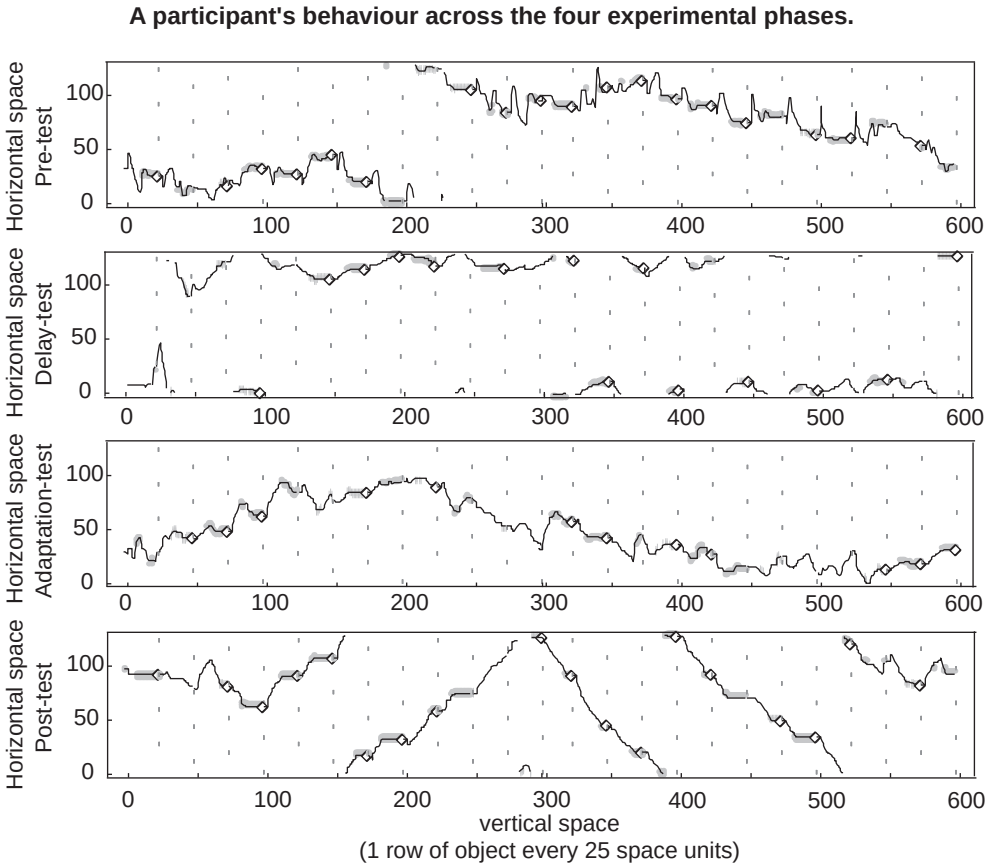

Fig. 9.4 Trajectories of an example participant over the course of the experiment (normalised for distance, not velocity between rows of objects; grey shades indicate tactile stimulation, diamonds indicate catch events). Even though the performance is identical in pre- and post test (75%), the behavioural strategy appears to change over training. During the pre-test and the adaptation-test, there is an ongoing online correction (swaying), whereas the delay-test and the post-test are marked by more careful slow movements and long periods of immobility, a change that is not reflected in the catch performance.

in sensorimotor dynamics that are easily overlooked from an open-loop and explicit design perspective. The following chapter presents a simulation model of the task, which provides some interesting general insights about the task and the functional role of delays in general (as analysed in chapter 11). The later part of the following chapter revisits the behavioural data presented here and presents some further tests and observations that are informed by the results from the simulation model and that confirm predictions about human data generated by the model.

Chapter 10

Simulating the Experiment on Tactile Delays

Alongside the experiment that was described in the previous chapter, an ER simulation model of the kind presented earlier in this book was conducted. Agents were evolved to perform the same behaviour as the experimental participants, i.e., to catch objects through simulated tactile feedback, in the environment described. The results from the simulation, which were conducted to aid experimental design, data analysis and interpretation, were in parts published in (Rohde and Di Paolo, 2007). This chapter presents these simulation results and then revisits the data presented in the previous chapter, in order to see in how far the insights gained in the simulation model apply to the experimental data. The model generates descriptive variables and concepts that are then tested against the data. However, the most significant results from the simulation are conceptual insights about the meaning of delays in different kinds of sensorimotor loops (*reflex-like, reactive* and *anticipatory*). These will only be discussed at length in the following chapter 11, which evaluates the data from both the simulation model and the experiment in the light of the larger picture of embodied time cognition and time perception given in chapter 8.

10.1 Model

The model presented here has a similar purpose as the models on perceptual crossing. It serves to explore the space of simple circuits that can bring about the required task. Thereby, it should point out the behaviours that are characteristic for a certain strategy, what they share in common and how their strategies compare (quantitatively and qualitatively) to the behaviour we observe in humans. The task posed to the agents was again very similar to the one posed to humans. As for previous models, the reader who is not interested in the technical details of the model is invited to move over to the results and discussion parts of this chapter.

The virtual task environment, in which the agents were evolved is in most respects identical to the one used for the experiment. Artificial agents can act by moving left or right in an infinite one-dimensional space (see Fig. 10.1), while evenly spaced objects (same sizes, distances, velocities *etc.* as in the human experiment) fall down in a direction vertical to the agent movement and have to be caught. Each trial consists of a sequence of 32 objects at variable random velocities (i.e., the agents were not tested on the fixed sequences across conditions that the participants were tested on). Even though the size of the agent's perceptive field is the same as the human participants' (16×8 units), the exact tactile input patterns the participants received are transformed in a way more suitable for minimal CTRNN controllers. A continuous input signal is fed into the controller that represents the horizontal distance from the centre when an object entered the receptive field ($I_1 = |d_h|/6$ if $|d_h| \leq 6 \wedge d_v \leq 16$). Signals to indicate the velocity of falling objects (akin to the auditory signals in the experiment) are fed into a second input neuron (I_2). The third input signal used (I_3) is the reward signal, in case an object is caught (rectangular input for 100 ms). Just as in the experiment, an object is caught if it is in the centre region of the agent's receptive field when reaching the bottom line ($|d_h| < 4 \wedge d_v = 0$).

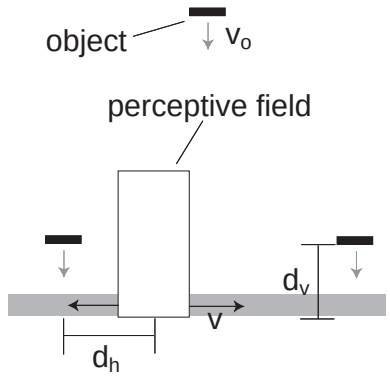

Fig. 10.1 Evolutionary Robotics simulation model of the experiment on adaptation to delays.

All three input signals are fed into the control network scaled by the sensory gain S_G and with a temporal delay. As explained in Sect. 9.2, in the condition 'without delay', there is a minimal processing delay (on average 35 ms) in the experiment, which is prolonged by 250 ms to 285 ms in the delay condition. The same values (i.e., 35 and 285 ms) are used in the simulation. The agents are controlled by a CTRNN (cf. Eq. (3.2)). The three input neurons feed forward into a fully connected layer of six hidden neurons, which feed

the two non-recursively coupled output neurons. A time step of 7 ms was chosen for both the simulation of the network dynamics and the task dynamics, which is a higher temporal resolution for the simulated environment than in the real experiment (ca. 15 ms). The basic velocity output v calculated by the network is $v = \text{sign}(\sigma(a_{M1}) - 0.5) \cdot M_G \cdot \sigma(a_{M2})$, so one neuron controls velocity and another one direction, the motor gain M_G scales the output.
The search algorithm used to evolve the parameters of the control network is the standard generational GA described in Sect. 3.3, vector mutation of magnitude $r = 0.6$ was used. The parameters evolved (145 parameters) are: $S_G \in [1,50]$, $M_G \in [0.001,0.1]$, $\tau_i \in [25,2000]$, $\theta_i \in [-3,3]$ and $w_{i,j} \in [-6,6]$. The fitness $F(i)$ of an individual i in each trial is given by the proportion of objects caught

$$F(i) = \frac{1}{32}\sum_{1}^{32} d_{hi}(T) < 4 \tag{10.1}$$

which is equivalent to the performance criterion used in the experiment (Eq. (9.1)).

10.2 Results

With only two exceptions out of 20 evolutionary runs (1000 generation), solutions for both conditions evolved to a high level of performance (see Fig. 10.2 (A)). On the level of behavioural strategy, the solutions evolved for both scenarios involve halting abruptly once the object is encountered, frequently slightly overshooting the target, to then invert velocity and slowly move back to place the object in the centre of the receptive field. Figure 10.3 (A) shows how this strategy, from different starting positions relative to the object, leads to a stabilisation of position by performing a temporally displaced stereotyped movement. This is a rather trivial strategy. It is probably due to tight temporal constraints on the task and the coarseness of the fitness function, that does not capture well the subtleties of sensorimotor perturbation and adaptation and thus does not encourage the evolution of adaptive or more variable behaviour (see following analysis).
Figure 10.2 (A) displays performance across the four conditions for agents evolved with and without delays, tested under both the delay and the no delay condition. These four tests can be seen as a metaphor for the conditions pre-test, delay-test, adaptation-test and post-test from the experiment: being evolved with or without delays corresponds to the situations in which participants are adapted to a certain condition, i.e. they correspond to pre-test and adaptation test. Testing the agents in a situation for which they are not evolved corresponds to the sudden introduction or removal of a delay, i.e., to the delay-test and post-test condition .In this sense, Fig. 10.2 (A) can be directly compared to Fig. 9.3 from

the experimental data. In this comparison, it can be seen that agents evolved with delays, which corresponds to the adaptation-test, achieve a much higher performance (similar level as without delays) than participants after training with delays.

Comparing how agent performance changes with introduction/removal of the delay, it is obvious that most of the solutions to the task with sensory delays are robust to the removal of the delay, while most of the solutions evolved without delays suffer a drastic breakdown in performance below chance level once the delay is introduced. This result is, to a degree, analogous to the experimental data, in which the delay condition was characterised by a catastrophic performance break-down, whereas removal of the delay led to the immediate recovery of original performance levels. If solving the task with delays in many cases subsumes solving it without in the given experimental set-up, we would have a very simple explanation for the fact that no negative after-effect could be measured in the experiment.

A closer look at the solutions evolved reveals that the velocity at which the object is first touched is on average twice as high for the controllers evolved without delays ($\bar{v} = 0.025$) than it is for the controllers evolved with delays ($\bar{v} = 0.014$). This difference suggests that the agents evolved may simply use the same strategy for both solutions, but slowing down their movement for the delay condition. Such a slowing down is exactly the strategy that the strict time pressure should have had made impossible in the simulation/experimental task. As argued in Sect. 9.1, slowing down to compensate for a delay interferes with semi-permanent adaptation. A very crude test of this possibility is to invert the M_G in agents evolved for either condition, i.e., to double it for agents evolved with delay and divide it by two for agents evolved without delays and investigate the effect of this inversion on performance on either condition. Figure 10.2 (B) depicts the performance profile of agents upon this modification of velocities, and they seem to confirm the apprehension: with this modification, the agents evolved without delays become generalists that perform alright under both conditions, whereas the agents evolved with delays, if sped up, lose their capacity to deal with delays but are still able to solve it without delays. Halving or doubling the velocity inverts the performance profile evolved for each agent originally (Fig. 10.2 (A)).

A closer look into the sensorimotor dynamics, however, shows that things are not quite this simple. As a first step into the analysis, it is established that all evolved controllers function independently of the reward signal and the pace at which the objects fall (I_2 and I_3). Agents simply try to put objects as quickly as possible into the centre of the perceptive field. Therefore, agents produce the same trajectories for different object velocities that are

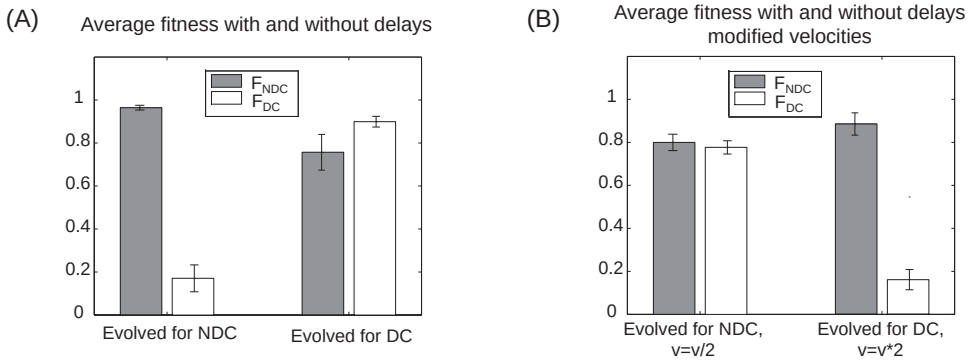

Fig. 10.2 Performance profile averaged over 9 evolutionary runs in an unperturbed condition as opposed to perturbation through scaling the velocity. (A) Unperturbed condition. (B) Scaled velocities (doubled for *DC*, divided by two for *NDC*) leads to an inversion of the performance profiles (error bars: standard error of the mean).

just cut off at different points in time. This simplifies analysis immensely, because object velocities can be largely ignored.

Initially, evolving agents with and without delays had been intended just as the first step for a series of simulation models, with the ultimate goal to evolve agents that adaptively switch strategy during their lifetime according to variable delays. However, most of the agents evolved produced no negative after-effect for shortening of delays and there was no selection pressure to evolve more interesting or adaptive mechanisms than just this robustness. The simulation experiment was not primarily intended as a theoretical study of the principles of adaptation to sensory delays but as a model of the empirical experiment. In this sense, limited adaptivity or sophistication of evolved solutions was actually a good thing, because it mirrored the problems encountered in the experiment with humans, who showed a similar robustness to the removal of delays.

10.2.1 *Systematic Displacements*

Probably the most important result from the analysis is the identification of systematic displacements depending on initial movement direction and velocity. Figure 10.3 depicts trajectories from different starting positions relative to the object position for two example individual agents, one evolved with delays (A) and one evolved without delays (B). The agents were tested without delay (top) and with delay (bottom).

Both achieve to locate the object in the centre of their receptive field for most possible starting positions in the respective condition they have been evolved for (bottom left for agent evolved with delays, top right for agent evolved without delays). Comparing, in con-

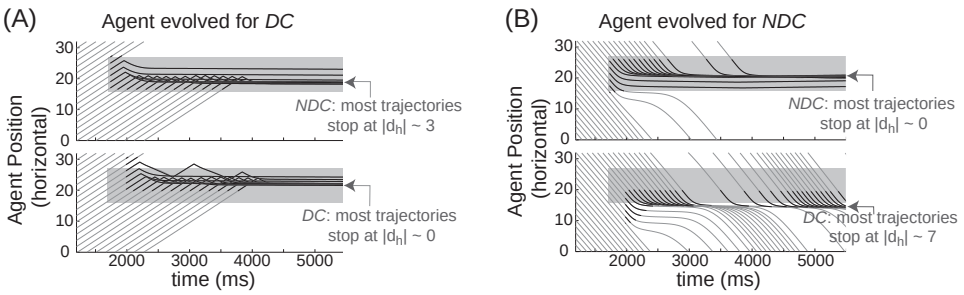

Fig. 10.3 Trajectories for different agent starting positions across time, presentation of a single object. Crossing the object (grey region) produces a (delayed) input stimulus I_1 (trajectories black during stimulation). Top: without delay, bottom: with delay. (A) An agent evolved with delays. (B) An agent evolved without delays.

trast, how the behaviour is altered by the introduction/removal of a delay (top left for agent evolved with delays, bottom right for agent evolved without delays), it can be seen that, in both cases, the trajectories are systematically displaced from the centre of the perceptive field. When the agent evolved without delays is exposed to a prolonged delay (bottom right) it overshoots its goal, while the agent evolved with delays stops too early if the delay is removed (top left). These systematicities are much closer to the behaviour predicted to occur in the experimental participants because both agents appear to be perturbed in their performance by alteration of sensorimotor latencies and one perturbation is the qualitative inversion of the other (negative after-effect).

Why is this systematic displacement disastrous to fitness in agents evolved without delays but interferes little with fitness of agents evolved with delays? As remarked earlier, agents evolved without delays move on average twice as fast. The magnitude of systematic displacements of the type described is proportional to the agents' velocities. The systematic displacement in the slow agent evolved with delay is small enough ($|d_h| < 4$) to still be close enough to the centre to be registered as success, as defined in the fitness function Eq. (10.1). For the agent evolved without delays, the displacement takes trajectories far away from the centre and outside its receptive field, as a direct consequence of the movement velocity when the object is sensed. Such systematic displacement of trajectories can be observed for most agents. The fitness function does not detect or punish such micro displacements. This appears to explain their robustness towards removal of the delay but not its introduction, which, therefore, does not appear to stem from a qualitative differences in functional impact, but rather from the magnitude of systematic displacements that relies on initial velocities.

In order to test this hypothesis, a new set of agents was evolved with a spatially more exact fitness function that measures the exact distance from the object centre, not only the distance if it exceeds 4 units.

$$F'(i) = \frac{1}{32}\sum_{1}^{32} 1 - \frac{\sqrt{d_{hi}(T)}}{4} \tag{10.2}$$

With this modification, solutions evolved with sensory delays cease to be robust to the removal of the delay (see Fig. 10.4), which confirms that robustness of agents evolved with delays is related to the fact that the original fitness function (10.1) is not sensitive to micro displacements. Applying this synthetic insight to the experimental study, which has the same coarse performance criterion, the model generates a possible explanation for why behavioural reaction to the removal of the delay was not reflected in a decrease in performance: if systematic displacements from the exact centre of the perceptive field occur, this suggests that maybe a behavioural after-effect to adaptation to delays occurred, but was not strong enough to trigger a break-down of performance.

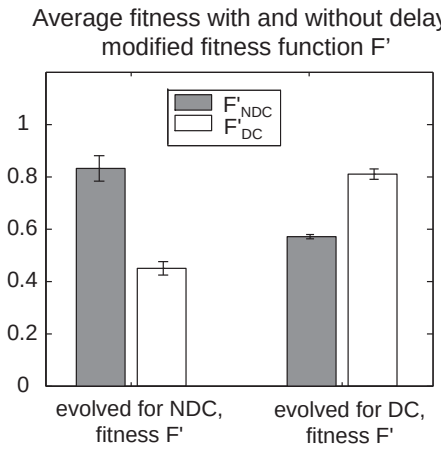

Fig. 10.4 Performance profile with the modified fitness function F' (50% performance chance level, 10 evolutionary runs, error bars: standard error of the mean).

10.2.2 *Stereotyped Trajectories*

Another interesting observation about the solutions evolved is the predominance of *reflex–like behaviour*. Looking at the steady state velocities for varying I_1 representing distance from the exact centre in evolved agents (Fig. 10.5), there is a strong tendency to output

$v^* = 0$ for values of I_1 that exceed a certain rather low threshold value of I_1. Behaviourally, this means that the agents are only sensitive to the onset of the stimulation when an object enters the receptive field, which triggers a rapid decay of v to 0. The exact magnitude of the input signal that represents the exact distance from the centre is not used for further adjustments. The variation in signal magnitude, as an agent moves to the exact position to stop, however, is without effect on agent behaviour. This is why the agent depicted in Fig. 10.3 (A), top, remains in its location displaced from the centre of the receptive field, rather than to actively search for the exact centre. Such strategies are reflex-like in that they produce stereotyped trajectories.

A common variation of this pattern is that deceleration is preceded by a movement direction inversion realised by negative peaks in the steady state profile: the negative peaks in v^* in Fig. 10.5 (left and right) realise this return behaviour (cf. Fig. 10.3 (A)). Such return strategies are, however, equally insensitive for exact signal magnitude.

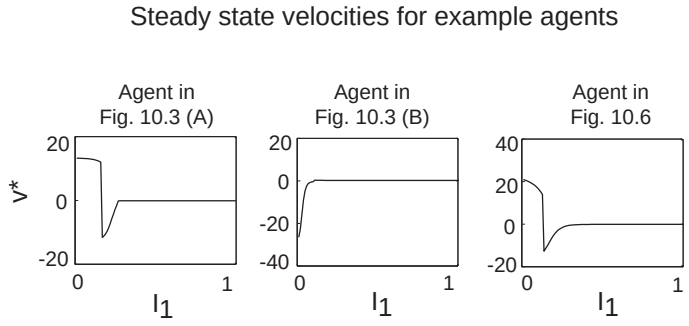

Fig. 10.5 Steady state velocities v^* for different I_1 for the analysed evolved agents in Fig. 10.3 (A) and (B) and Fig. 10.6.

Reflex-like behaviour evolved in all agents but one. The agent evolved without delays whose behaviour is depicted in Fig. 10.6 is one of the two agents that maintain a relatively high level of performance when sensory delays are introduced (cf. Fig. 10.2 (A)). Even though the strategy evolved is also reflex-like in its 'native condition' (i.e., without delay), it allows adjustment of behaviour to a certain degree after performing the first reflex-like positioning: crossing the object, the target is overshot by a large amount and the first movement inversion (induced by lower negative peak in the steady state velocity profile in Fig. 10.5 right) positions the object in the centre in the condition without delay. In the condition with delay, however, this reflex happens to bring the object back into the outside margin of the perceptive field where the other negative peak in the steady state ve-

locity profile is situated (Fig. 10.5 right). Therefore, another return reflex is triggered that brings the trajectory into the centre. In this sense, the behaviour is more *reactive*, because it is sensitive to changes in magnitude of the signal caused by ongoing behavioural dynamics (Fig. 10.6 top vs. bottom).

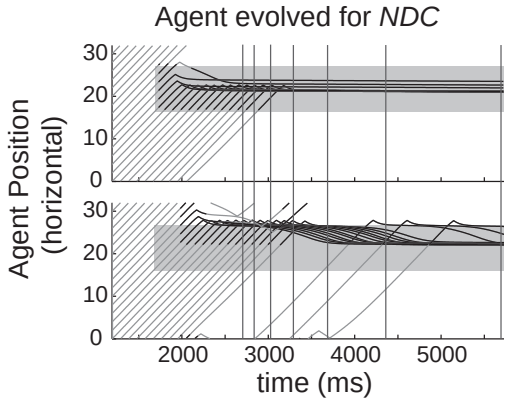

Fig. 10.6 Trajectories for different agent starting positions across time, presentation of a single object. The agent that has been evolved without delays uses a reactive sensorimotor strategy. Crossing the object (grey region) produces a (delayed) input stimulus I_1 (trajectories black during stimulation). Top: without delay; bottom: with delay. Vertical lines: time at which presentation is cut off depending on v_o.

This reactive strategy is, however, plainly accidental and not the outcome of artificial evolution: if the magnitude of the return trajectory or the initial velocity were a bit different, the second inversion of velocity would not be realised in the delay condition that the agent was not evolved on. Reactive strategies did not evolve systematically because the deliberate inherent time pressure in the task does not allow for online correction. The cut off time for trials with the top three velocities is 1000, 1142 and 1333 ms after the objects become perceptible, which corresponds to the vertical lines at $t = 2701, 2843$ and 3033 in Fig. 10.6. A reactive online mechanism to bring back overshooting trajectories needs more time to come into effect. The time window is just big enough to execute a reflex, not for reactive behaviour correction. Agents have to induce the right behaviour immediately when the object is perceived.

10.2.3 *Velocity*

The question remaining is why the solutions evolved for the task without delays are so much faster than those evolved with delays. The intuitive answer to this question is the

wrong answer: slowing down seems the obvious way of coping with a delay – this intuition is, however, only directly true for reactive strategies, in which ongoing behaviour correction is informed by and has to wait for the delayed signal representing the effect of one's own previous actions. For the execution of a reflex there is no immediate disadvantage to high velocity when faced with delayed sensation. Three other possible explanations were explored.

A first explanation would be that velocity is optimised for a simple circuit to drift back to the point of contact: if very fast time constants are used in the output neuron responsible for direction, and very slow time constants are used in the velocity neuron, this difference in τ could explain a stereo-typed reflex-like trajectory that overshoots and comes back. However, the τs evolved in motor neurons show a general trend towards minimal τ, irrespective of the condition or the function of the motor neuron.

The second possible explanation was that the minimal reaction time t_r in the task is a function of the sensory delay $t_r(d) = t_n + d$ (where t_n is the controller-internal reaction time) and that the networks would optimise velocity in order to use this minimal reaction time to localise the centre of the object (6 units). Were this the case, v should be such that $t_n = 6/v - d$ is near constant across evolved networks. Calculating this value as a function of the evolved velocities, however, several orders of magnitude of variation between and within networks evolved for both conditions result. This means that there is a lot of variation as regards the time occupied to arrive at the centre, and that selection pressure does not operate to optimise velocities in the described way.

The third and last possible explanation tested was whether the shortening of the absolute time window in which to solve the task by 250 ms in the trials with delay makes a difference and gives the networks evolved without delay more freedom to deviate further from the centre before focusing. However, testing the networks evolved without delay with faster object velocities to compensate for this difference in time window led only to a marginal (5.6%) decrease in performance. There seems to be no simple answer for the question why there is a discrepancy in velocities for agents with and without delay, even if the answer may well be a combination of several of these simple factors tested.

10.3 Summary

The model generates a number of insights into the task and the constraints it imposes on the strategy space, which are in the following tested against the human data from the experiment presented in chapter 9. Most noticeably, the model shows that, given the task

design, it is impossible to solve the task because the possibilities for predicting object location are too limited. The reasons for which the model fails to exhibit the kind of adaptation processes expected provide further insights into the sensorimotor requirements for delay adaptation. Time pressure had been introduced to force agents (and subjects) to adapt to the systematic delays, rather than to just compensate by slowing down. However, in the experiment/simulation designed, time pressure was so fierce and sensory information was so impoverished that the only possible and sub-optimal way to solve the task is to perform a ballistic movement without online control once the object is first perceived. The effort to minimise task complexity to its absolute basics has taken us one step too far. The following chapter 11 expands on these theoretical issues about necessary sensorimotor complexity for delay adaptation.

However, other insights gained about possible strategies and agent behaviour may also help to better understand the human data. The model shows that the coarse fitness function is unable to register subtle systematic displacements that result from a shortening of sensorimotor latencies as unsuccessful behaviour. These displacements can be seen as analogues of a behavioural after-effect of adaptation to sensory delays that is not reflected in the task performance. Movement velocity could be shown to play a role in explaining differences in displacement magnitude that result in differences in performance. The simulation model suggests that a similar behavioural adaptation, undetected by the performance criterion, could have occurred in the human participants, too. In this sense, the model predicts that the participants in the experiment should *overshoot* their target when the delay is introduced, that this overshooting decreases over training, and that they should *stop earlier* when the delay is removed. As part of the findings on systematic displacements, the model predicts that velocities decrease between pre- and post-test.

Another factor the simulation suggests for analysis is that the behaviour should be reflex-like. From the simulation we expect that, since a delay prolongs the absolute temporal duration of a closed sensorimotor loop for online control (from perception, to action, to perception), the strategies become more reflex-like over training with delays. It is not straight forward to define or measure whether movement is reflex-like or not in such a simple task. A measure explored in the data analysis below is the intra-participant self-similarity of trajectories. The simulation also predicts that systematic displacements should be more pronounced in strategies identified as reflex-like in this sense.

The evolved agent controllers have very simple strategies that rely on only few sensorimotor invariances. Factors that do not matter to evolved strategies are the velocity of the

objects (catch as fast as you can), the history of previous object presentations, the exact magnitude of the tactile input and the auditory reward signal. These factors are assumed to be irrelevant in the following analysis as a consequence (some support for this assumption is presented in (Rohde, 2008)).

10.4 Revisiting the Human Data

The combined experimental and modelling study is presented here as an example of how minimalist behavioural experiments with humans and minimal ER simulation modelling can be combined and mutually inform each other. This section revisits the human data to test whether there is evidence that the human failure to exhibit the hypothesised adaptation to tactile delays relies on similar processes and factors as the analogous failure of evolved agents (i.e., it tests the occurrence of systematic displacements, stereo-typedness of trajectories and a decrease in velocity after adaptation).

As a first step, the motion data was re-structured. Human movement of a computer mouse is not a symmetrical behaviour (due to arm morphology); at least some subjects appeared to use different strategies for catching an object they approached from the left than they did for catching an object they approached from the right. Therefore, subjects' attempts to catch an object were separated into left and right attempts (according to movement direction before first contact with the object) and analysed separately, as if they were generated by a different person (even if, in the following analysis, the data is sometimes again collapsed, assuming approximate symmetry of strategies).

The data from different object presentations was segmented and normalised in time with respect to the point and moment of first contact. Assuming simple sensorimotor strategies, i.e., either simple reactive feedback circuits or ballistic stereo-typed trajectories, other factors, such as the velocity of the objects or the history of previous catch attempts, were not taken into consideration. With this kind of normalisation, i.e., location and time of first tactile contact when approaching from one direction, reflex-like ballistic movement should be exactly congruent, like in the simulation, whereas reactive behaviour should be contingent on the ongoing sensory flow.

Data was filtered, removing first those catch attempts in which contact with the object was not established or in which the participants did not move (possibly due to outside events distracting their concentration). For the remaining data, average trajectories were calculated using a cumulative average of change in position Δp between each two points of measurement (measurements every 20 ms, missing or irregular data points were filled

in by linear interpolation). Using an iterative method, outliers were eliminated if the mean squared error of the average trajectory was larger than three standard deviations σ from the average trajectory. This quantity is referred to in the following analysis as *MSE* of a trajectory P

$$MSE(P) = \frac{1}{(T-1)} \sum_{t=1}^{T-1} ((p(t+1) - p(t)) - (p_{mean}(t+1) - p_{mean}(t)))^2 \qquad (10.3)$$

where $[1,T]$ is the sequence of measurements (taken every 20 ms) for which all trajectories in a set are defined (different lengths and overlapping parts), p is the position relative to the position of first touching the object and P_{mean} is the sequence of changes in position that characterises the average trajectory. After removing the data for one participant, because remaining data was sparse, the processed data sets per subject and movement direction contained on average 14 trajectories from which the average trajectories were calculated, and none contained less than five.

This normalised movement data allows to test for systematic displacements and differences in velocity. Furthermore, the average trajectories and the mean squared deviation from it across trials (*MSE*), which was used to eliminate outliers, can also be used in order to measure and compare stereo-typedness of trajectories (see Sect. 10.4.2).

10.4.1 *Systematic Displacements*

The main expectation derived from the simulation model is that a clear negative after-effect occurs in terms of changes in systematic relative displacements between the centre of the receptive field and the centre of the object to be caught at the end of an object presentation that depend on initial movement direction and velocity. Due to the coarseness of the performance criterion (Eq. (9.1)), if such systematic displacements are small enough in magnitude, they are not necessarily reflected in catch-performance, which could explain the lack of support for the main hypothesis.

Displacements were calculated by the distance $d_{t\prime}$ of the receptive field at the end of an object presentation from the exact object centre. We studied the change in displacement across the different phases: $(d_{t\prime}^{pre} - d_{t\prime}^{delay}), (d_{t\prime}^{delay} - d_{t\prime}^{adap}), (d_{t\prime}^{adap} - d_{t\prime}^{post})$. Displacements were multiplied by the sign of initial movement direction. The simplifying assumption here is that, independent of movement strategy, overshooting corresponds to a displacement in the direction of movement, whereas stopping early manifests as a displacement in the opposite direction. This assumption, which is not valid in a more general context, is justified by the fact that the time pressure encourages ballistic reflex-like catch motion (cf. Sect. 10.4.2)

and only affords limited possibilities for online correction or more complex sensorimotor transformations. The main prediction then is that subjects should overshoot the goal when the delay is introduced and and that the opposite change should occur first during the adaptation phase (error correction) and then in the post-test (undershooting). This translates to the expectation that: $\text{sign}(d_{t\prime}^{pre} - d_{t\prime}^{delay}) = -\text{sign}(d_{t\prime}^{delay} - d_{t\prime}^{adap}) = -\text{sign}(d_{t\prime}^{adap} - d_{t\prime}^{post})$.

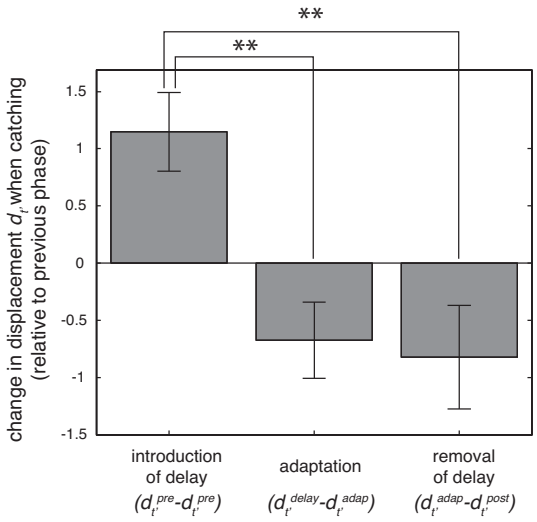

Fig. 10.7 Change in systematic displacements from the object centre across the phases of the experiment (pooled for subjects; n=19; errorbars: standard error of the mean).

This prediction is supported by the collapsed data from left and right attempts (in Cochran's Q on the signs of displacement: $p = 0.001$; in a repeated measures ANOVA on the changes in displacement from phase to phase with time as factor: $F(2,36) = 10.68; p < 0.0002$).[1] Figure 10.7 shows the average change in displacement across the phases. Pairwise comparison confirms that the significant differences in this comparison are that $(d_{t\prime}^{pre} - d_{t\prime}^{delay})$ is different in both sign and in value from the changes occurring in the other two phases $(d_{t\prime}^{delay} - d_{t\prime}^{adap}), (d_{t\prime}^{adap} - d_{t\prime}^{post})$, all $p < 0.02$. However, $(d_{t\prime}^{delay} - d_{t\prime}^{adap})$ and $(d_{t\prime}^{adap} - d_{t\prime}^{post})$ are not significantly different, as hypothesised.[2]

[1] Displacements were multiplied by the sign of the initial movement direction, assuming symmetry of strategies. However, running the same tests on the non-collapsed data (i.e., for both movement directions separately, in case they are not symmetrical) confirms all these effects, such that the collapsed data is presented for simplicity.

[2] In the dissertation on which this book is based (Rohde, 2008), the numbers presented, as well as the conclusions, are slightly different. This is partially due to a computational mistake in calculating the systematic displacements, and partially due to the application of unsuitable statistical tests (paired t-tests). The latter mistake in statistical testing also concerns the other variables investigated below, but, in these other cases, there is no difference in the main results if the correct test is used.

This confirmation of the prediction generated from the model suggests that subtle adaptation effects, undetected in the performance measure, also occur in the human participants. Note, however, that the difference in absolute displacement from the centre between pre- and post-test ($d_{t\prime}^{pre} = -0.28,\ d_{t\prime}^{post} = -0.63$) is not significant ($p = 0.29$). Therefore, an alternative and equally valid interpretation of the data is that the shift in displacements induced by the delays is only partially compensated during training and that removing the delay implies a return to the initial strategy, which corresponds to another decrease in displacement. In this interpretation, the non-significant additional displacement would be merely the result of noise. However, given that the other variables identified also change in the ways predicted by the simulation (see analysis below), it is not unreasonable to assume that theoretical insights gained from the model can be applied to and tested in humans, because both may undergo the same kind of transformations in sensorimotor behaviour.

10.4.2 *Stereotyped Trajectories*

For the pre-processing and filtering of the data, the inter-participant average of trajectories during each phase of the experiment and the mean squared deviation *MSE* (see Eq. (10.3)) from these mean trajectories had been computed. $MSE(P)$ can be taken as a measure for reflex-like or ballistic strategies: given that the trajectories were normalised with respect to the moment and location of first perceptual contact, perfectly stereo-typed trajectories would be exactly congruent ($MSE = 0$), whereas more reactive or variable trajectories would be contingent on ongoing sensory flow (high *MSE*). The model predicts that trajectories should get more reflex-like over training with delays, as a consequence of decreased possibility for online control.

A repeated measures ANOVA of the variation ($\ln(MSE)$) of trajectories with experimental phase as factor confirms that this is the case for the experimental data ($F(3,57) = 4.51; p < 0.007$). Figure 10.8 shows how variability $\ln(MSE)$ decreases across the phases of the experiment. Pairwise comparison of the values between each condition show that the significant reduction takes place during training with delays and is maintained on that level during the post-test (all $p < 0.05$, compare also Fig. 10.8). Again, the computations had been performed on the collapsed data, assuming symmetry of trajectories from left and right catching attempts. Analysis of the data separated for left and right attempts confirms the general effects.

There are a number of problems associated with using $\ln(MSE)$ as measure for stereotypedness. Firstly, the *MSE* measure is affected by the accidental application of the mouse

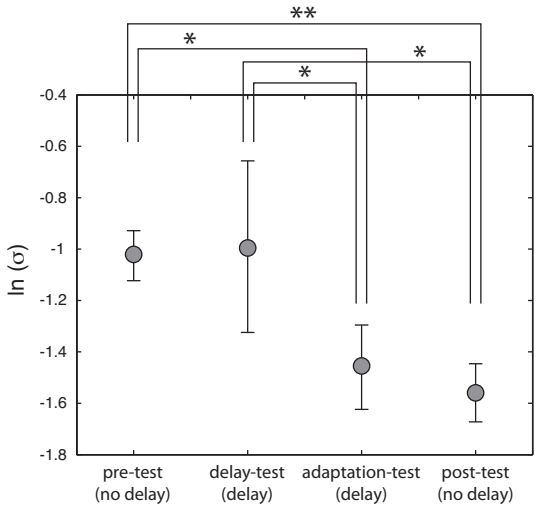

Fig. 10.8 Logarithm of the *MSE* from mean trajectories throughout the phases of the experiment. This change towards more stereotyped behaviour happens during training with delays.

acceleration function to the human movement data, as the *MSE* is computed on the basis of changes in space for each measurement interval. Arguably, this would be the same for any measure of stereotypedness. Also, the number of trajectories used to calculate P_{mean} and their respective length could possibly play a role in the effect but had not been controlled for. Thirdly, this measure does not include a way to quantify in how far trajectories with a high *MSE* are not just variable but instead *reactive*, i.e., the sensory flow is not considered in any way other than that the onset is normalised to the moment of first stimulation. Cross-correlation had been applied in order to explore the role of the sensorimotor flow, not just the motion, but the results had not been very indicative. Despite these limitations, the differences found in the *MSE* give further evidence that it may be reasonable to draw analogies between the simulation model and the human data.

10.4.3 *Velocity*

In the evolved agents, the magnitude of systematic displacements is dependent on initial movement velocity. This difference in magnitude impacts on performance, thus explaining why no negative after-effect of removing delays is measured. This section investigates whether human subjects also decrease the velocity of their catching behaviour. Even though the previous chapter 9 already investigated velocity and could not find significant effects, it only looked at general average velocity, not specifically at velocity before contact. As

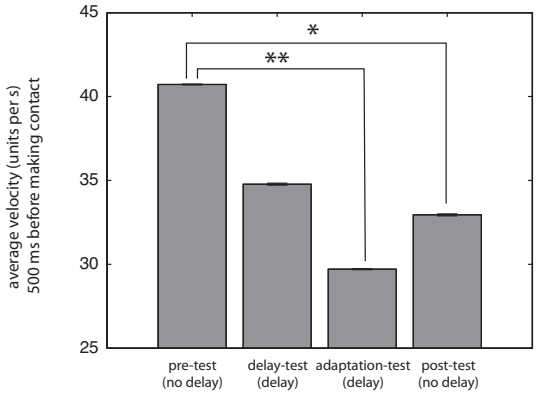

Fig. 10.9 Average velocity before making contact with the object to be caught across the phases of the experiment. The direction of change in velocity is rather variable across subjects (comparably weak main effect despite very low errors).

previously remarked, due to the unintended application of mouse acceleration, the change in position from which velocity is computed is not an accurate measurement of real mouse velocity, but the distortion of spatial data is deemed negligible in the analysis.

Velocity $\bar{v}$ was computed by the mean absolute difference in distances covered per measurement interval of 20 ms during the last 500 ms before touching the object. Again, velocity was computed on the basis of the collapsed data from left and right approaches, as absolute velocities were used, i.e., direction did not figure in. Figure 10.9 displays how this value changes across the different phases of the experiment. A repeated measures ANOVA on the velocity with phase as factor shows that the main effect of change in velocity is significant ($F(3,57) = 3.87; p = 0.0137$). Pairwise comparison shows that, as predicted by the simulation model, there was a significant decrease in initial velocity that took place during training ($p = 0.0003$) and that was carried over to the post-test ($p = 0.0217$). This finding adds another confirmed prediction of the human data from the ER simulation to the set.

10.5 Discussion

The simulation model has generated a number of insights about the sensorimotor dynamics of the catch task used to study adaptation to tactile delays. Most significantly, it points out that the strategies afforded by the given sensorimotor task only allow stereo-typed ballistic catch movements or, in exceptional cases, minimal reactive online control. What this means

in the context of time cognition, perceived simultaneity and the possibility to recalibrate is discussed in the following chapter 11.

On a more practical level, the simulation has generated a number of descriptive concepts and variables along which the human behavioural data can be analysed to test whether human subjects really are subject to the same kind of processes and factors. Following up to the presentation of the simulation model, some of these factors have been tested in the human data, confirming the predictions from the model.

- *Systematic Displacements* have been found to follow the direction of change suggested by the model (i.e. increase when the delay is introduced, decrease during adaptation and decrease even further when the delay is removed). The difference in absolute position of these displacements comparing pre-test and post-test shows a trend into the expected direction (i.e., stopping too early, at a larger distance from the object centre), but this trend is not significant. Therefore, it is not clear if the systematic displacements are just a sign of the restoration of the original situation and strategy or evidence for humans following patterns found in evolved agents.
- As an approximation of the *stereotypedness or reflex-likeness* of strategies, the intra-subjective similarity of trajectories, measured as the logarithm of the mean square error from the average trajectory $\ln(MSE)$ could be shown to follow the pattern predicted, i.e., to decrease significantly during training with the delays. This decrease entails a significant decrease from pre-test to post-test. This measure relies on a number of assumptions, such as that the ongoing flow of sensory information is irrelevant for identifying a tendency towards ballistic movements.
- Concerning the movement velocity, the prediction that velocity before making contact with an object would decrease between pre- and post-test is confirmed. However, it is not clear what this decrease in velocity implies, since its functional role in the simulation model is unclear as well.

In many senses, this analysis is rather crude, compared to dynamical analyses such as those provided by (Beer, 2003). As stated previously, the presented application of simulation results to the data serves as an example how the combined behavioural-experimental and ER simulation modelling approach proposed can work. Given that the data does not support the main experimental hypothesis, it seems unreasonable to spend more energy in analysing the simulation model or the human data.

In order to do justice to the emphasis that the enactive approach places on closed-loop interaction, in a different situation, further dynamical analysis in the closed sensorimotor loop should have been conducted. Sceptics of embodied approaches frequently find it difficult to imagine what such an analysis would look like. There are indeed no simple recipes about how such an analysis should be undertaken yet and the tools for analysis in many senses still need to be developed – even tools to explain the simulation models in the first place. In principle, however, the possibilities for analysing either the simulated or the human data are open-ended, and there are vast possibilities to be inspired by other approaches. It is important to recall that the enactive approach is a change in perspective, a paradigm, not a radically new method, different from anything before. To name but a few examples, for analysing the evolved controllers (Beer, 2003) provides a paradigm case. In terms of analysing time series and physiological data, simple measures, such as cross-correlation, can be applied, as well as more sophisticated relational measures such as Granger causality (e.g., Seth and Edelman, 2007). In terms of extracting sensorimotor invariances, a lot can be gained from ecological approaches (e.g., Gibson, 1979; Lee, 1998). Any tool from any science with similar data can, in principle, be used to describe phenomena that are interesting from an enactive view. The enactive approach really has to be seen as a new paradigm rather than as a new method.

Chapter 11

Perceived Simultaneity and Sensorimotor Latencies

What can be learned from the behavioural experiment and the simulation model presented in the previous two chapters about the question of delay adaptation and recalibration of perceived simultaneity (Sect. 9.1)? Can we derive a new experimental hypothesis from the study, a new experimental paradigm? How can the failure to reproduce the hypothesised effects reported by (Cunningham *et al.*, 2001a) be integrated into the more general picture of embodied time perception and time cognition? This chapter evaluates the empirical and the simulated data in this larger context. After providing a concise summary in Sect. 11.1, Sect. 11.2 proposes an interpretation that ties in with the conceptual analysis of temporality in general given in chapter 8.

11.1 Summary of the Results

The experimental paradigm was inspired by (Cunningham *et al.*, 2001a)'s findings on semi-permanent adaptation to visual delays that lead to a negative after-effect in task performance and, anecdotally, to a recalibration of perceived simultaneity. The author's hypothesis that inherent time pressure in the task is the necessary and sufficient factor for yielding such an interesting adaptation effect, which distinguishes their experiment from similar earlier studies, was tested in this experiment. The experimental study follows the minimalist approach described in chapter 3. The visuomotor avoidance task used by Cunningham *et al.* was simplified, turned into a catch task and transferred to the audio-tactile platform Tactos (Gapenne *et al.*, 2003) in order to be more tractable, controllable and suited for dynamical analysis. The objective was to first reproduce the adaptation effect reported by Cunningham *et al.* in a minimalist set-up and to then identify the defining factors that bring the effect to break-down. Thus, the sensorimotor basis of experienced simultaneity and presentness should have been elucidated by describing and analysing not only the qualitative

adaptation of performance, but also the changes in sensorimotor dynamics and strategy that bring it about.

The main hypothesis tested in the experiment was that the participants' performance profile would follow the same pattern as reported in (Cunningham *et al.*, 2001a). This is, a decrease of initial performance level upon introduction of delay, full or partial recovery over training with delays, and a decrease of performance as compared to the initial performance levels once the delay is removed. This hypothesis is not supported by the data. There is no significant recovery of performance with training and no significant after-effect. In this sense, the results are closer to those obtained in earlier studies, in which subjects slowed down their movement to compensate cognitively for the sensory delays, yielding only partial compensation for the delays. Such compensatory strategies do not produce negative after-effects. The repeated failure to produce semi-permanent adaptation had led (Smith and Smith, 1962) to conclude that delay adaptation is impossible in principle.

The agents evolved in the ER simulation model of the experiment presented in chapter 10 were similarly inapt of exhibiting the expected performance profile. Analysis of the strategies evolved led to insights about the sensorimotor properties afforded by the task. Most importantly, it came out that the time pressure, which had been implemented to catalyse delay adaptation, in reality restricts viable strategies to ballistic reflex-like catch movements, in which shortening or lengthening of sensorimotor latencies manifests as a systematic displacement of the agent with respect to the object it should catch (overshooting when delay is introduced and stopping too early when it is removed). Three variables, i.e., these systematic displacements, a reduction in variability of trajectories as an indicator of ballistic movements and a reduction of velocity before touching an object, were investigated in the human data to test if human behaviour is subject to similar factors and constraints. This *post hoc* analysis guided by ER simulation modelling gave evidence that there may be adaptation processes of the same kind, i.e., spatial modulation of ballistic movements. The after-effect that this adaptation produces does not lead to a decrease in performance, because the performance criterion is spatially not accurate enough to pick up such subtle modulations.

In principle, the insights gained about the discrepancy between behavioural adaptation effects and task performance should make it possible to design a better experiment in which these variables concur. In simulation, using a fitness function that is spatially more exact had exactly this effect, i.e., a negative after-effect occurred. Such a modification of the original experimental paradigm is not feasible because the temporal sampling rate and

the spatial resolution have fierce limits imposed by the fact that the experimental platform needs to work in real-time.

However, even if these technical limitations could be mitigated, the analysis of the human and the simulated data propose a different direction for further experimentation. The key issue is that the behavioural compensation obtained was not of the hypothesised kind. The spatial modulation of ballistic movements that agents (and possibly humans) adopt in order to compensate for the delays is very specific to the experimental set-up. It would not work as a delay compensation technique in a ecologically more sophisticated scenario. The following section analyses this further-reaching question that relates back to the conceptual insights gained in chapter 8.

11.2 The Sensorimotor Basis of Present-Time

In an attempt to explain the failure to produce delay adaptation in certain scenarios but not others, on the basis of different effects that sensory delays have in different sensorimotor tasks, we have proposed a classification of sensorimotor feedback loops into *reactive*, *reflex-like* and *anticipatory* (cf. Rohde and Di Paolo, 2007). *Reactive* feedback loops are those in which the motor output is, at any point in time, a result of the most recent sensory input (i.e., such strategies do not rely on internal state). Phototaxis in a Braitenberg vehicle (Braitenberg, 1984) like agent is the paradigm example of a reflex-like strategy. As discussed in chapter 10, the circuits evolved in the model of the experimental catch task are not reactive but instead ballistic and *reflex-like*. The motor output that defines an action is only sensitive to stimulus onset, and are not sensitive to the moment to moment variation of stimulus magnitude as the movement unfolds. A third class of sensorimotor feedback loops called *'anticipatory'* is marked by the characteristic that motor outputs depend on both, the moment-to-moment sensory flow and the history of previous interactions (as internal state). This distinction should not be seen as bindingly formal, even though it could possibly be formalised. However, these concepts should capture our intuitive understanding of different kinds of strategies and how they are affected by certain sensorimotor perturbations on a functional level, and a formal account of state-sensitivity of behaviour runs into danger of not adequately capturing such dependencies. On this level, any real (or evolved) sensorimotor behaviour is likely to be *more* or *less* reactive, reflex-like or anticipatory, not strictly a member of one of these classes.

The important point is that sensory delays play different functional roles in these different kinds of behavioural feedback loops. In a reactive sensorimotor loop, a sensory delay

implies that behaviour has to be slowed down. If action relies on online correction on the basis of current sensory state, increased sensorimotor latencies mean that the agent has to wait, in order to sense the outcome of its previous action. A Braitenberg vehicle with increased sensorimotor latencies will turn past the light source, correct, overshoot again and thus start oscillating, unless it has means to slow down its movement to compensate for the perturbation. This kind of behaviour is reminiscent of subject's behaviour in (Smith and Smith, 1962)'s outline-drawing task.

Such overshooting as a consequence of sensory delays in a reactive sensorimotor loop, which results in oscillatory movement, is behaviourally very similar to what we experience as the consequence of an increase in inertia, e.g., when driving a larger car or when canoeing. The way we compensate for an increase in inertia, on a day to day basis, is to slow down. Therefore, in a reactive sensorimotor loop, a delay manifests as a discrepancy that is akin to the much more ecologically common increase in inertia. From such an ecological perspective, it appears logical to adopt the same compensation strategy – particularly, if it is successful in mitigating the suffered perturbation. To an extent, this was already recognised by (Cunningham *et al.*, 2001a) and led them to conjecture that negative after-effects did not occur in previous studies because it was possible to compensate by slowing down.

In reflex-like behavioural loops, such as those evolved in the artificial agents, a delay does not manifest in a discrepancy akin to increased inertia, but, instead, to a discrepancy akin to a fixed spatial offset, whose magnitude depends on the self-movement velocity before contact. Just as increases in inertia, spatial offsets are ecologically much more common than prolonged sensorimotor latencies. Therefore, the compensatory strategy adopted is the one suitable for dealing with displacements, i.e., to produce an inverse systematic displacement. As intended, time pressure in the task made it impossible to compensate to the delay by slowing down, treating the delay as an increase in inertia. However, in the paradigm studied, a different way to avoid real delay adaptation was afforded, i.e., spatial counter-displacement.

Both reactive and reflex-like strategies allow to conceptualise the experienced discrepancy induced by the delay as something different and more common. We are very used to compensate for increases in inertia in the described ways that do not produce negative after-effects. The proposal here is that this is what happened in paradigms that report a failure to adapt to delays in reactive tasks (e.g., Smith and Smith, 1962; Kennedy *et al.*, 2009; Thompson *et al.*, 1999; Ferell, 1965). The experience of delays as displacements in reflex-like behaviours, by contrast, induce semi-permanent adaptation of behaviour (i.e., inverse

spatial displacement of motion). These displacements (stopping earlier) as negative after-effects occurred in the simulation model and, arguably, as well in the human subjects. However, these after-effects are not the ones we are after, as the recalibration obtained was one of space, not of time.

It makes sense that no recalibration of experienced simultaneity occurs if more likely alternative compensatory techniques are possible – if the delay is not experienced as a delay in the first place, it cannot cease to be experienced as a delay over training. Which brings us to the – somewhat constructivist – question of what characterises a delay in an ecological context and how is it different, in terms of sensorimotor contingencies, from a displacement or an increase in inertia. The difference between an increase in inertia and a delay is that, in a high inertia system, it is impossible to change the direction of movement fast. In a system with sensorimotor delays, on the other hand, the possibility to change movement direction fast is still given – only the possibilities for fast online control of such fast behaviour is eliminated. For a delay not to be conceptualised as an increase in inertia, time pressure is thus indeed a necessary component that brings this difference to the subject's attention. Only under time-pressure will the subject realise that he can still change direction of movement fast.

Time pressure was implemented using fast object velocities in the experiment presented, which, indeed, suppressed reactive strategies. However, subjects used instead ballistic stereo-typed catch movements, a strategy, in which a delay manifests as an offset, not as a delay, and, as a consequence, did not lead to the hypothesised patterns of delay adaptation. The difference is that, unlike in the minimal task used here, (Cunningham *et al.*, 2001a)'s task affords the possibility to exercise anticipatory control. Their visual task forces subjects to produce fast sequences of motion with variations in velocity and direction, during which the regular structure of the visual environment has to be exploited continually. Only in the presence of such longer term structural links between perception and action that are directly relevant for the online modulation of behaviour, delay adaptation in the strong sense is possible and required. Such anticipatory behaviour is, however, only possible if the signal is sufficiently structured and allows anticipation over a longer time-course, which was not the case in our experiment. There needs to be a cohesion between momentary signal structure, own movement possibilities, and future signal structure. This new hypothesis about a combination of requirements for delay adaptation, i.e., time-pressure combined with the possibility for longer term anticipation, needs to be empirically tested, in an experiment that may well be another simplified version of the task used in (Cunningham *et al.*, 2001a),

but not simplified to the point that it is impossible to register the delay as a delay in the first place.

What does this analysis teach us about the sensorimotor dynamics of time perception? What is the role of sensorimotor latencies in constituting the experience of present, past and future? Some tentative ideas about the length of sensorimotor loops (i.e., the time it takes, from the observer perspective, for a sensation to result in motion and then again in sensation) and their role in defining primitive past, present and future are now developed in the light of the cross-disciplinary analysis in chapter 8. They link to the use of spatial and temporal language in the Aymaran language and the role of knowledge and agency in their conceptual *time is space* metaphor (cf. Sect. 8.5).

When a sensorimotor loop is enacted, i.e., the causal chain from sensation to action back to sensation, it is extended in time from the observer perspective. However, this time extension is not *a priori* known to the agent itself. This is because the causal chain of executing this sensorimotor loop is not something the agent itself can still interrupt or influence in a controlled way, whereas, the observer, can. In the light of what it means for something to be past or to be future (cf. chapter 8), this period, in which I await the confirmation of the expected outcome of my action, does not qualify for either. What happens during the execution of a sensorimotor loop is neither past, if we recall that the past is what is known, done and unchangeable, nor is it future, as the future is still open to volitional change. Therefore, the time it takes for a sensation to be transformed into a motion that again leads to a sensation is the present, jammed between the past and the future in the just-mentioned sense. As soon as the subject's expectations are matched by the reafferent sensation, this present turns into past just like any previous sensations. Once the causal chain is initiated by the agent, it loses its own possibility to further influence what happens, but it is only once the reafferent signal arrives that external forces are equally unable to interfere with the agent's expectation of the outcome of its actions.

The tricky thing is that, at any moment in time, infinitely many such sensorimotor loops, continuous in time, are being realised. The diagram depicted in Fig. 11.1 tries to capture this idea, in which subjective time, from the observer perspective, takes the form of a tube. In this tube, change of variables in the eye of the observer forms the x-axis and the helices running around this tube are the causal sensorimotor-loops, also in the eye of the observer. The experience of subjective presentness is then a chunk of this tube, a chunk that advances in discrete overlapping steps, which gives rise to the discrete flow of chained events.

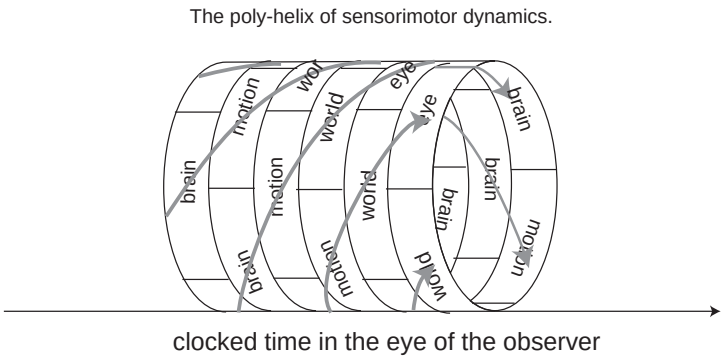

Fig. 11.1 Illustration of ideas on the relation between temporal experience and sensorimotor loops from the observer's perspective.

These ideas are not yet fully developed and it is not clear how they could be tested empirically. However, both (Libet, 2004)'s and (Cunningham *et al.*, 2001a)'s counterintuitive results on disruptions of experienced presentness make sense in this view. In Libet's experiments, the 500 ms that elapse between a peripheral stimulus and the build up of correlated cerebral activity, as well as the 500 ms between the Readiness Potential and the onset of movement, form part of a sensorimotor loop in the process of completion, outside the subject's volitional control. Therefore, these 500 ms, time-extended in the eye of the observer, do not exist from the subject perspective in a temporal sense. They are neither future (changeable), nor past (confirmed truth). Only through the use of technology, they can come into meaningful existence, as it was done in Walter's experiment (as reported in Dennett and Kinsbourne, 1992) by cutting short the inherent sensorimotor latencies. This short-cut induced a breakdown of perceived ownership of the action in the subjects.

Similarly, in (Cunningham *et al.*, 2001a)'s experiment, by imposing a delay, the sensorimotor loop was stretched such that the extra 200 ms to await visual feedback became a meaningless time span and was therefore banned from temporal experience. This corresponds to inflating the tube of temporal experience depicted in Fig. 11.2. Through the anticipatory nature of the task, it was brought to the subject's attention that the time-span during which it is still possible to intervene with the course of events was shortened, which was not the case in the present and in previous studies, which allowed a different more ecologically plausible conceptualisation of the discrepancy induced by the delay as increase in inertia. If the delay is removed, the tube is shrunk and the subject is suddenly afforded an extra 200 ms to influence the unfolding course of events. This shortening of sensorimotor

latencies is experienced as an inversion of the temporal order associated with causal chains, reminiscent of the one reported by Walter.

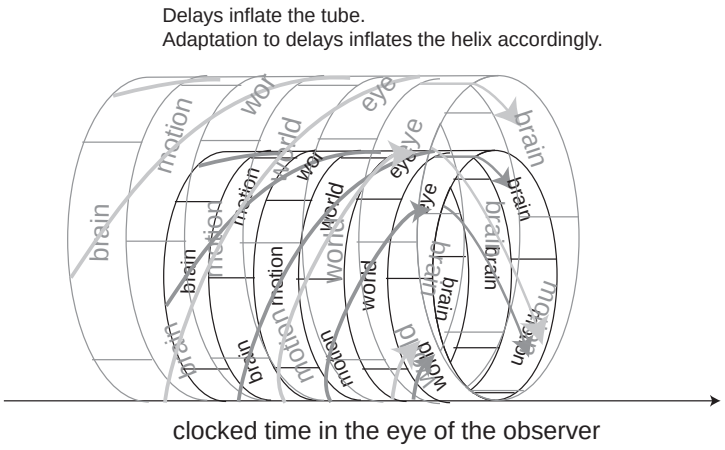

Fig. 11.2 Illustration of how adaptation to increased sensorimotor latencies may change experienced simultaneity (inflation of the tube sketched in Fig. 11.1).

Linking the results to (Varela, 1999)'s neurophenomenology of present-time consciousness, it is worthwhile pointing out that both the visual delay of 200 ms used by Cunningham *et al.* and the tactile delay of 250 ms used in our experiment are at the intersection between the time scales associated with the primitive and the immanent flow of time. Possibly, for that reason the delays are perceptible, yet can still be integrated into experienced presentness, and this would not be true for delays of arbitrary length. Neurophysiology poses constraints on the construction of reality; the tube in Fig. 11.2 cannot be inflated indefinitely. Evidence to support this assumption comes from (Cunningham *et al.*, 2001b)'s experiment on visual delays in a driving simulator. Comparing adaptation to a 130 ms, a 230 ms and a 430 ms delay, they only found the kind of effect reported in (Cunningham *et al.*, 2001a) in the condition with a 230 ms delay. This suggests that the 130 ms delay is too small to be registered and the 430 ms delay is too large to be integrated.

A last issue to be mentioned here, which has already been addressed implicitly, is the role of unity and causality in delay adaptation. The disruption of temporal experience in (Cunningham *et al.*, 2001a)'s experiment surprises participants because multi-modal aspects of one unified action are temporally torn apart. What is experienced visually (reafference) precedes what is experienced proprioceptively (movement). Therefore, the experiential effect is not actually the distortion of temporal order between two simultaneous temporal

object-events (different in space but identical in time), but the disruption of unity (in both time and space). The question to be asked is whether, without the destruction of perceived unity, the same surprise would have occurred. In other words, it is possible that such illusory reversals of experienced temporal order much more frequent than we think, between separate external objects or events. However, if such a reversal does not coincide with a break-down of our perceived temporal self-cohesion, we do not detect such inconsistencies, because they are irrelevant to us. Such inherently meaningful determinants of perception in the world is difficult to explain from a computationalist perspective. In such a view, causal links and temporal relations would be inferred constantly and automatically, indiscriminate of the basis of structured inputs. However, more work is necessary to turn these ideas into a model or to derive hypotheses that can be tested against empirical data.

Chapter 12

Outlook

This book set out to mould out a space for computational methods within the enactive paradigm in cognitive science. It promotes simulation models, not as thinking machines, but as machines for thinking. It presents case studies, to give concrete examples of how simple simulation models can contribute to the explanation of mind, without the accompanying claim that they would be minds themselves, embedded in the context of conceptual methodological debate, making explicit the shift of perspective that marks the enactive approach, which frequently results in asking the unusual and non-obvious questions. Hopefully, even if the reader does not want to go all the way with me, he or she now understands the characteristics of the enactive paradigm and its assets in terms of scientific explanation. This last chapter summarises and evaluates the presented collection of facts, ideas and results and returns to the methodological theme of the book: can a post-cognitivist science of human level cognition be informed by simple ER simulation models? What can such simple models contribute, what is their role in scientific explanation? Section 12.1 summarises the material presented in this book, Sect. 12.2 evaluates them before the concluding remark in Sect. 12.3.

12.1 Summary

In trying to move beyond the paradigmatic struggle in cognitive science, this book promotes and develops the enactive approach to cognition and behaviour. For historical reasons, the metaphor of cognition as computation is closely tied to the idea of the interdisciplinary and scientific study of mind and cognition, in particular, if it involves the use of computer models. Over the past decades, however, the computational metaphor turned out to be empirically limited and conceptually harmful. The enactive approach rejects this metaphor in favour of an embodied, situated, dynamical and constructivist perspective inspired by the

metaphor of the living organism as cognitive system. It focuses on autonomous dynamics on several emergent levels of biological organisation, on experience and on the genuine meaningfulness of mind and mindful behaviour. Chapter 2 summarises this debate and points out the differences between the enactive approach and other alternative approaches in cognitive science, many of which are related to the enactive approach. This chapter also identifies the big challenges that Enactivism faces in the coming decades. Processes of abstract, symbolic and high-level cognition count as representationalist strongholds and pose the biggest challenge to the enactive paradigm to demonstrate its explanatory potential. How can minimal ER simulation modelling as a technique for enactive cognitive science be used to elucidate any aspects of human cognition and behaviour, and particularly those identified as representationalist strongholds?

From there, the repertoire of methods underlying the research in this book (ER simulation modelling, CTRNN controllers, DST analysis and PS experiments) are introduced in chapter 3. This chapter also sees an extensive debate on methodological issues, such as the role of the scientist as an observer in constructivist approaches, on the scientific value of ALife simulation models and on the possibility of scientifically studying experience by combining first, second and third person methods in a non-reductionist fashion. Crucially, this chapter also develops the interdisciplinary methodological framework that was put to use in some of the work presented, i.e., the application of ER simulation modelling to minimalist experiential and experimental research on perception and sensorimotor adaptation (PS research). This kind of research is only truly interdisciplinary if modelling, experimentation and subjective experience are brought together, in a mutually informative polylogue (see Fig. 12.1). The results presented in the subsequent chapters highlight individual parts of this diagram.

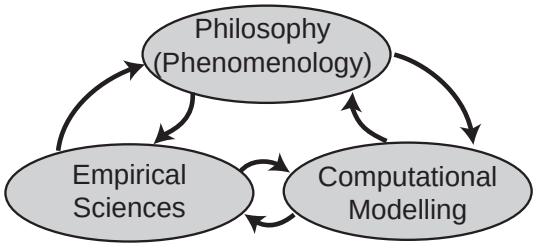

Fig. 12.1 Illustration of the interdisciplinary enactive framework proposed.

Chapter 4 presents a model of directional reaching in an idealised human arm to investigate the principle of linear synergies in motor organisation. This model shows that imposing this kind of constraint on a motor system can enhance evolvability. This benefit is not just due to the smaller search space: reducing the task to two dimensions means an equal decrease of parameter space but has the opposite effect on evolvability. Concerning the diagram illustrating the scientific role of simulation modelling (Fig. 12.1), this study successfully implements the links between simulation modelling and the experimental sciences.

The simulation model of value system architectures presented in chapter 5 investigates a research question of a much more abstract and philosophical nature, i.e., it illustrates logical problems with a certain type of neural or cognitive architecture and points out the implicitly held modelling assumptions underlying such approaches. The model criticises *a priori* semantics of dedicated meaning-generating modules to supervise life-time learning in an embodied context. Such models have been proposed as solution to problems encountered with more rigid, fully disembodied approaches. The ER simulation points out how such meaning generating modules, if no further processes that ensure their intact functioning are provided, are unlikely to explain adaptivity as a general phenomenon. This model demonstrates the mutual methodological links between philosophical theory building and simulation modelling in the diagram in Fig. 12.1.

The following two simulation models on perceptual crossing in a one-dimensional (chapter 6) and a two-dimensional (chapter 7) simulated environment apply ER modelling to experiments in PS. The simulation models contribute to the understanding and interpretation of the experiments on different levels, generating concrete predictions about descriptive variables involved in perceptual distinction or about morphological aspects of observed behaviour, but also providing abstract proofs of concept about dynamical principles at work and implicit premises held by experimenters or subjects. Given that the experimental PS approach addresses both questions of perceptual experience and simple sensorimotor behaviour, these models succeed at implementing all four methodological links between simulation modelling and the other disciplines in Fig. 12.1.

Chapter 8 provides a conceptual interlude on the issue of time perception and time cognition that brings together material from a variety of sources and disciplines. It concludes with a refined view of dimensions along which time can be studied (levels of temporal experience, methodological approaches and their scopes and limits) to prepare for the subsequent study on delay adaptation and recalibration of experienced simultaneity.

This interdisciplinary project, which directly combines behavioural experiments with humans and ER simulation modelling is presented in chapters 9-11. The experiment tests the hypothesis that recalibration of perceived simultaneity results from adaptation to sensory delays in simple sensorimotor tasks, provided that these tasks are marked by a time pressure that forces subjects to move fast. This hypothesis, which is based on (Cunningham *et al.*, 2001a), is not supported by the data presented in chapter 9. Guided by the ER simulation model of the task presented in chapter 10, the behavioural data is analysed to search for reasons for this failure. On the basis of both the simulated and the human data, chapter 11 extends the tested hypothesis, proposing that delay adaptation does not only require time pressure, but also temporal structure in the sensorimotor interactions with the environment that allow anticipation. This extended hypothesis is based on an ecological analysis of the effect of delays in reactive, reflex-like and anticipatory sensorimotor loops, which concludes that only in the latter the delay will really manifest as a delay. Returning to the more abstract, conceptual and general view on time cognition given in chapter 8, it is proposed that the present-time experience corresponds to the sensorimotor behaviour currently enacted, over which an agent does not exercise volitional control.

In this project on experienced simultaneity, all mutual links in the diagram in Fig. 12.1 are active: work done using all three methods – the empirical, the computation and the conceptual – was conducted by the same person (myself, the author – albeit with the help of experienced collaborators). This acid test of the methodological framework proposed in chapter 3 helps to point towards the merits and demerits of this approach, an evaluation that is performed in the following section.

12.2 Evolutionary Robotics Simulations for a Post-cognitivist Science of Mind

The material presented in this book, in its diversity, has hopefully convinced the sceptical reader that simple simulation models have merit for the study of human level cognition and behaviour, even if he or she may not want to go along with each and every of the claims brought forward. The subsequent evaluation focuses on three core issues more profoundly: the question of the recognition and incorporation of simulation results in empirical science (Sect. 12.2.1), the question of advancing the enactive paradigm by conquering representationalist strongholds (Sect. 12.2.2) and a more detailed critique of the interdisciplinary framework proposed in chapter 3 on the basis of the results presented (Sect. 12.2.3).

12.2.1 *Reception in the Scientific Community*

In a recent provocative article, (Webb, 2009) attacks 'animat' modelling, i.e., simple agent research that does not explicitly link its results to empirical phenomena, and questions the scientific (biological) relevance of this kind of approach, a view which is debatable (cf. Rohde, 2009). After all, the model on value system architectures presented in chapter 5 can be seen as exactly the kind of conceptual, theory-driven approach to modelling that she criticises.

However, it cannot be denied that, in the field of ALife, there is the potential danger that relevant results get lost in a nexus. A simulation model may be inspired by a real biological phenomenon, it may then model this phenomenon, and even generate useful results, both for other synthetic approaches and for the scientific domain studying the phenomenon that inspired the model. However, many times the results do not receive the attention and acknowledgement they deserve.

Fortunately, the simulation models presented in this book have not fallen victim to this trend. Both groups working on motor synergies that had inspired the model presented in chapter 4 were very positive about the model, encouraged us to keep up the work and cited the simulation research as a consequence (Shemmell *et al.*, 2007). The models of perceptual crossing in a one-dimensional and a two-dimensional simulated environment have been well received by the CRED group who have conducted the original study and cited the work as a relevant contribution (Auvray *et al.*, 2009). The simulation results from the one-dimensional variant were published in a domain-specific (i.e., a psychological) journal (Di Paolo *et al.*, 2008). The later models were conducted in direct collaboration with the empirical researchers (Lenay, Rohde & Stewart, in preparation; interdisciplinary study of delay adaptation, chapters 9-11). Those models that were also not published to a wider audience (i.e., the value system model and the study on adaptation to delays), naturally, did not produce the same kind of resonance in the relevant scientific communities.

Supporting (Webb, 2009) at least in some parts of her criticism, it is important to mention that such positive responses do not come for free. It requires work to apply simulation results to real-world phenomena, to identify concrete predictions and relevant conceptual insights and to communicate those to the relevant communities. The encouraging sympathy with which the work presented in this book was received suggests that the time and effort to do so are well spent.

12.2.2 *Representationalist Strongholds*

In chapter 2, high-level, abstract and symbolic domains of cognition have been identified as representationalist strongholds – as scientific problem areas for which enactive approaches are still struggling to generate powerful results, models and explanations. In how far did the research presented in this book contribute to the invasion of representationalist strongholds? The answer to this question already stumbles over the problem that what is considered high-level and low-level from either perspective overlaps, but is not *a priori* congruent. In a representationalist view, high-level cognition is the kind of symbol manipulation performed in the most decoupled and homuncular modules that are furthest away from the sensory and motor periphery. In an enactive perspective, it is not fully clear how high or low-level cognition should be defined other than in phylogenetic advances, new forms of value generation and more mediated meanings emerging from new levels of autonomous self sustaining dynamics (see chapters 2 and 5). From this embodied perspective, peripheral systems of the organism can be equally essential for explaining a high-level cognitive capacity as cortical brain areas.

The work on motor synergies (chapter 4) would be considered more low-level from both perspectives. From the representationalist perspective, it is low-level because it is concerned with the realisation of motion, not with motor planning or reasoning. From the enactive perspective, it is low-level because the processes described are not embedded in a meaning generating context, if investigated by themselves. This is not to say that the study of principles in motor control is irrelevant for cognitive science. As argued in the introduction (chapter 1), our human cognition, our concepts and experiences, probably relies much more on such simple processes of sensorimotor self-organisation than traditional approaches acknowledge. However, in order to be able to make claims about higher levels of cognition, the role that simple motor self-organisation plays in our mental lives or in enabling our logical capacities has to be explicitly addressed.

The model of value system architectures (chapter 5), by contrast, dives straight into questions of neural organisation and its role in realising general purpose adaptivity. TNGS associates value system function with neural populations in the limbic system and the brainstem, which are phylogenetically old brain regions that are traditionally not linked to higher cognitive function. However, given the generality of the criticism, which applies to proposals of localised *a priori* semantics in hybrid and semi-homuncular approaches in general, the model is relevant for both higher and lower levels of cognition. Probably, representationalist and enactivist researchers would agree on this matter. However, since

the point is so general, the model is largely a conceptual model. It criticises a certain type of cognitive architecture, but has nothing to put in its place. This failure to provide concrete and empirically testable ideas is a shortcoming of many simple ER simulation models. Considering cognition as a global and dynamically complex phenomenon, you lose the benefit of a representationalist perspective to add simple functional models as building blocks, assuming they interact linearly. By doing justice to the possibility of nonlinear interactions, the enactive modeller faces a trade-off between the applicability of the model to a concrete real-world phenomenon and the generality of the question it can address. The model of value system architectures is an example of a more general but less applicable, theory-driven approach to modelling.

Concerning the models of simple behavioural experiments with humans that use the methods of psychophysics and the psychology of perception, the question of low-level vs. high-level is more complicated. From a computational perspective, this kind of research is about 'just perception', i.e., the generation of internal representations for cognition to work on, a process that is not deemed cognitive itself. Empirically, whenever this strict separation breaks down, i.e., when factors other than stimulus energy or directly measurable physical or peripheral-physiological variables impact on perceptual judgment behaviour, a black box labelled 'attention' or 'higher-level process' is invoked for explanation. Remnants of this computationalist division of cognition and 'just perception' have snuck into the work here presented as well, such as in the distinction between true perceptual learning that produces a negative after-effect (semi-permanent) and 'cognitive' adjustment that does not lead to such after-effects. Is this use of language not in tension with the aspiration of this book to tackle questions of high-level human cognition? As usually, from the enactive perspective things are not that simple. The point is not to debate that a conceptual distinction between perception and cognition in terms of reasoning is useful in many situations. The point is to question their strict separation, to emphasise that there is a continuum in both mechanism and function. What is further questioned is the assumption that explaining the perceptual bits is the easy part, whereas explaining the 'cognitive' bits is the real problem. The conjecture put forward in this book is that once self-organisation of perceptuo-motor invariants and the physical and social constraints on our sense-making are explained, the reflexive symbolic processes that build on such ongoing coping, and which manifest in our abstract cognition and conscious awareness, will fall into place naturally.

In this sense, both the research on agency perception in minimalist environments and the study on recalibration of experienced simultaneity, though low-level from a computational-

ist perspective, can be seen as progress on problems on high-level human cognition. Both perceptual phenomena are of a highly abstract nature, in the sense that the meaning involved is highly mediated, i.e., very far away from the physical form of the stimulus (cf. Sect. 5.5). Both are important factors in how we subjectively experience our worlds. By proceeding, step by step, on the explanation of such abstract mental phenomena, a more coherent, complete and parsimonious picture will be gained. In this sense this book has seen a shift in focus, away from those cognitive phenomena that computationalists set as goal-posts for enactive accounts and towards the kind of phenomena that computationalist approaches struggle to account for, which they tend to downplay and ignore, but which are equally pressing: open-ended meaning generation, the construction of time and space, participatory sense-making – all these are problems that traditional AI and autonomous agent models struggle with. The construction of self and Body Image could be added to this list of core problems in traditional AI (cf. Rohde and Ikegami, 2009). Advancing on 'representation-hungry' problems of symbol use, reflexivity and image making remain as challenges for the enactive approach, but there is no need to have our pace and our focus dictated by the sceptics. A shift towards an enactive perspective entails asking questions differently and attending to non-obvious problems.

12.2.3 *Simulating Human Perceptual Behaviour*

The research presented as part of this book activated increasing numbers of conceptual links in the diagram in Fig. 12.1, which coincided with an increase in methodological novelty. The models of motor synergies and value system architectures (chapters 4 and 5) were strictly in the spirit of previous ER simulation models, and the scientific role of such models has been analysed extensively (e.g., Harvey *et al.*, 2005; Di Paolo *et al.*, 2000; Beer, 1996). On the other hand, the application of ER modelling to psychophysics or PS research, closely matching the experiment and model, following the agenda laid out in Sect. 3.6, is novel. This section evaluates the application of the approach. Special attention is paid to questions of the experiential dimension of the work and the interdisciplinary polylogue.

As concerns the aim to include equal proportions of all three disciplinary bubbles in Fig. 12.1, concerning the work presented here, the experiential dimension has been neglected. When the research was conducted, the idea of perceptual judgements as first or second person methods for a crude neurophenomenology (cf. Sect. 3.5) had not yet been fully developed. However, it is easy to envision research along the same lines that puts these ideas to good use.

Another issue to mention is that the interdisciplinary study on adaptation to delays had also tested the assumption that it is necessary or beneficial for one and the same person to realise all tasks involved in the interdisciplinary polylogue depicted in Fig. 12.1 herself and in parallel. The underlying assumption was that performing both the experiment and implementing the model in person would lead to a much closer interaction between the two, such that modelling and scientific practice would constantly mutually inform each other and keep growing alongside one another. In practice, however, this was not the case. There were phases of work that were strictly dedicated to modelling and others strictly dedicated to experimentation, and the application of one to the other (i.e., the 'communication' of results) was not always working. Work on methodologically different aspects of a complex project requires different mindsets, which can only be exercised at the same time to a limited extent. This insight resonates with the classical idea of the hermeneutic circle of understanding of a text described, for instance, by (Gadamer, 1994), in which understanding is advanced by alternating phases of closure and prejudice, from the global perspective, and phases of thorough investigation of detail, in which our ideas are open to change (see Fig. 12.2).

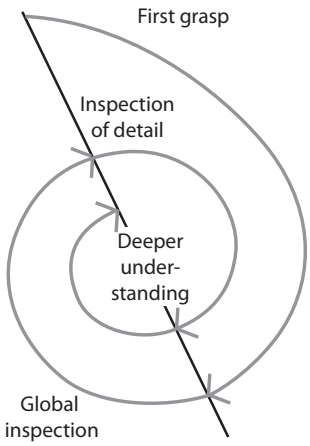

Fig. 12.2 Illustration of the hermeneutic circle of understanding.

The analogy is a bit flawed as both simulation and experimental planning/measurement are relevant in both the global and local phases of understanding. However, a similar diagram can be used in order to illustrate how a phase of modelling can aid experimental design, be pushed into the background during piloting, return to the foreground for elaboration

of the set-up, become irrelevant during conduction of the experiment, but later aid interpretation of the results, *etc.* Therefore, the benefits (if any) of performing all tasks in the interdisciplinary framework in person, as in the study on delay adaptation, as opposed to contributing with simulation modelling to existing experimental research, as in the modelling of perceptual crossing, are only of a quantitative nature, not of a qualitative nature. This means that in a well managed collaboration with working communication the kind of simulation modelling proposed can be equally effective. In Sect. 3.6, the computationalist approach was criticised for being multidisciplinary, rather than genuinely interdisciplinary. Effectively, this means that the collaborative demands are comparably higher in enactive approaches, not that the individual scientists need to be generalists.

12.3 Conclusion

This book starts by recalling the long gone optimism of the early days of AI and computationalist cognitive science. It finishes with the appraisal of a new optimism of a dawning era, the era of enactive, embodied and dynamical cognitive science. Work from across disciplines and areas that forms part of this movement was presented, both own and other, indicating avenues for future research and pointing out gaps in the methodological inventory, which wait to be filled.

The enactive view is not a simple view, one that paints black and white. Problems that look deceptively simple reveal their true complexity under the enactive scrutiny. The global perspective on the conceptual level is in stark opposition to the simplicity of the ER models presented as case studies in this book, as well as the modesty about their scientific function or explanatory potential. However, it is exactly this modesty that allows the enactivist to think big without becoming delusional about what it is that can be feasibly achieved. It may be useful to write in a grant proposal that our computers will soon be our best mates and our robots indistinguishable from our pets. However, it is not satisfactory on a personal level, if what you care for is understanding that we have minds – that we *are* minds. Bold claims about our possible achievement are not true in a strict sense. The loose sense, in which they are true (i.e., computers will sometimes make us smile or our robots can take the role of a pet if we are happy to endorse the illusion) does not help us to advance on the real issue, for the reasons given throughout this book.

Computer models in the enactive approach are not thinking machines, like in the computationalist approach, they are machines for thinking, like in any other of the natural sciences. One important point that was not addressed at depth in this book should be briefly

mentioned here in this outlook: giving up the dream of the thinking machine (thinking computer) does not imply giving up the dream of the synthesis of intelligent or cognitive systems. Maybe, one day, we will be able to synthesise a system that is genuinely cognitive. After all, we are all naturalists. But this system, a product of the hard work of many inspired scientists, studying what it is about our organisation as organisms that makes us cognitive, is not going to be a software program or a computer or a machine in the strict sense, and it will not do abstract decoupled information processing. Probably, this system would be an artificial organism of some sort, involving chemical, energetic or other real physical processes that are spatially extended, dynamically embedded and whose meaning would be intrinsic.

Appendix A

List of Abbreviations and Symbols

α_i	Joint angle of i^{th} joint
a_i	Activation of unit n_i
AI	Artificial Intelligence
ALife	Artificial Life
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BBR	Behavior-Based Robotics
C, c_{ij}	Network connectivity matrix in which $c_{ij} \in \{0,1\}$ indicates existence of a connection from unit n_j to unit n_i
CCNR	Centre for Computational Neuroscience and Robotics, University of Sussex
CPG	Central Pattern Generator
CRED	Cognitive Research and Enaction Design Group, Université de Technologie de Compiègne
CTRNN	Continuous-Time Recurrent Neural Network
δ, Δ	Parameters of RBF (see chapter 4)
d	Delay (of sensory inputs to CTRNN controller)
$d(x)$	A distance function (locally defined)
DS, DST	Dynamical System, Dynamical System Theory
DoF	Degree-of-Freedom
ε	Noise or a very small constant (locally defined)
ER	Evolutionary Robotics
ϕ	Required pointing direction signal (see chapter 4)
$F(i)$	Fitness function for individual i, performance for participant i
FLE	Flash-Lag-Effect in psychophysics

GA	Genetic Algorithm
GOFAI	Good-Old-Fashioned Artificial Intelligence
h	Simulation time step
I_i	External input to n_i
k_i	A constant
$K(\phi)$	Linear synergy function (see chapter 4)
M_i	Motor signal
M_G	Motor gain
MSE	Mean Square Error
n_i	The i^{th} unit (neuron) in an ANN/CTRNN
ω	Angular velocity
ODE	Open Dynamics Engine (C++ library)
PDP	Parallel Distributed Processing
PS	Perceptual Supplementation
r	Magnitude of vector mutation in GA
RBF, RBFN	Radial Basis Function, Radial Basis Function Network
σ	Standard deviation
$\sigma(a)$	Standard logistic (sigmoidal) function (Eq. (3.3))
S_i	Sensory signal
S_G	Sensor gain
θ_i	Bias of unit n_i
τ_i	The time constant of decay of a_i
t, T, t_0	t = time, T = length of task, t_0 = initial/reference time
TM	Turing Machine
TNGS	Theory of Neuronal Group Selection
TVSS	Tactile Visual Sensory Substitution
v	velocity
W, w_{ij}	Network weight matrix in which w_{ij} gives the connection weight from unit n_j to unit n_i
x^*	Fixed point or steady state activity (numerically established) of variable x in a DS

Bibliography

Allen, J. (1984). Towards a general theory of action and time, *Artificial Intelligence* **23**, pp. 123–154.

Amedi, A., Stern, W., Camprodon, J. A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P. and Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex, *Nature Neuroscience* **10**, pp. 687 – 689.

Arbib, M. (1981). Perceptual structures and distributed motor control, in V. Brooks (ed.), *Handbook of Physiology*, Vol. II, Motor Control, Part 1 (American Physiological Society), pp. 1449–1480, section 2: The Nervous System.

Ashby, W. (1954). *Design for a Brain* (Chapman and Hall Ltd., London).

Auvray, M., Lenay, C. and Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment, *New Ideas in Psychology* **27**, pp. 79–97.

Bach-y Rita, P., Collins, C., Sauders, F., White, B. and Scadden, L. (1969). Vision substitution by tactile image projection, *Nature* **221**, pp. 963–964.

Bach-y Rita, P., Tyler, M. and Kaczmarek, K. (2003). Seeing with the brain, *Int. J. Human-Computer Interaction* **15**, pp. 285–295.

Baird, J. and Noma, E. (1978). *Fundamentals of Scaling and Psychophysics* (John Wiley & Sons, New York), Wiley Series in Behavior.

Barandiaran, X. (2007). Mental Life: Conceptual models and synthetic methodologies for a post-cognitivist psychology, in B. Wallace, A. Ross, T. Anderson and J. Davies (eds.), *The World, the Mind and the Body: Psychology after cognitivism* (Imprint Academic), pp. 49–90.

Barandiaran, X., Di Paolo, E. A. and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry and spatio-temporality in action, *Adaptive Behavior* **17**, pp. 367–386, special issue on Agency in Natural and Artificial Systems.

Barandiaran, X. and Ruiz-Mirazo, K. (2008). Modelling autonomy: simulating the essence of life and cognition, *BioSystems* **91**, pp. 295–304, editorial introduction to the special issue.

Beer, R. (1995). On the dynamics of small Continuous-Time Recurrent Neural Networks, *Adaptive Behavior* **3**, 4, pp. 469–509.

Beer, R. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior, in P. Maes, M. Mataric, J. Meyer, J. Pollack and S. Wilson (eds.), *From Animals to Animats 4* (MIT press), pp. 421–429, URL `citeseer.ist.psu.edu/article/beer96toward.html`.

Beer, R. (2000). Dynamical approaches to cognitive science, *Trends in Cognitive Sciences* **4**, pp. 91–99.

Beer, R. (2003). The dynamics of active categorical perception in an evolved model agent, *Adaptive Behavior* **4**, 11, pp. 209–243.

Beer, R. (2006). Parameter space structure of Continuous-Time Recurrent Neural Networks, *Neural Computation* **18**, pp. 3009–3051.

Bernstein, N. (1967). *The Coordination and Regulation of Movements* (Pergamon, Oxford), Russian original published in 1935.

Bertschinger, N., Olbrich, E., Ay, N. and Jost, J. (2008). Autonomy: An information theoretic perspective, *BioSystems* **91**, 2, pp. 331–45, special issue on modelling autonomy.

Bitbol, M. (1988). The concept of measurement and time symmetry in quantum mechanics, *Philosophy of Science* **55**, pp. 349–375.

Bitbol, M. (2001). Non-representationnalist theories of cognition and quantum mechanics, *SATS (Nordic journal of philosophy)* **2**, pp. 37–61.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology* (MIT Press, Cambridge, MA).

Brooks, R. (1991). Intelligence without reason, in J. Myopoulos and R. Reiter (eds.), *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence, San Mateo, CA* (Morgan Kaufmann), pp. 569–595.

Cantwell-Smith, B. (1996). *On the Origin of Objects* (MIT Press, Cambridge MA).

Chalmers, D. (1995). Facing up to the problem of consicousness, *Journal of Consciousness Studies* **2**, pp. 200–220.

Chrisley, R. (2003). Embodied Artificial Intelligence, *Artificial Intelligence* **149**, pp. 131–150.

Churchland, P. M. and Churchland, P. S. (1998). *On the Contrary: Critical Essays 1987-1997* (MIT Press, Cambridge MA).

Clark, A. (1997). *Being there: Putting brain, body, and world together again* (MIT Press, Cambridge MA).

Clark, A. (1998). Time and mind, *The Journal of Philosophy* **95**, 7, pp. 354–376.

Clark, A. and Grush, R. (1999). Towards a cognitive robotics, *Adaptive Behavior* **7**, pp. 5–16.

Cliff, D. (1991). Computational Neuroethology: A provisional manifesto, in J. Meyer and S. Wilson (eds.), *Proc 1st Int. Conf. on Simulation of Adaptive Behaviour: From Animals to Animats* (MIT Press, Cambridge MA), pp. 29–39.

Cole, P. (ed.) (1981). *Radical pragmatics* (Academic Press).

Cunningham, D., Billock, V. and Tsou, B. (2001a). Sensorimotor adaptation to violations of temporal contiguity, *Psychological Science* **12**, pp. 532–535.

Cunningham, D., Chatziastros, A., von der Heyde, M. and Bülthoff, H. (2001b). Driving in the future: Temporal visuomotor adaptation and generalization, *Journal of Vision* **1**, 2, pp. 88–98, URL `http://www.journalofvision.org/1/2/3/`.

Cunningham, D., Kreher, B., von der Heyde, M. and Bülthoff, H. (2001c). Do cause and effect need to be temporally continuous? Learning to compensate for delayed vestibular feedback, Abstract, *Journal of Vision* 1(3): 135a.

Dassonville, P. and Bala, J. K. (2004). Perception, action, and roelofs effect: A mere illusion of dissociation, *PLoS Biol* **2**, p. e364.

Dassonville, P., Sanders, T. and Capp, B. (2009). The rod-and-frame and simultaneous tilt illusions: Perception, action and the two-wrongs hypothesis, Abstract at the Annual Meeting of the Vision Sciences Society VSS 2009, *Journal of Vision*.

De Jaegher, H. (2007). *Social Interaction Rhythm and Participatory Sense-Making. An Embodied, Interactional Approach to Social Understanding, with Implications for Autism*, Ph.D. thesis, Department of Informatics.

Dennett, D. (1985). *Elbow Room: The Varieties of Free Will Worth Wanting* (Clarendon Press).

Dennett, D. (1989). *The intentional stance* (MIT Press, Cambridge MA).

Dennett, D. C. and Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain, *Behavioral and Brain Sciences* **15**, pp. 183–201.

Di Paolo, E. (2000). Behavioral coordination, structural congruence and entrainment in acoustically coupled agents, *Adaptive Behavior* **8**, pp. 27–47.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency, *Phenomenology and the Cognitive Sciences* **4**, 4, pp. 429–452.

Di Paolo, E. and Iizuka, H. (2008). How (not) to model autonomous behaviour, *BioSystems* **91**, pp. 409–423, special issue on modelling autonomy.

Di Paolo, E., Noble, J. and Bullock, S. (2000). Simulation models as opaque thought experiments, in *Artificial Life VII: The Seventh International Conference on the Simulation and Synthesis of Living Systems, Reed College, Portland, Oregon, USA, 1-6 August (Proceedings)* (MIT Press, Cambridge, MA), pp. 497–506.

Di Paolo, E., Rohde, M. and De Jaegher, H. (forthcoming). Horizons for the enactive mind: Values, social interaction, and play, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Di Paolo, E., Rohde, M. and Iizuka, H. (2008). Sensitivity to social contingency or stability of interaction? Modelling the dynamics of perceptual crossing, *New Ideas in Psychology* **26**, pp. 278–294, special issue on Dynamics and Psychology.

Dienes, Z. and Seth, A. (forthcoming). The conscious and the unconscious, in G. Koob, R. F. Thompson and M. Le Moal (eds.), *Encyclopedia of Behavioral Neuroscience* (Elsevier).

Doya, K. (2002). Metalearning and neuromodulation, *Neural Networks* **15**, pp. 495–506.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason* (Harper & Row).

Eagleman, D. and Sejnowski, T. (2002). Untangling spatial from temporal illusions, *Trends in Neurosciences* **25**, p. 293.

Eagleman, D. M., P.U., T., Janssen, P., Nobre, A. C., Buonomano, D. and Holcombe, A. O. (2005). Time and the brain: how subjective time relates to neural time, *Journal of Neuroscience* **25**, pp. 10369–71.

Edelman, G. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic Books, New York).

Edelman, G. (1989). *The Remembered Present: A Biological Theory of Consciousness* (Basic Books, New York).

Edelman, G. (2003). Naturalizing consciousness: A theoretical framework, *Proc Natl Acad Sci USA* **100**, pp. 5520–5524.

Egbert, M. and Di Paolo, E. (2009). Integrating autopoiesis and behavior: An exploration in computational chemo-ethology, *Adaptive Behavior* **17**, pp. 387–401, special issue on Agency in Natural and Artificial Systems.

Ehrenstein, W. and Ehrenstein, A. (1999). Psychophysical methods, in U. Windhorst and H. Johansson (eds.), *Modern techniques in neuroscience research* (Springer, Berlin, Heidelberg), pp. 1211–1241.

Elman, J. (1998). Connectionism, Artificial Life, and Dynamical Systems: New approaches to old questions, in W. Bechtel and G. Graham (eds.), *A Companion to Cognitive Science* (Basil Blackwood, Oxford).

Evans, V. (2004). How we conceptualise time: Language, meaning and temporal cognition, *Essays in Arts and Sciences* **XXXIII**, pp. 13–44, issue theme: time.

Eysenck, M. and Keane, M. (2000). *Cognitive Psychology: A student's handbook*, 4th edn. (Psychology Press, Hove).

Fechner, G. (1966). *Elements of Psychophysics. Volume I* (Holt, Rinehartand Winston, Inc., New York), translated by H. E. Adler. Edited by D.H. Howes and E. G. Boring. German original published in 1860.

Ferell, W. (1965). Remote manipulation with transmission delay, *IEEE Trans, hum. Factors Elect.* **HFE-6**, pp. 24–32.

Fodor, J. (2000). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology* (MIT Press, Cambridge MA).

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis, *Cognition* **28**, pp. 3–71, special issue on Connections and Symbols, edited by S. Pinker and J. Mehler.

Fröse, T. (2007). On the role of AI in the ongoing paradigm shift within the Cognitive Sciences, in M. Lungarella (ed.), *50 Years of AI* (Springer), pp. 63–75.

Fröse, T. and Di Paolo, E. A. (2008). Stability of coordination requires mutuality of interaction in a model of embodied agents, in M. Asada, J. C. T. Hallam, J.-A. Meyer and J. Tani (eds.), *From Animals to Animats 10: Proc. of the 10th Int. Conf. on the Simulation of Adaptive Behavior*, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg), pp. 52 – 61.

Fujisaki, W., Shimojo, S., Kashino, M. and Nishida, S. (2004). Recalibration of audiovisual simultaneity, *Nature Neuroscience* **7**, pp. 773 – 778.

Gadamer, H.-G. (1994). *Truth and Method* (Continuum, New York), translated by J. Weinsheimer and D.G. Marshall. German original published in 1960.

Gallagher, S. (2005). *How the Body Shapes the Mind* (Clarendon Press, Oxford, UK).

Gapenne, O. (forthcoming). Kinaestheses and the construction of perceptual objects, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Gapenne, O., Rovira, K., Ali Ammar, A. and Lenay, C. (2003). Tactos: Special computer interface for the reading and writing of 2D forms in blind people, in C. Stephanidis (ed.), *Universal Access in HCI: Inclusive Design in the Information Society* (Lawrence Erlbaum Associates, London), pp. 1270–1274.

Gepshtein, S. and Kubovy, M. (2007). The lawful perception of apparent motion, *Journal of Vision* **7**, pp. 1–15.

Gibbon, J. and Church, R. (1984). Sources of variance in an information processing theory of timing, in H. Roitblat, T. Bever and H. Terrace (eds.), *Animal Cognition* (Lawrence Erlbaum, Hillsdale, NJ), pp. 465–488.

Gibson, J. (1982). The problem of temporal order in stimulation and perception, in E. Reed and R. Jones (eds.), *Reasons for Realism: Selected Essays of James J. Gibson* (Lawrence Erlbaum, Hillsdale NJ), pp. 171–179, originally published in the *Journal of Psychology*, 1966, 62, 141-149.

Gibson, J. J. (1979). *The ecological approach to visual perception* (Houghton Mifflin, Boston).

Gottlieb, G., Song, Q., Almeida, G., Hong, D. and Corcos, D. (1997). Directional control of planar human arm movement, *Journal of Neurophysiology* **78**, pp. 2985–2998.

Grossberg, S. and Paine, R. (2000). A neural model of corticocerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements, *Neural Networks* **13**, pp. 999–1046.

Grush, R. (2007). Skill theory v2.0: dispositions, emulation, and spatial perception, *Synthese* **159**, pp. 389–416.

Harnad, S. (1990). The symbol grounding problem, *Physica D* **42**, pp. 335–346.

Harvey, I. (1996). Untimed and misrepresented: Connectionism and the computer metaphor, *AISB Quarterly* **96**, pp. 20–27.

Harvey, I., Di Paolo, E., Wood, R., Quinn, M. and Tuci, E. A. (2005). Evolutionary Robotics: A new scientific tool for studying cognition, *Artificial Life* **11**, 1-2, pp. 79–98.

Haugeland, J. (1981). Semantic engines: An introduction to mind design, in J. Haugeland (ed.), *Mind design: Philosophy - Psychology - Artificial Intelligence* (MIT Press, Cambridge MA), pp. 1–34.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea* (MIT Press, Cambridge, MA).

Havelange, V. (forthcoming). The ontological constitution of cognition and the epistemological constitution of cognitive science: Phenomenology, enaction and technology, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Hebb, D. (1949). *The Organization of Behavior. A Neuropsychological Theory* (John Wiley & sons inc., New York).

Heidegger, M. (1963). *Sein und Zeit*, tenth unmodified edn. (Max Niemeyer Verlag, Tübingen), original published in 1927.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar* (Blackwell, Oxford).

Held, R., Efstathiou, A. and Greene, M. (1966). Adaptation to displaced and delayed visual feedback from the hand, *Journal of Experimental Psychology* **72**, pp. 887–891.

Herzog, M. (2007). Spatial processing and visual backward masking, *Advances in Cognitive Psychology* **3**, pp. 85–92.

Hinton, G. and Nowlan, S. (1987). How learning can guide evolution, *Complex Systems* **1**, pp. 495–502.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor).

Holland, O. (2002). Grey Walter: The imitator of life, in R. Damper and D. Cliff (eds.), *Biologically-Inspired Robotics: The Legacy of W. Grey Walter. Proceedings of the EPSRC/BBSRC International Workshop WGW-02*, pp. 32–48.

Hurlburt, R. and Schwitzgebel, E. (2007). *Describing Inner Experience? Proponent Meets Skeptic* (MIT Press, Cambridge MA).

Hurley, S. (1998). *Consciousness in Action* (Harvard University Press, London).

Hurley, S. and Noë, A. (2003). Neural plasticity and consciousness, *Biology and Philosophy* **18**, pp. 131–168.

Hutchins, E. (forthcoming). Enaction, imagination, and insight, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Iizuka, H. and Di Paolo, E. (2007). Minimal agency detection of embodied agents, in F. Almeida e Costa, L. Rocha, E. Costa, I. Harvey and A. Coutinho (eds.), *Proceedings of the 9th European Conference on Artificial Life*, Lecture Notes in Artificial Intelligence (Springer, Berlin, Heidelberg), pp. 485–494.

Iizuka, H. and Ikegami, T. (2004). Adaptability and diversity in simulated turn-taking behavior, *Artificial Life* **10**, pp. 361–378.

Ikegami, T. and Suzuki, K. (2008). From a homeostatic to a homeodynamic self, *BioSystems* **91**, pp. 388–400, special issue on modelling autonomy.

Ivry, R. and Schlerf, J. (2008). Dedicated and intrinsic models of time perception, *Trends in Cognitive Sciences* **12**, pp. 273–280.

Izquierdo-Torres, E. and Harvey, I. (2007). Hebbian learning using fixed weight evolved dynamical 'neural' networks, in H. Abbass, M. Bedau, S. Nolfi and J. Wiles (eds.), *Proceedings of the First IEEE Symposium on Artificial Life*, Series on Computational Intelligence (IEEE, Honolulu, Hawaii), pp. 394–401.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, *Science* **304**, 5667, pp. 78–80.

James, W. (1890). *The Principles of Psychology*, Vol. 1 (Holt and Macmillan, New York, London), chapter 15: The Perception of Time. Retrieved: 14.03.2008 from 'Classics in the History of Psychology. An internet resource developed by Christopher D. Green. URL: http://psychclassics.asu.edu/James/Principles/prin15.htm.

Jonas, H. (1966). *The phenomenon of life: Towards a philosophical biology* (Northwestern University Press, Evanston, IL).

Kandel, E., Schwartz, J. and Jessel, T. (eds.) (2000). *Principles of Neural Science*, 4th edn. (McGraw-Hill, New York).

Kant, I. (1974). *Kritik der reinen Vernunft, Wissenschaft, die drei Kritiken*, Vol. 1 (Suhrkamp, Frankfurt a. M.), edited by W. Weischedel. Original published in 1781.

Karmarkar, U. R. and Buonomano, D. V. (2007). Timing in the absence of clocks: encoding time in neural network states, *Neuron* **53**, pp. 427–438.

Kelso, S. (ed.) (1982). *Human Motor Behavior: An Introduction* (Lawrence Erlbaum, Hillsdale, NJ).

Kennedy, J., Buehner, M. and Rushton, S. (2009). Adaptation to space and to time, Abstract at the Annual Meeting of the Vision Sciences Society VSS 2009, *Journal of Vision*.

Kirsh, D. (1991). Today the earwig, tomorrow man? *Artificial Intelligence* **47**, pp. 161–184.

Kohler, I. (1962). Experiments with goggles, *Scientific American* **206**, pp. 62–72.

Krichmar, J. and Edelman, G. (2002). Machine psychology: autonomous behavior, perceptual categorization and conditioning in a brain-based device, *Cereb. Cortex* **12**, pp. 818–30.

Kurthen, M. (1994). *Hermeneutische Kognitionswissenschaft. Die Krise der Orthodoxie* (DJRE Verlag, Bonn).

Lakoff, G. and Johnson, M. (2003). *Metaphors We Live By* (University of Chicago Press), with an afterword; originally published in 1980.

Lakoff, G. and Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being* (Basic Books, New York).

Langton, C. (ed.) (1997). *Artificial Life: An overview*, first mit press paperback edn. (MIT Press, Cambridge MA).

Le Van Quyen, M. (forthcoming). Neurodynamics and phenomenology in mutual enlightenment: The example of the epileptic aura, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Le Van Quyen, M. and Petitmengin, C. (2002). Neuronal dynamics and conscious experience: An example of reciprocal causation before epileptic seizures, *Phenomenology and the Cognitive Sciences* **1**, pp. 169–180.

Lee, D. (1998). Guiding movement by coupling taus, *Ecological Psychology* **10**, pp. 221–250.

Lenay, C. (2003). Ignorance et suppléance : La question de l'espace, HDR, Université de Technologie de Compiègne.

Lenay, C., Gapenne, O., Hanneton, S., Marque, C. and Genouëlle, C. (2003). Sensory Substitution: Limits and perspectives, in Y. Hatwell, A. Streri and E. Gentaz (eds.), *Touching for Knowing* (Benjamins Publishers, Amsterdam), pp. 275–292, chapter 16, English translation.

Levine, J. (1983). Materialism and qualia: The explanatory gap, *Pacific Philosophical Quarterly* **64**, pp. 354–61.

Li, W. and Matin, L. (2005). Two wrongs make a right: linear increase of accuracy of visually-guided manual pointing, reaching, and height-matching with increase in hand-to-body distance, *Vision Research* **45**, 5, pp. 533 – 550.

Libet, B. (2004). *Mind time. The temporal factor in consciousness*, Perspectives in Cognitive Neuroscience (Harvard University Press, Cambridge MA and London), edited by S. Kosslyn.

Maass, W., Natschläger, T. and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations, *Neural Comput* **14**, pp. 2531–60.

Markram, H. (2006). The blue brain project, *Nat Rev Neurosci* **7**, pp. 153–160.

Martius, G., Nolfi, S. and Herrmann, J. (2008). Emergence of interaction among adaptive agents, in M. Asada, J. C. T. Hallam, J.-A. Meyer and J. Tani (eds.), *From Animals to Animats 10: Proc. of the 10th Int. Conf. on the Simulation of Adaptive Behavior*, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg), pp. 457–466.

Maturana, H. (1978). Kognition, in P. Hejl, P. Köck and G. Roth (eds.), *Wahrnehmung und Kommunikation* (Peter Lang, Frankfurt), pp. 29–49.

Maturana, H. and Varela, F. (1980). *Autopoiesis and cognition: The realization of the living* (D. Reidel, Boston, MA).

Maturana, H. and Varela, F. (1987). *The tree of knowledge: The biological roots of human understanding* (Shambhala, Boston, MA).

McCarthy, J., M.Minsky, Rochester, N. and Shannon, C. (1955). A proposal for the Dartmouth summer research project on Artificial Intelligence, Funding proposal, (First documented use of the term 'Artificial Intelligence'). Retrieved from: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, retrieval date: 20.03.2008.

McClelland, J., Rumelhart, D. and Hinton, G. (1986). The appeal of parallel distributed processing, in D. Rumelhart, J. McClelland and the PDP Research Group (eds.), *Parallel Distributed Processing*, Vol. 1 (MIT Press, Cambridge MA), pp. 3–40.

McGann, M. (forthcoming). Perceptual modalities: Modes of presentation or modes of interaction? *Journal of Consciousness Studies* .

Melchner, L. v., Pallas, S. and Sur, M. (2000). Visual behavior induced by retinal projections directed to the auditory pathway, *Nature* **404**, pp. 871–875.

Merleau-Ponty, M. (2002). *Phenomenology of perception*, Routledge Classics (Routledge, London and New York), translated by C. Smith. French original published in 1945.

Metzinger, T. (ed.) (2000). *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, Cambridge MA).

Millikan, R. (1984). *Language, thought and other biological categories: New foundations for realism* (MIT Press, Cambridge MA).

Milner, A. and Goodale, M. A. (1995). *The visual brain in action* (Oxford University Press, New York).

Minsky, M. (1961). Steps toward artificial intelligence, *Proceedings of the IRE* **49**, pp. 8–30.

Minsky, M. and Papert, S. (1969). *Perceptrons. An Introduction to Computational Geometry* (MIT Press, Cambridge MA).

Morasso, P., Mussa Ivaldi, F. and Ruggiero, C. (1983). How a discontinuous mechanism can produce continuous patterns in trajectory formation and handwriting, *Acta Psychologica* **54**, pp. 83–98.

Moreno, A. and Etxeberria, A. (2005). Agency in natural and artificial systems, *Artificial Life* **11**, pp. 161–176.

Morrone, M. C., Ross, J. and Burr, D. (2005). Saccadic eye movements cause compression of time as well as space, *Nature Neuroscience* **8**, pp. 950 – 954.

Myin, E. and O'Regan, J. K. (2002). Perceptual consciousness, access to modality and skill theories: A way to naturalise phenomenology? *Journal of Consciousness Studies* **9**, pp. 27–45.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D. and Reserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds, *Developmental Science* **2**, pp. 164–174.

Nagel, S., Carl, C., Kringe, T., Märtin, R. and König, P. (2005). Beyond sensory substitution: Learning the sixth sense, *J. Neural Eng.* **2**, pp. R13–R26.

Newell, A. and Simon, H. (1963). GPS: A program that simulates human thought, in E. Feigenbaum and J. Feldman (eds.), *Computers and Thought* (R. Oldenbourg KG), pp. 279–293.

Nijhawan, R. (1994). Motion extrapolation in catching, *Nature* **370**, pp. 256–257.

Nijhawan, R. (2004). Motor space, visual space and the flash-lag effect, in C. Koch, R. Adolphs, T. Bayne, D. Leopold, G. Rees, S. Shimojo, P. Stoerig and P. Wilken (eds.), *Proceedings of the 9th annual meeting of the Association for the Scientific Study of Consciousness ASSC9, 24.-27.06.2004, Pasadena, California*, (Abstract).

Nijhawan, R. and Kirschfeld, K. (2003). Analogous mechanisms compensate for neural delays in the sensory and the motor pathways: Evidence from motor flash-lag, *Current Biology* **13**, pp. 749–753.

Noë, A. (2004). *Action in perception* (MIT Press, Cambridge, MA).

Nolfi, S. and Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines* (MIT Press, Cambridge MA).

Núñez, R. and Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time, *Cognitive Science* **30**, pp. 401–450.

Núñez, R. E. (forthcoming). Enacting infinity: Bringing transfinite cardinals into being, in O. Stewart, J. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

O'Regan, K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences* **24**, pp. 939–1011.

Pariyadath, V. and Eagleman, D. (2007). The effect of predictability on subjective duration, *PLoS ONE* **2**, p. e1264.

Petitmengin, C. (2005). Un exemple de recherche neuro-phénoménologique : L'anticipation des crises d'épilepsie, *Intellectica* **40**, pp. 63–89.

Petitmengin, C. (2006). Describing one's subjective experience in the second person. an interview method for the Science of Consciousness, *Phenomenology and the Cognitive Sciences* **5**, pp. 229–269.

Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence* (MIT Press, Cambridge MA).

Piaget, J. (1936). *La naissance de l'intelligence chez l'enfant* (Delachaux et Niestlé, Neuchátel-Paris).

Piaget, J. (1969). *The child's conception of time* (Routledge & Kegan Paul, London), translated by A. J. Pomerans. French original published in 1946.

Port, R. and van Gelder, T. (eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition* (MIT Press, Cambridge MA).

Prinz, J. (2006). Putting the brakes on enactive perception, *Psyche* **12**, pp. 1–19.

Quinn, M. (2001). Evolving communication without dedicated communication channels, in J. Kelemen and P. Sosik (eds.), *Advances in Artificial Life: Sixth European Conference on Artificial Life (ECAL01)* (Springer), pp. 357–366.

Revonsuo, A. and Newman, J. (1999). Editorial: Binding and consciousness, *Consciousness and Cognition* **8**, pp. 123–127.

Rodriguez, E., George, N., Lachaux, J.-P., Martinerie, J., Renault, B. and Varela, F. J. (1999). Perception's shadow: Long-distance synchronization of human brain activity, *Nature* **397**, pp. 430–433.

Rohde, M. (2003). Dynamical properties of self-regulating neurons, Bachelor's thesis, Institute for Cognitive Science, University of Osnabrück. BSc in Cognitive Science.

Rohde, M. (2008). *Evolutionary Robotics Simulation Models in the Study of Human Behaviour and Cognition*, Ph.D. thesis, University of Sussex, Department of Informatics.

Rohde, M. (2009). No need for intellectual straightjackets, *Adaptive Behavior* **17**, pp. 334–337, reply to a target article by B. Webb.

Rohde, M. and Di Paolo, E. (2005). *t* for two: Linear synergy advances the evolution of directional pointing behaviour, in M. Capcarrere, A. Freitas, P. Bentley, C. Johnson and J. Timmis (eds.), *Advances in Artificial Life: 8th European Conference, ECAL 2005, Canterbury, UK, September 5-9, 2005, Proceedings*, *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, Vol. 3630 (Springer, Heidelberg), pp. 262–271.

Rohde, M. and Di Paolo, E. (2006). 'Value signals' and adaptation: An exploration in Evolutionary Robotics, Tech. Rep. 584, Centre for Research in Cognitive Science, University of Sussex, UK.

Rohde, M. and Di Paolo, E. (2007). Adaptation to sensory delays: An Evolutionary Robotics model of an empirical study, in F. Almeida e Costa, L. Rocha, E. Costa, I. Harvey and A. Coutinho (eds.), *Proceedings of the 9th European Conference on Artificial Life*, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg), pp. 193–202.

Rohde, M. and Di Paolo, E. (2008). Embodiment and perceptual crossing in 2D: A comparative Evolutionary Robotics study, in M. Asada, J. Hallam, J.-A. Meyer and J. Tani (eds.), *Proceedings of the 10th International Conference on the Simulation of Adaptive Behavior SAB'08 in Osaka, Japan 7.-12.7.2008*, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg), pp. 83–92.

Rohde, M. and Ikegami, T. (2009). Editorial: Agency in natural and artificial systems, *Adaptive Behavior* **17**, pp. 363–366, editorial introduction to the special issue.

Rohde, M. and Stewart, J. (2008). Ascriptional and 'genuine' autonomy, *BioSystems* **91**, 2, pp. 424–433, special issue on modelling autonomy.

Rosenfield, I. (1988). *The Invention of Memory: A New View of the Brain* (Basic Books, New York).

Ross, S. (1984). *Differential Equations*, 3rd edn. (John Wiley & Sons, New York).

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach* (Prentice Hall, New Jersey).

Rutkowska, J. (1997). What's value worth? Constraining unsupervised behaviour acquisition, in P. Husbands and I. Harvey (eds.), *Proceedings of the Fourth European Conference on Artificial Life EACL97, Brighton UK* (MIT Press), pp. 290–298.

Searle, J. (1980). Minds, brains and programs, *Behavioral and Brain Sciences* **3**, pp. 417–424.

Seth, A. (2007). Measuring autonomy by multivariate autoregressive modelling, in F. Almeida e Costa, L. Rocha, E. Costa, I. Harvey and A. Coutinho (eds.), *Proceedings of the 9th European Conference on Artificial Life*, Lecture Notes in Artificial Intelligence (Springer, Berlin, Heidelberg), pp. 475–484.

Seth, A. and Edelman, G. (2007). Distinguishing causal interactions in neural populations, *Neural Computation* **19**, pp. 910–933.

Shanon, B. (2001). Altered temporality, *Journal of Consciousness Studies* **8**, pp. 35–58.

Shemmell, J., Hasan, Z., Gottlieb, G. and Corcos, D. (2007). The effect of movement direction on joint torque covariation, *Experimental Brain Research* **176**, pp. 150–158.

Smith, K. and Smith, W. (1962). *Perception and motion: An analysis of space-structured behavior* (Saunders).

Smith, R. (2004). Open Dynamics Engine (0.5 release), Retrieved: 10.1.2005, URL: http://ode.org.

Snel, M. and Hayes, G. M. (2008). Evolution of valence systems in an unstable environment, in M. Asada, J. Hallam, J.-A. Meyer and J. Tani (eds.), *Proceedings of the 10th International Conference on the Simulation of Adaptive Behavior SAB'08 in Osaka, Japan 7.-12.7.2008*, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg), pp. 12–21.

Sporns, O. and Edelman, G. (1993). Solving Bernstein's problem: A proposal for the development of coordinated movement by selection, *Child Dev.* **64**, pp. 960–981.

Steiner, U. (ed.) (1997). *Husserl*, Philosophie jetzt! Edited by Peter Sloterdijk (Eugen Diederichs Verlag, München), selected writings from Husserl's lifework 1901-1936.

Stetson, C., Cui, X., Montague, P. and Eagleman, D. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation, *Neuron* **51**, pp. 651–659.

Stewart, J. (2004). *La vie : Existe-t-elle?* (Vuibert, Paris).

Stewart, J. (forthcoming). Foundational issues in enaction as a paradigm for cognitive science: From the origin of life to consciousness and writing, in J. Stewart, O. Gapenne and E. Di Paolo (eds.), *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Stewart, J. and Gapenne, O. (2004). Reciprocal modelling of active perception of 2-D forms in a simple tactile-vision substitution system, *Minds and Machines* **14**, pp. 309–330.

Stewart, J., Gapenne, O. and Di Paolo, E. (eds.) (forthcoming). *Enaction: Towards a New Paradigm for Cognitive Science* (MIT Press, Cambridge, MA).

Stilling, N., Weisler, S., Chase, C., Feinstein, M., Garfield, J. and Rissland, E. (1998). *Cognitive Science: An Introduction*, 2nd edn. (MIT Press, Cambridge MA), original in 1995.

Strogatz, S. (1994). *Nonlinear Dynamics and Chaos. With Applications to Physics, Biology, Chemistry and Engineering* (Perseus Books, Cambridge MA).

Thelen, E. and Smith, L. (1994). *A dynamic systems approach to the development of cognition and action* (MIT Press, Cambridge, MA).

Thompson, E. (2005). Sensorimotor subjectivity and the enactive approach to experience, *Phenomenology and the Cognitive Sciences* **4**, pp. 407–427.

Thompson, J., Ottensmeyer, M. and Sheridan, T. (1999). Human factors in telesurgery: Effects of time delay and asynchrony in video and control feedback with local manipulative assistance, *Telemedicine Journal* **5**, pp. 129–137.

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity, in M. Bullowa (ed.), *Before speech* (Cambridge University Press, Cambridge), pp. 39–52.

Tuci, E., Quinn, M. and Harvey, I. (2002). Evolving fixed-weight networks for learning robots, in *Congress on Evolutionary Computation CEC2002 (Proceedings)* (IEEE Press), pp. 1970–1975.

Turing, A. (1950). Computing machinery and intelligence, *Mind* **59**, pp. 433–460.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science, *Behavioral and Brain Sciences* **21**, pp. 615–628.

Varela, F. (1991). Organism: A meshwork of selfless selves, in A. Tauber (ed.), *Organism and the origin of the self* (Kluwer Academic, Netherlands), pp. 79–107.

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem, *Journal of Consciousness Studies* **3**, pp. 330–350.

Varela, F. (1997). Patterns of life: Intertwining identity and cognition, *Brain and Cognition* **34**, pp. 72–87.

Varela, F. (1999). The specious present: The neurophenomenology of time consciousness, in J. Petitot, F. Varela, B. Pachoud and J.-M. Roy (eds.), *Naturalizing Phenomenology* (Stanford University Press, Stanford), pp. 266–314.

Varela, F., Maturana, H. and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model, *BioSystems* **5**, pp. 187–196.

Varela, F., Thompson, E. and Rosch, E. (eds.) (1991). *The embodied mind: Cognitive science and human experience* (MIT Press, Cambridge, MA).

Vermersch, P. (1994). *L'entretien d'explicitation en formation initiale et en formation continue* (E.S.F., Paris).

Verschure, P., Wray, J., Sporns, O., Tononi, G. and Edelman, G. (1995). Multilevel analysis of classical conditioning in a behaving real world artifact, *Robotics and Autonomous Systems* **16**, pp. 247–265.

von Holst, E. and Mittelstaedt, H. (1950). Das Reafferenzprinzip, *Naturwissenschaften* **37**, pp. 464–76.

Webb, B. (2009). Animals versus animats: or why not model the real iguana? *Adaptive Behavior* **17**, pp. 269–286.

Weber, A. (2003). *Natur als Bedeutung. Versuch einer semiotischen Theorie des Lebendigen* (Königshausen und Neumann, Würzburg).

Weber, A. and Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality, *Phenomenology and the Cognitive Sciences* **1**, pp. 97–125.

Weiss, P. and Jeannerod, M. (1998). Getting a grasp on coordination, *News Physiol. Sci.* **13**, pp. 70–75.

Welch, R. (1978). *Perceptual Modification: Adapting to Altered Sensory Environments* (Academic Press, New York).

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step* (MIT Press, Cambridge MA).

Yamauchi, B. and Beer, R. (1994). Sequential behaviour and learning in evolved dynamical neural networks, *Adaptive Behavior* **2**, pp. 219–246.

Yarrow, K., Haggard, P., Heal, R., Brown, P. and Rothwell, J. C. E. (2001). Illusory perceptions of space and time preserve cross-saccadic perceptual continuity, *Nature* **414**, pp. 302–305.

Zaal, F., Daigle, K., Gottlieb, G. and Thelen, E. (1999). An unlearned principle for controlling natural movements, *Journal of Neurophysiology* **82**, pp. 255–259.

Author Index